Simulation of the phenol and chlorophenol removal using combined adsorption and biodegradation: regression analysis and data-mining approach

1Parthiban P., 2Malarvili T., 3Ashutosh Das, 4Mukesh Goel*, 4Abhishek Asthana., 5B.V. Ramanaiah
1Research Scholar, Department of Biochemistry, Rajah Serfoji government college (Autonomous),
Affiliated to Bharathidasan University
2Principal, Bharathidasan University Constituent Arts & Science College,
Navalurkuttapattu-620 009
3Centre for Environmental Engineering, PRIST Deemed University, Thanjavur-613 403,
Tamilnadu, India
4Department of Engineering and Mathematics, Sheffield Hallam University,
Sheffield, UK - S1 1WB
5Department of Chemical Engineering, MVGR College of Engineering,
Vizianagaram, Andhra Pradesh 535005

*Corresponding author: Dr. Mukesh Goel
Email: Mukesh.Goel@shu.ac.uk

Abstract:

It is widely known that phenols and chlorophenols (CP) are two of the most toxic chemicals and a versatile treatment is imperative to tackle this evil of industrialization. Adsorption using low-cost adsorbents is advantageous and economical; however, it has not turned out to be feasible technology. Biological treatment is much more flexible, useful, and environmentally friendly and a combination of biological treatment and adsorption has yielded much better results compared to using them individually. However, very few works are applying statistical methods in elucidating the importance of various options in such a combined study. This work focused on the effect of temperature, initial concentration of chemicals and adsorbent dosage on the removal of these chemicals. Furthermore, it compares various processes, viz., biological treatment (bio), sequential biological and adsorption (seq), and simultaneous biological and adsorption (sim) methods in treating phenols and chlorophenols. A range of linear regression models was developed to predict the percentage reduction for each of the processes used (bio, sim & seq), and each of these models was statistically significant as evident from R-square values and the ANOVA table for regression parameters. A data-mining tree-classifier for
modelling the phenol and CP removal was also developed. The data mining study indicates the initial concentration of the solvent and temperature to be the primary classifying parameters.

**Keywords:** Biodegradation, Combined adsorption and biodegradation, Phenols, Chlorophenols, linear regression, data-mining tree-classifier
**Introduction**

Phenol and chlorophenol are two of the major waste products in a wide range of manufacturing industries, viz. chemical and pharmaceutical, paint and textiles, paper and pulp, plastics and polymer, oil, and gasoline as well as cooking ovens and metallurgical furnaces (Loncar et al., 2011). Most of the wastewaters have a very high organic load as well as toxicity, and these certainly present a clear danger to aquatic organisms. Most of the methods that are generally engaged to treat the wastewater are meant for removing and reducing contaminants to a safe level (Tetteh et al., 2019). However, toxic chemicals treatments are not encouraging, as even so-called clean water seems to have a significant amount of these chemicals. Thus, it is important to improve existing technologies to eliminate these compounds from wastewater. Researchers have tried advanced oxidation processes (AOPs), coagulation-flocculation, ozonation, wet air oxidation, photocatalysis, constructed wetlands, membrane filtration, adsorption, etc (Amor et al., 2019; Goel et al., 2011; Karunakaran et al., 2012). There are also reports of integrating two or more than two methods to produce a complementary result. Goel et al. (2010) tried the integrated biological and photochemical methods for the treatment of chlorophenol and found a substantial increase by combining both methods. Similar other methodologies are reported for the integrated treatment of various toxic effluents (Goel et al., 2012; Parthiban et al., 2019, 2020; Pookpoosa et al., 2015; Orozco et al., 2016; Zielinska et al., 2014).

Statistics is being applied in all the fields of engineering and technology and slowly gaining its due attention in wastewater treatment. Recent computational advancements provide a rare opportunity for data-driven control of the treatment and also have a huge scope for performance improvement. However, big data in wastewater treatment plants are not adequately explored areas, which is mainly due to the lack of knowledge in data science in the wastewater treatment industry (Alam et al., 2007). Saleh et al (2018) emphasised the implementation of statistical
optimization of the parameters and conditions in evaluating the performance of adsorbents in water treatment to refine the usage of chemicals and reagents in experimental work. The treatment processes exhibit non-linear, non-stationary, random correlations, which are not easy to be modelled. Some of the researchers working on wastewater treatment have screened key determinants such as phosphate, iron sulphate, calcium chloride, etc. using Plackett–Burman design, which was further optimized by response surface methodology (Basaleh et al., 2021, Subhasis et al., 2013, Garg et al., 2008, Silva et al., 2007). Sivasubramanian et al., (2014) have also used response surface methodology to assess the effect of various parameters like initial phenol concentration, temperature, pH and time on phenol degradation. Similarly, Ray et al. (2009) developed a four-factor three-level Box-Benkhen design (BBD) to describe the photocatalytic degradation of phenol in an aqueous media.

Extraction of knowledge from data is called data mining. Data mining performed on the collected data helps in different ways. It is also progressively applied in various fields of science and technology (Han et al., 2011). Its application in wastewater treatment is increasing in the last decade (Haimi et al., 2013). They are found to be significant for refining wastewater treatment operation and automation. It is especially applicable in the characterisation of inlet wastewater (Kim et al., 2016). This work aims to correlate phenol and chlorophenol degradation using statistical models to optimize the interactive factors involved in the combined adsorption and biodegradation of phenol and CP. Subsequently, models concerning sequential adsorption and biodegradation of phenol and CP in terms of COD are developed through a data-mining tree-classifier for modelling the toxicity removal.

Materials and Methods

Reagents

Both phenol and chlorophenol were supplied by Merck. Analysis was carried out using the 4-Amino Antipyrine method using UV spectrophotometer (Make: Systronics, Model: 2202).
Chromatographic measurements were carried out using HPLC (Make Shimadzu, Model: 10AS). Distilled water was used for all the experimental studies. All the reagents used were of analytical grade.

**Preparation of adsorbents**

Jatropha deoiled cake (JDC) used for adsorption was procured from Bannari Amman sugar mills, Erode, Tamilnadu, India. They were dried overnight at 105°C in a hot air oven, followed by crushing and sieved with a 125 μm mesh. The powdered samples were dried using a desiccator and stored in airtight containers. The dried adsorbents were impregnated with H₃PO₄ in the ratio of 1:2. The activation method involves the use of 100 g of adsorbents with 120 ml of H₃PO₄ mixed and kept overnight at room temperature. It was then kept in a muffle furnace in the presence of a nitrogen atmosphere at 500°C for 2 hours. The samples so generated were washed with distilled water until the pH becomes neutral. Then the powder obtained was dried and sieved (at 125 microns) for the adsorption process (Parthiban et al., 2019).

**Adsorption experiments**

Adsorption studies were conducted routinely by batch technique. For this purpose, 0.2 g of JDC was added to 50 mL phenol solution in 100 mL flasks. The experiments with phenol and CP were conducted for six different initial concentrations ranging from 25 ppm to 400 ppm and 20 ppm to 150 ppm respectively. The flasks were placed on a rotating shaker (Systronics) with constant agitation, 150 rpm, at room temperature for 24 hrs. Samples were centrifuged and the phenol or chlorophenol concentration was determined using a spectrophotometer. All the experiments were carried out in triplicates.

**Biological experiments**

Phenol and CP were subjected to combined treatment by using adsorption and biological degradation. Integrated adsorption and biological treatment were conducted in sequential
mode. The adsorbed sample was made up to 30 ml by adding the required quantity of media containing glucose and inoculum. The conical flasks (100 mL) were kept in an orbital shaker at the desired temperature and neutral pH was maintained during the biological experiments. The glucose concentrations used were 1 g/L, 2 g/L and 3 g/L and the temperatures maintained were 25, 30, 35 and 40°C. Both the COD and phenol, CP concentrations were recorded every 24 hours.

**Multivariate analysis using statistics tools**

Multivariate analysis is utilized to analyse complex sets of data, which are not possible with univariate analysis and even bivariate analysis. The analysis involves several dependent variables yielding a singular outcome. The real-world data such as weather data, machine performance data, human-behaviour data and so on are all multivariate. This analysis is generally performed with statistical software, such as SPSS or SAS. The conclusion drawn using multivariate analysis is more accurate and realistic (Clautier et al., 2008; Edet et al., 2013). In our work, process parameters like temperature, dosage and initial concentrations were selected as the independent variables. COD reduction was the dependent variable. The significance of the data obtained during the degradation of phenol and chlorophenol was further analysed, and empirical models were developed for percentage COD reduction as a function of glucose concentration, initial concentration, and temperature.

The multiple regression modeling, employed in the present work, has been computed by solving the following equations:

\[
\hat{Y} = b_0 + b_1 x_{i1} + b_2 x_{i2} \ldots \ldots \ldots b_n x_{in}
\]

where \(\hat{Y}\) is the calculated Y value, \(x_{i1}, x_{i2} \ldots \ldots x_{in}\) refer to the independent variables of \(i^{th}\) equation and \(b_1, b_2 \ldots \ldots b_n\) refer to the corresponding coefficients computed using least square method.
For logistic regression modelling, the following equations have been used to solve:

\[ \log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_{i1} + b_2 x_{i2} \ldots \ldots b_n x_{in} \]

where P is the probability of having the outcome and \( P / (1-P) \) is the odds of the outcome.

**Development of a data-mining tree-classifier for modelling the phenol and CP removal**

Data mining is gaining a lot more attention in wastewater treatment (Mannina et al., 2011). Decision tree (or classifier) learning is regularly applied in data mining. They are some of the learning methods of a concept, produced by a learning algorithm, given the instances of that notion. Decision trees are straightforward to understand, and the model generated can be conveyed as a set of decision rules. It produces the output as a binary tree-like structure, resulting in relatively easy explanation to the end-user and effortless classification of important variables. The tree model will have the rules to forecast the target variable (Balakrishnan et al., 2019; Peter and Petermann, 2018). The flowchart consists of root, intermediate and leaf nodes. The root divides all data into two or compound subgroups. The intermediate node articulates a condition for each independent variable and the leaf node signifies the consequences of a blend of decisions (Deepnarain et al., 2019). In this work, the Weka classifier was used for developing a data-mining tree-classifier. Weka is a data-mining workbench comprising several algorithms for classification, data pre-processing, clustering, etc. It is a very broad workbench and is free, open-source software (Hall et al., 2009). A chemical-based classifier system (C 4.5 algorithm) was also developed in this study using J48 classifier (Ross Quinlan, 1993), with confidence factor = 0.25 and minimum number of objects = 2 and number of folds =3 (Pruned) and evaluated the performance of empirical models regarding the baseline classifier (ZeroR) and the combined classifier. This study also aims at exploring the levels of contribution and threshold values of the process parameters used (namely, glucose concentration, initial concentrations, and temperature).
The objective function used here is maximization of information gain, i.e., minimization of difference in entropies \( h(x) \) across the branches about any given node, defined by:

\[
h(x) = - \sum_{i=1}^{n} P(x_i) \log P(x_i)
\]

where, the discrete random variable, with possible outcomes \( x_1, x_2, \ldots, x_n \), occurs with corresponding probabilities \( P(x_1), P(x_2), \ldots, P(x_n) \)

**Results and Discussion**

*Exploring the relationship of experimental conditions with phenol and CP removal*

The experimental parameters (i.e., temperature, dosage, and concentration) were used to draw a scatterplot against the percentage removal of phenol and chlorophenol, to estimate the strength of their interrelationship. There is a definitive inverse correlation between concentration and percentage reduction and a mild direct correlation between temperature and percentage removal in both cases (phenol and chlorophenol) using all methods (Figure 1) and concentration seems to have a direct relationship with the removal of both phenol and chlorophenol.

Hence, to estimate the variation of removal percentage with concentration, descriptive statistics for all the concentrations were evaluated and the means of percentage reduction were plotted against various initial concentrations of phenol and chlorophenol. The representations are detailed in Figure 2.

*Estimation of the significance of phenol and CP removal methods*

To evaluate whether the variation amongst the methods used for the reduction of phenol and chlorophenol is statistically significant or just a random outcome, a one-way Analysis of Variance (ANOVA) was conducted between reductions of the solvents (i.e., phenol or chlorophenol) about the methods used (bio, sim and seq) (Table-1)
As indicated by table 1, the methods involved show statistically significant variation at a 95% confidence interval, as indicated by p-value < 0.05. Interestingly, sequential removal seems to be statistically most significant, compared to others (i.e., bio and sim), as indicated by high F-statistics in both chlorophenol and phenol, the former being 3-times more significant compared to the latter.

Combined adsorption and biodegradation treatment was found to be especially suitable for the recalcitrant chemicals, such as phenol and CP. Ananthanarayan et al. (2018) found that the greater success of combined mode could be due to improved adsorption of non-biodegradable compounds. Biodegradation generates many intermediates which can easily attract adsorbents facilitating the adsorption of these intermediates.

**Linear regression for prediction of phenol and CP removal**

In order to predict the percentage removal of phenol and CP, linear regression was carried out with reduction percentage as the dependent variable and temperature, dosage and initial concentration as independent variables. The statistical software IBM SPSS Statistics 24.0 was chosen to execute the regression analyses. It is a highly popular and reliable program and is generally used by several academicians and researchers. The residuals of regression were found to be normally distributed through histogram and P-P plots (data not shown). The results of the regression study are presented in Table 2. However, there are regression modelling for the adsorption of various chemicals processes, the combined adsorption and biological work has not been explored using regression modelling (Das et al., 2015). Wu et al. (2011) studied the bioavailability of herbicides in the soil after the adsorption. Detailed regression modelling of the adsorption process was presented, but only kinetic parameters of biological degradation were obtained. Lin (2017) developed an unsteady state mathematical model for the kinetics of 2-chlorophenol adsorption and biodegradation.
All the models presented in Table 2 are found to have F-statistics higher than table values, as indicated by the ANOVA table for regression (not shown here). All the models developed have $r^2$ values of more than 0.75, and each parameter is significant in all the models provided, except dosage in sequential approach for chlorophenol (and hence there is no loss in model accuracy in dropping dosage from the regression model in this case only, although, in all other models, dosage does play an important role). The sequential removal yielded a higher value especially in the case of phenol removal.

The actual and predicted values (through regression) are presented in Figure 3. As discussed before, the linear regression model best fits the experimental values. It was observed that each parameter is significant in all the models provided, except dosage in a sequential approach for chlorophenol. Hence there is no loss in the model accuracy in dropping dosage from the regression model in this case only, although in all other models, dosage does play an important role.

**Logistic regression modelling of the phenol and CP removal for the different methods studied**

To avoid the experimental variability and to utilize the categorical variables as well in prediction, the percentage reduction of the solvents used were codified as ordinal values, namely, $>90$ (highest), 90 to 75 (high), < 75 to 50 (average), < 50 to 25 (low) and < 25 (lowest) and multinomial logistic regression was attempted.

As observed, in all the cases studied, -2 Log Likelihood of reduced model for all the parameters is highest for concentration, followed by the temperature. These are significantly higher than an intercept-only model, except dosage, which is close to the intercept-only model. The same pattern is also observed by their chi-square values. In fact, only in the case of average and low-level reduction, a logistic model was found to be applicable. It was found that the logistic
regression was less robust than the linear regression and indicates that only in the case of average and low-level reduction, a logistic model could be applicable.

**Data-mining tree-classifier for modelling the phenol and CP removal**

C4.5 classifier using Weka 3.8 was used to develop the data-mining tree-classifier for modelling the phenol and CP removal. J48 algorithm was employed in the WEKA for initiation of decision trees. The data were normalized between 0.05 and 0.95 before the model was implemented. The input parameters were the effect of initial concentration, dosage, and temperature in all three processes, viz., biological treatment, sequential biological and adsorption, and simultaneous biological and adsorption methods in treating phenols and chlorophenols. The data set is allocated to the learning set, which was employed for model building and the test set. The test set determines the model precision. Predicted classes are in the leaves of the decision trees. The models are compared with ZeroR model, which was used as the benchmark. The model estimates are given in Table 3. Fig. 4-6 shows the biological treatment, sequential treatment, and simultaneous biological & adsorption methods in the J48 model tree with three input parameters respectively. The results of the J48 model tree of all the three treatment methods of chlorophenol shows a significant number of domains. The data mining study indicates the initial concentration of the solvent and temperature to be the primary classifying parameters. The dosage plays a role only at a higher temperature and/or concentration. It can also be noted that sequential treatment is preferred by both phenol and chlorophenol which has the most significant number of domains than other tree regression models.

**Conclusions**

This paper optimizes the synergistic factors affecting the combined adsorption and biodegradation of phenol and chlorophenol (CP) using statistical models. A data-mining tree-classifier was then used for the sequential adsorption and biodegradation of phenol and CP in
terms of COD. The result shows there is a definitive inverse correlation-ship between concentration and percentage reduction and mild direct correlation-ship between temperature and percentage removal in both cases (phenol and chlorophenol) using all methods as indicated by scatterplots, plots, and data of variation of mean percentage reduction with the initial concentration of the solution. One-way ANOVA results show statistically significant variation at a 95% confidence interval for reductions of the solvents (i.e., phenol or chlorophenol) with regard to the methods used (bio, sim and seq). Based on the three methods, sequential removal seemed to be statistically most significant, compared to others (i.e., bio and sim), as indicated by high F-statistics in both chlorophenol and phenol, the former being 3-times more significant compared to the latter. Linear regression models were also developed involving temperature, concentration and dosage to predict the percentage reduction. These models were found to be statistically significant as apparent from high R-square values and the ANOVA table. The results are matching with the experimental values. Data mining reveals the importance of the initial concentration of the solvent and temperature as the primary classifying parameters. The adsorbent dosage seems to have a presence only at a higher temperature and/or concentration. Future work involves blending neural networks, fuzzy logic, and other data analysis tools to the created expert system, that would result in an improved expert system for parameters governing the wastewater treatment. Detailed mathematical models would also be developed to predict the effluent quality.

**Data Availability Statement**

All data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.
References


sludge wastewater treatment plant. Process Safety and Environmental Protection, 126: 25-34.


Subhasis Mandal, Biswanath Bhunia, Anuj Kumar, Dalia Dasgupta, Tamal Mandal, Siddhartha Datta and Pinaki Bhattacharya. 2013. A statistical approach for optimization of media components for phenol degradation by Alcaligenes faecalis using Plackett–Burman and response surface methodology, Desalination and Water Treatment, 51:31-33, 6058-6069, DOI: 10.1080/19443994.2013.769746


