JOURNAL OF ENVIRONMENTAL SCIENCES 75 (2019) 201–208

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jes

JES
JOURNAL OF ENVIRONMENTAL SCIENCES
www.jesc.ac.cn

Statistical regression modeling for energy consumption in wastewater treatment

Yang Yu[1], Zhihong Zou[1,*], Shanshan Wang[1,2,*]

1. School of Economics and Management, Beihang University, Beijing 100191, China
2. Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing 100191, China

ARTICLE INFO

Article history:
Received 29 September 2017
Revised 19 March 2018
Accepted 20 March 2018
Available online 10 April 2018

Keywords:
Energy consumption modeling
Wastewater treatment
Semi-parametric model
Bayesian quantile regression

ABSTRACT

Wastewater treatment is one of critical issues faced by water utilities, and receives more and more attentions recently. The energy consumption modeling in biochemical wastewater treatment was investigated in the study *via* a general and robust approach based on Bayesian semi-parametric quantile regression. The dataset was derived from a municipal wastewater treatment plant, where the energy consumption of unit chemical oxygen demand (COD) reduction was the response variable of interest. *Via* the proposed approach, the comprehensive regression pictures of the energy consumption and truly influencing factors, *i.e.*, the regression relationships at lower, median and higher energy consumption levels were characterized respectively. Meanwhile, the proposals for energy saving in different cases were also facilitated specifically. First, the lower level of energy consumption was closely associated with the temperature of influent wastewater, and the chroma-rich wastewater also showed helpful in the execution of energy saving. Second, at median energy consumption level, the COD-rich wastewater played a determinative role in the reduction of energy consumption, while the higher quality of treated water led to slightly energy intensive. Third, the higher level of energy consumption was most likely to be attributed to the relatively high temperature of wastewater and total nitrogen (TN)-rich wastewater, and both of the factors were preferably to be avoided to alleviate the burden of energy consumption. The study provided an efficient approach to controlling the energy consumption of wastewater treatment in the perspective of statistical regression modeling, and offered valuable suggestions for the future energy saving.

Introduction

As a significant area in environmental sciences, wastewater treatment has been attracting considerable research in recent decades (Gao et al., 2017; Han et al., 2013; Wang et al., 2016). It serves a double purpose of resolving the water pollution and simultaneously motivating the water reuse, hence it protects the environment and alleviates the shortage of water. Nevertheless, wastewater treatment is an energy-intensive industry that the high cost of energy consumption hinders its development to some degree. Thereinto, the electric energy consumption accounts for the largest ratio (Jin and Yang, 2012), and the biochemical treatment, which consumes about 60% of the total energy (Li, 2010; Jin et al., 2009; Bai, 2012), is the most energy-consuming unit in the entire process. Therefore, it is very necessary to model the electric energy consumption of

---

[*] *Corresponding authors. E-mail:* zouzhihong@buaa.edu.cn (Zhihong Zou), sswang@buaa.edu.cn (Shanshan Wang).

https://doi.org/10.1016/j.jes.2018.03.023

biochemical treatment, analyze the effect of influencing factors and present proposals for energy saving in a mathematical point of view.

Owing to the complicated relationships among numerous influencing factors, the mathematical modeling for electric energy consumption in wastewater treatment is still an open problem. The existing methods mainly focused on the mechanism modeling, neural network and statistical regression. Whereas, most of the mechanism models only considered single influencing factor and cannot explain the law of energy consumption correctly (Yi et al., 2009; Kusiak et al., 2013; Han et al., 2016). Jin et al. (2014) employed back-propagation (BP) neural network in the prediction of energy consumption in anoxic–oxic (A/O) process. Huang et al. (2013) utilized Elman neural network to model the relationship between energy consumption and discharged water quality, yet the system dynamics was not considered in the method. Han et al. (2016) proposed an adaptive regressive kernel function to deal with the problem and obtained well model precision. Statistical regression modeling is an approach which explores the function relationship between response variable and predictors based on data. Jiang et al. (2014) utilized statistical methods to analyze the influencing factors of energy consumption in municipal wastewater treatment plant. Yang et al. (2008), Liang (2014) and Ren et al. (2015) applied power function regression, multivariate linear regression and exponential regression respectively to modeling the energy consumption in wastewater treatment, while the parametric methods presume fixed function form before estimation and usually lead to misspecification of the model.

Fortunately, semi-parametric regression methods avoid the above concerns naturally. It is a data-driven approach which relaxes the presumptions on parameter space and function form, and thus obtains better flexibility and robustness. In particular, partial linear single-index model (PLSIM) (Carroll et al., 1997; Liang et al., 2010; Boente and Rodriguez, 2012) is a widely used semi-parametric regression model. It retains the advantages of interpretability in linear regression and generality in non-parametric models. Furthermore, the dimensionality of the model is reduced significantly *via* the index term. In short, incorporating the PLSIM into the energy consumption can not only detect the unknown relationship between energy consumption and influencing factors, but also lead to explicit interpretations of the impact of factors.

However, there exist numerous potential factors that affect the variation of energy consumption in wastewater treatment, some of the factors are highly correlated and exert repetitive impacts, and some of them are actually insignificant to the energy consumption. Consequently, a crucial step before figuring out the function relationship is to find out the truly contributing factors and exclude the unnecessary ones. Indicator model selection (IMS) is a popular variable selection technique in Bayesian framework (Araki et al., 2015; O'Hara and Sillanpaa, 2009). It uses binary variables to indicate which predictors are significant in the true model, and in terms of prior knowledge and observed data, the posterior inference of binary indicators could be achieved with the approximation to posterior distributions by Markov Chain Monte Carlo (MCMC) simulation. With the most direct spike-and-slab prior, the estimates of insignificant coefficients would be exactly zeros, and the corresponding predictors are regarded as excluded.

Nevertheless, the above mentioned methods depend largely on conditional mean regression and can only reflect the regression information at mean level of the energy consumption. Actually, the significance of influencing factors and their impact on energy consumption might change with different levels of energy consumption. Additionally, the estimates from the conditional mean regression are sensitive to outliers, and the estimation efficiency might be impaired in the case of usual non-normal error distributions.

In these circumstances, quantile regression serves an important alternative (Koenker and Bassett, 1978). The methodology could characterize the function relationship at any interested quantiles of the distribution of response variable, and can also capture the significant factors for corresponding quantile of response variable; therefore it provides more comprehensive regression information and better adaptability to non-normal random errors. Particularly, with the asymmetric Laplace distribution (ALD) assigned to the error term, quantile regression could be implemented straightforwardly *via* Bayesian approaches (Yu and Moyeed, 2001; Koenker and Machado, 1999). One of the attractions in ALD is that it fits the real world data more suitably than common symmetric distributions in virtue of the peaked mode and fat tail of the distribution. The Bayesian quantile estimates under ALD errors are inferred from the Markov chain of parameters using the MCMC mechanism.

Motivated by those considerations, the aim of the study was to model the electric energy consumption in biochemical wastewater treatment at different levels of energy consumption, and present proposals for energy saving in terms of the results. To achieve that, a semi-parametric PLSIM was adopted to approximate the function relationship between energy consumption and influencing factors, with the ALD being assigned as error distribution to formulate Bayesian quantile regression for the model and the IMS technique being applied to finding out truly contributing factors. By means of the proposed approach, the factors which are significant with respect to the change of energy consumption at lower, median and higher consumption levels were identified respectively, and the different trends that show how they affect the energy consumption across various levels were also depicted. Finally, the proposals with respect to the energy saving were summarized based on the mathematical models.

1. Energy consumption modeling based on Bayesian semi-parametric quantile regression

1.1. *Energy consumption and potential influencing factors*

To efficiently measure the actual energy consumption for pollutant removal, the electric energy consumption of unit chemical oxygen demand (COD) reduction in biochemical treatment was taken as the interested response variable to observe (*Energy_Consmp*, kWh/kg). The potential influencing factors included: influent pH (*pH*), influent biochemical oxygen demand concentration (*BOD*, mg/L), influent chemical oxygen demand concentration (*COD*, mg/L), influent suspend solid concentration (*SS*, mg/L), influent chroma (*Chrom*), influent total phosphorus concentration (*TP*, mg/L), influent total nitrogen concentration (*TN*, mg/L), influent NH_3-N concentration (*NH_3-N*, mg/L), treatment scale per day (*Scale*, m^3/day), national discharge standard that the treated water reaches (*Stand*), degree of coldness and hotness of influent wastewater (*Degree*).

The reason to utilize the influent pollutant concentration is that it is an observable and dominant factor in wastewater treatment, additionally, the influent concentrations are far greater than the effluent values, and most effluent pollutant concentrations have no obvious changes due to the restriction of national discharge standard, hence the influent pollutant concentration is used to represent the pollutant removal efficiency. In the study, *Stand* and *Degree* were handled as categorical variables. Specifically, *Stand* = 0,1,2,3 denote the standard of the third class, the secondary class, the first class B and the first class A respectively, which accords with the discharge standard of pollutants for municipal wastewater treatment plant GB18918–2002. For another factor, *Degree* was abstracted from the influent wastewater temperature for a more straightforward concept, and it was designed as three levels: *Degree* = 0,1,2, which point to the ranges 10–15°C, 15–25°C and 25–30°C respectively.

1.2. Bayesian semi-parametric quantile regression model for energy consumption

The function relationship between energy consumption and influencing factors is of crucial importance in the characterization for the effect of truly influencing factors on the energy consumption. Compared with the common parametric regression method, which determines the function form artificially and often leads to model misspecification, semi-parametric regression approach yields better generality and flexibility. It relaxes presumptions on the model form and utilizes useful available information to simplify the model. Hence, semi-parametric regression model was adopted in the study to deal with the function relationship between energy consumption and influencing factors.

PLSIM is one of the most popular semi-parametric models. It reconciles the parametric linear part and non-parametric single-index component. Among the influencing factors, the categorical variables *Stand* and *Degree* were assigned to the linear part, while the other factors were involved into the single-index term. The PLSIM facilitated easy-interpretably linear effect of *Stand* and *Degree* and flexibly nonlinear effect of *pH*, *BOD*, *COD*, *SS*, *Chrom*, *TP*, *TN*, *NH_3-N* and *Scale*. The function relationship between the energy consumption and influencing factors was fitted in the following PLSIM:

$$\begin{aligned} Energy\_Consmp_i = {} & \beta_1 Stand_{1_i} + \beta_2 Stand_{2_i} + \beta_3 Stand_{3_i} + \beta_4 Degree_{1_i} \\ & + \beta_5 Degree_{2_i} + g(\alpha_1 pH_i + \alpha_2 BOD_i + \alpha_3 COD_i \\ & + \alpha_4 SS_i + \alpha_5 Chrom_i + \alpha_6 TP_i + \alpha_7 TN_i \\ & + \alpha_8 NH_3\text{–}N_i + \alpha_9 Scale_i) + \varepsilon_i, \end{aligned} \tag{1}$$

where $\{Energy\_Consmp_i; X_i, V_i\}_{i=1}^{n}$ are the observations, ε_i is the independent and identically distributed (iid) random error, the observations $X_i = (Stand_{1_i}, Stand_{2_i}, Stand_{3_i}, Degree_{1_i}, Degree_{2_i})^T$, $V_i = (pH_i, BOD_i, COD_i, SS_i, Chrom_i, TP_i, TN_i, NH_3 - N_i, Scale_i)^T$. Here, $(Stand_{1_i}, Stand_{2_i}, Stand_{3_i})^T$ and $(Degree_{1_i}, Degree_{2_i})^T$ are dummy vectors to treat the categorical variables *Stand* and *Degree* respectively, with dummy variable $Stand_{k_i} = 1$ indicating that the category k is observed in the ith observation, 0 otherwise, and similar to $Degree_{j_i}$, $k = 1, 2, 3$, $j = 1, 2$. The unknown quantities to be estimated in the above model include the linear term $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)^T$, the index vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_9)^T$ and the univariate link function $g(.)$.

For brevity purpose, Model (1) is simplified as the framework below

$$Energy\_Consmp_i = X_i^T \boldsymbol{\beta} + g(V_i^T \boldsymbol{\alpha}) + \varepsilon_i, \tag{2}$$

There into, $X_i^T \boldsymbol{\beta}$ is the parametric linear part that serves straightforward interpretability, while $V_i^T \boldsymbol{\alpha}$ is the single-index term that twists the high-dimensional factors to a 1-dimensional quantity and affects the *Energy_Consmp* nonparametrically *via* $g(.)$.

In the semi-parametric energy consumption model, not all the potential influencing factors exert working effect on *Energy_Consmp*. In order to find out the truly contributing factors, IMS based algorithm was applied to the model. In detail, independent binary indicators r_{α_j} and r_{β_l} were introduced to the coefficient of each influencing factors, and the coefficient could be rewritten as $\alpha_{r_j} = r_{\alpha_j}\alpha_j$, $\beta_{r_l} = r_{\beta_l}\beta_l$, where the binary variables r_{α_j}, $r_{\beta_l} \in \{0,1\}$ and the value of one implies the corresponding factor is included in the model and zero otherwise, $j = 1, \ldots, 9$, $l = 1, \ldots, 5$. And the linear term $\boldsymbol{\beta}_r = (\beta_{r_1}, \ldots, \beta_{r_5})^T$, the index vector $\boldsymbol{\alpha}_r = (\alpha_{r_1}, \ldots, \alpha_{r_9})^T$, and the indicator vectors $\mathbf{r}_\beta = (r_{\beta_1}, \ldots, r_{\beta_5})^T$, $\mathbf{r}_\alpha = (r_{\alpha_1}, \ldots, r_{\alpha_9})^T$.

The link function $g(.)$ acts as an important role in Model (2), for it characterizes the nonlinear relationship between *Energy_Consmp* and $V_i^T \boldsymbol{\alpha}_r$. In the estimation for $g(.)$, the free knot spline provides a quite flexible and robust method. It searches for the optimal number of knots and location of knots automatically to determine the model based on data-driven correction in Bayesian term (Denison et al., 1998; Poon and Wang, 2013). Owing to the merits, a 3-degree free knot spline was utilized here to achieve better estimation for the link function.

$$g(V_i^T \boldsymbol{\alpha}_r) \approx \sum_{j=0}^{3} \theta_j^{(1)} (V_i^T \boldsymbol{\alpha}_r)^j + \sum_{l=1}^{K} \theta_l^{(2)} (V_i^T \boldsymbol{\alpha}_r - \eta_l)_+^3,$$

where K is the number of knots, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^T$ denotes the location vector of knots, K and $\boldsymbol{\eta}$ are unknown parameters that are randomized as random variables, the function $u_+ = \max(0, u)$, and $\theta_j^{(1)}$, $\theta_l^{(2)}$ are the coefficients of the spline. By the methodology, model (2) could be approximated as

$$Energy\_Consmp_i \approx \sum_{j=0}^{3} \theta_j^{(1)} (V_i^T \boldsymbol{\alpha}_r)^j + \sum_{l=1}^{K} \theta_l^{(2)} (V_i^T \boldsymbol{\alpha}_r - \eta_l)_+^3 + \varepsilon_i, \tag{3}$$

and it can be rewritten as a linear combination: $Energy\_Consmp_i \approx \mathbf{X}_i^T \mathbf{B} + \varepsilon_i$, where $\mathbf{X}_i = (Stand_{1_i}, Stand_{2_i}, Stand_{3_i}, Degree_{1_i}, Degree_{2_i}, 1, V_i^T \boldsymbol{\alpha}_r, \ldots, (V_i^T \boldsymbol{\alpha}_r)^3, (V_i^T \boldsymbol{\alpha}_r - \eta_1)_+^3, \ldots, (V_i^T \boldsymbol{\alpha}_r - \eta_K)_+^3)^T$, $\mathbf{B} = (B_1, \ldots, B_{q+K+L+1})^T = (\beta_{r_1}, \ldots, \beta_{r_5}, \theta_0^{(1)}, \ldots, \theta_3^{(1)}, \theta_1^{(2)}, \ldots, \theta_K^{(2)})^T$.

In order to identify the corresponding significant factors for no matter higher level of energy consumption or lower level of energy consumption, and compare the differences of function relationships across various levels, the idea of quantile regression analysis was introduced to the model. Firstly, note

Q_τ(*Energy_Consmp*) the τth quantile of *Energy_Consmp*, if it meets the cumulative probability P(*Energy_Consmp* ≤ Q_τ(*Energy_Consmp*)) = τ, where P(*U*) is the probability of the event *U*. Secondly, the ALD was designated as the error distribution to formulate quantile regression for *Energy_Consmp* that $\varepsilon_i \sim$ ALD $(0,\sigma,\tau)$, with the scalar parameter $\sigma > 0$, the skewness parameter $0 < \tau < 1$, and the location parameter $\mu = 0$. Critically, the location parameter is also the τth quantile of the random error that $Q_\tau(\varepsilon_i) = \mu = 0$. In terms of properties of ALD, the conditional τth quantile function of *Energy_Consmp*$_i$ follows

$$Q_\tau\left(Energy\_Consmp_i|X_i,V_i\right) = \mathbf{X}_i^T\mathbf{B}. \tag{4}$$

By means of the above equation, the contributing factors at the τth quantile of *Energy_Consmp* could be captured by $\mathbf{r}_\alpha(\tau)$ and $\mathbf{r}_\beta(\tau)$, with their effect being measured *via* $\boldsymbol{\alpha}_r(\tau)$, $\boldsymbol{\beta}_r(\tau)$. Correspondingly, the nonparametric link function $g_\tau(.)$ could be determined by K_τ, $\boldsymbol{\eta}_\tau$ and $\mathbf{B}_\tau$. For brevity purpose, the notation τ in the parameters will be omitted hereafter. Hence, given an interested quantile τ, the semi-parametric quantile regression function of *Energy_Consmp*$_i$ is finally represented by $\mathbf{X}_i^T\mathbf{B}$, and the aim of the study is to achieve it by estimating the parameters $\mathbf{r}_\alpha$, $\mathbf{r}_\beta$, $\boldsymbol{\alpha}_r$, K, $\boldsymbol{\eta}$ and $\mathbf{B}$.

To characterize the energy consumption models at lower, median and higher consumption levels, the quantiles τ = 0.2, 0.5, 0.8 were chosen respectively to investigate. Targeting at the parameters mentioned above, the posterior inference for them proceeded by the prior distributions and the data information. The reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm (Green, 1995; Lindstrom, 2002; Yu, 2002) was employed in the model to address the estimation of K. Meanwhile, the posterior update for $\mathbf{B}$, w, σ and $\boldsymbol{\xi}$ was implemented *via* Gibbs sampler, while $\mathbf{r}_\alpha$, $\mathbf{r}_\beta$ and $\boldsymbol{\alpha}_r$ were modified through Metropolis-Hastings (M-H) algorithm. The posterior estimates for them were extracted from the Markov chain of the parameters which was generated according to the above.

1.3. *Data description*

The dataset in the study was derived from the daily records of a municipal wastewater treatment plant, and there added up to 363 samples (from 25th December, 2015 to 24th December, 2016) after removing 3 invalid samples.

The descriptive statistics for *Energy_Consmp* and influencing factors are reported in Table 1. Comparing the median Q_2 with the minimum and maximum, it can be found that the probability distributions of *Energy_Consmp*, *BOD*, *COD* and *SS* tend to be right-skewed with fat right tail, and all of them suffer outliers, which indicates that there might exist severe water pollution incidents during the period of the records. In contrast, both *pH* and *Scale* vary upon quite small dispersion without extreme values. In addition, the strange phenomenon on *Chrom* that all the quartiles Q_1, Q_2, Q_3 equal to 16 mainly because *Chrom* just contains three different values {8, 16, 32} in the dataset, and 16 accounts for a relatively larger proportion. For further illustration, the histogram and probability density graph of *Energy_Consmp* are presented in Fig. 1. The sharp peak and heavy right tail of the density graph intuitively corroborate the conclusion in Table 1.

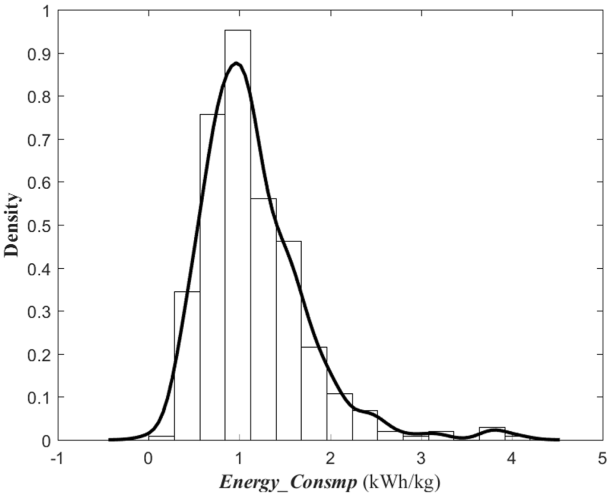

Fig. 1 – The histogram and probability density graph of *Energy_Consmp*.

Table 2 displays the Spearman rank correlation matrix among *Energy_Consmp* and influencing factors. It can be observed that *pH* and *Scale* are nearly uncorrelated with other influencing factors, while *BOD*, *COD*, *TP*, *TN* and *NH*$_3$*-N* are highly correlated with each other, especially for the groups of {*BOD*, *COD*} and {*TN*, *NH*$_3$*-N*}, the information between the two factors could be seemed as replaceable. Except for the weak correlation with *pH* and *Scale*, *Energy_Consmp* appears to be negatively correlated with all of the other influencing factors, particularly for *BOD* and *COD*. The strongly negative

Table 1 – Descriptive statistics of Energy Consmp and influencing factors.

	Energy_Consmp	*pH*	*BOD*	*COD*	*SS*	*Chrom*	*TP*	*TN*	*NH*$_3$*-N*	*Scale*
Mean	1.16	7.38	88.01	211.41	161.40	17.04	3.71	33.28	24.55	7.64
SD	0.58	0.13	46.68	103.71	111.88	6.11	1.15	8.91	7.04	0.82
Min	0.01	7.00	12.00	50.00	46.00	8.00	1.21	13.30	7.11	0.00
Q_1	0.77	7.30	57.00	142.00	96.00	16.00	2.96	26.40	19.20	7.34
Q_2	1.06	7.40	79.00	187.00	131.00	16.00	3.70	34.10	24.80	7.63
Q_3	1.44	7.50	109.50	250.50	197.00	16.00	4.32	39.70	29.65	7.95
Max	4.08	7.70	354.00	751.00	990.00	32.00	9.10	54.10	41.70	9.52

SD: standard deviation; Min: minimum; Max: maximum; Q_1, Q_2, Q_3: the quartiles, *i.e.*, the 25%, 50% and 75% quantiles respectively. *Energy_Consmp*: energy consumption; *pH*: influent pH; *BOD*: influent biochemical oxygen demand concentration; *COD*: influent chemical oxygen demand concentration; SS: influent suspend solid concentration; *Chrom*: influent chroma; *TP*: influent total phosphorus concentration; *TN*: influent total nitrogen concentration; *NH*$_3$*-N*: influent NH_3-N concentration; *Scale*: treatment scale per day.

Table 2 – Spearman rank correlation matrix of *Energy_Consmp* and influencing factors.

Correlation	*pH*	*BOD*	*COD*	*SS*	*Chrom*	*TP*	*TN*	*NH$_3$-N*	*Scale*	*Energy_Consmp*
pH	1.000	−0.051	−0.024	0.061	−0.045	0.093	−0.064	−0.089	−0.226	0.054
BOD	−0.051	1.000	0.892	0.505	0.294	0.649	0.609	0.486	−0.039	−0.867
COD	−0.024	0.892	1.000	0.566	0.353	0.714	0.629	0.484	−0.041	−0.962
SS	0.061	0.505	0.566	1.000	0.373	0.493	0.227	0.064	−0.020	−0.533
Chrom	−0.045	0.294	0.353	0.373	1.000	0.362	0.176	0.085	0.086	−0.327
TP	0.093	0.649	0.714	0.493	0.362	1.000	0.727	0.625	−0.133	−0.638
TN	−0.064	0.609	0.629	0.227	0.176	0.727	1.000	0.893	−0.098	−0.554
NH$_3$-N	−0.089	0.486	0.484	0.064	0.085	0.625	0.893	1.000	−0.073	−0.414
Scale	−0.226	−0.039	−0.041	−0.020	0.086	−0.133	−0.098	−0.073	1.000	−0.044
Energy_Consmp	0.054	−0.867	−0.962	−0.533	−0.327	−0.638	−0.554	−0.414	−0.044	1.000

BOD: influent biochemical oxygen demand concentration; *COD*: influent chemical oxygen demand concentration; *SS*: influent suspend solid concentration; *Chrom*: influent chroma; *TP*: influent total phosphorus concentration; *TN*: influent total nitrogen concentration; *NH$_3$-N*: influent NH$_3$-N concentration; *Scale*: treatment scale per day.

correlation between *Energy_Consmp* and *COD* is straightforward that *Energy_Consmp* measures the energy consumption in unit COD reduction, therefore the higher the *COD*, the lower the *Energy_Consmp*.

In the dataset, several missing values occurred in *COD* and SS, which might cause information loss and undermine the inference. The imputed data for missing values was generated from the normal distributions $N(\overline{COD}, \sigma^2_{COD})$ and $N(\overline{SS}, \sigma^2_{SS})$, where $\overline{\delta}$, σ^2_δ signify the mean and variance of δ accordingly. At last, *Energy_Consmp* and all the influencing factors were standardized to have zero mean and unit deviation for the purpose of comparing the significance of influencing factors in a uniform standard.

2. Results and discussion

For different energy consumption levels, the truly contributing factors and their effect on *Energy_Consmp* were captured by virtue of the Bayesian semi-parametric quantile regression approach, and the proposals for energy saving *via* adjusting the key factors were put forward correspondingly.

For practical application purpose, given current value of *Energy_Consmp*, its energy consumption level could be roughly judged according to the following quantiles. Taking the dataset of the study as an example, the *Energy_Consmp* that less than 0.73, *i.e.*, the 0.2-quantile of $\{Energy\_Consmp_i\}_{i=1}^{n}$, could be considered as the lower level of energy consumption, while the *Energy_Consmp* around 1.06 is approximately the median level of energy consumption. Likewise, the *Energy_Consmp* that larger than 1.54, that is, the 0.8-quantile of $\{Energy\_Consmp_i\}_{i=1}^{n}$, could symbolize the higher level of energy consumption. In practical application, the current energy consumption level can be judged by the above thresholds, and the energy consumption might decrease *via* adjusting key factors on the corresponding level.

Explicitly, in Model (2), *X* and *V* could be considered as the two branches of influencing factors. For the first branch, *Stand* and *Degree* exert linear effect on *Energy_Consmp*, and the effect could be directly reflected *via* β_{r_l}, while in another branch, influencing factors affect the *Energy_Consmp* nonlinearly, with the effect being analyzed in terms of both α_{r_j} and $g'(V^T\boldsymbol{\alpha}_r)$, that is, the partial derivative $\partial Energy\_Consmp/\partial V_j = \alpha_{r_j} g'(V^T\boldsymbol{\alpha}_r)$, $j = 1, \ldots, 9$, $l = 1, \ldots, 5$. Accordingly, the posterior quantile estimates $\hat{\boldsymbol{\alpha}}_r$ and $\hat{\boldsymbol{\beta}}_r$ are summarized in Table 3, and the estimated curve of link function $\hat{g}(V^T\boldsymbol{\alpha}_r)$ at quantiles $\tau = 0.2$, 0.5, 0.8 are outlined in Figs. 2-4 respectively. The small standard deviations in Table 3 imply the stable estimates and well estimation efficiency.

2.1. Regression analysis at lower energy consumption level

At the quantile $\tau = 0.2$, the lower level of energy consumption is closely related to *Degree*$_1$ and *Chrom*. In detail, *Degree*$_1$ shows a negative impact on *Energy_Consmp* as reported in Table 3, which demonstrates that the *Energy_Consmp* with the normal

Table 3 – Bayesian quantile estimates for $\boldsymbol{\alpha}_r$, $\boldsymbol{\beta}_r$. Standard deviations are reported in parentheses.

	$\hat{\alpha}_{r_1}$ *pH*	$\hat{\alpha}_{r_2}$ BOD	$\hat{\alpha}_{r_3}$ COD	$\hat{\alpha}_{r_4}$ SS	$\hat{\alpha}_{r_5}$ *Chrom*	$\hat{\alpha}_{r_6}$ TP	$\hat{\alpha}_{r_7}$ TN	$\hat{\alpha}_{r_8}$ NH$_3$-N	$\hat{\alpha}_{r_9}$ *Scale*	$\hat{\beta}_{r_1}$ *Stand*$_1$	$\hat{\beta}_{r_2}$ *Stand*$_2$	$\hat{\beta}_{r_3}$ *Stand*$_3$	$\hat{\beta}_{r_4}$ *Degree*$_1$	$\hat{\beta}_{r_5}$ *Degree*$_2$
$\tau = 0.2$	0	0	0	0	1	0	0	0	0	0	0	0	−0.1660 (0.0572)	0
$\tau = 0.5$	0	0	1	0	0	0	0	0	0	0	0.0933 (0.0248)	0.1425 (0.0496)	0	0
$\tau = 0.8$	0	0	0	0	0	0	1	0	0	0	0	0	0	0.2002 (0.0610)

BOD: influent biochemical oxygen demand concentration; *COD*: influent chemical oxygen demand concentration; *SS*: influent suspend solid concentration; *Chrom*: influent chroma; *TP*: influent total phosphorus concentration; *TN*: influent total nitrogen concentration; *NH$_3$-N*: influent NH$_3$-N concentration; *Scale*: treatment scale per day.

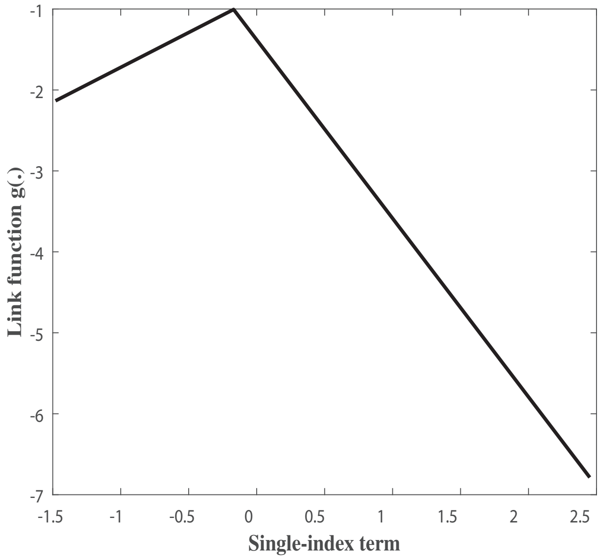

Fig. 2 – Graph of the link function about single-index term at the 0.2-quantile.

wastewater temperature tends to be less than that with lower wastewater temperature. The phenomenon accords with the principle that the proper temperature stimulates the reaction of microorganism and therefore alleviates the burden of energy consumption. Meanwhile, *Stand* is not active in the model, and the inactiveness manifests that the quality of treated water might have no obvious effect on the change of *Energy_Consmp* in this case.

For the second branch of influencing factors, incorporating $\hat{\boldsymbol{\alpha}}_r$ in Table 3 and $\hat{g}(V^T\boldsymbol{\alpha}_r)$ in Fig. 2, it can be found that *Chrom* is the unique significant factor, and *Energy_Consmp* goes up with it at first, then goes down greatly. The trend signifies that enhancing the *Chrom*-rich wastewater (*e.g.*, dyeing mill effluent or paper mill effluent) might help reduce the energy consumption. Noteworthily, the graph of $\hat{g}(V^T\boldsymbol{\alpha}_r)$ is a folding line with one sharp point rather than a smooth curve, the strange performance is mainly related to the fact that *Chrom* has only three different values in the dataset, and hence $\{(V_i^T\boldsymbol{\alpha}_r, g(V_i^T\boldsymbol{\alpha}_r))\}_{i=1}^{n}$ just has three different pairs of points in the figure.

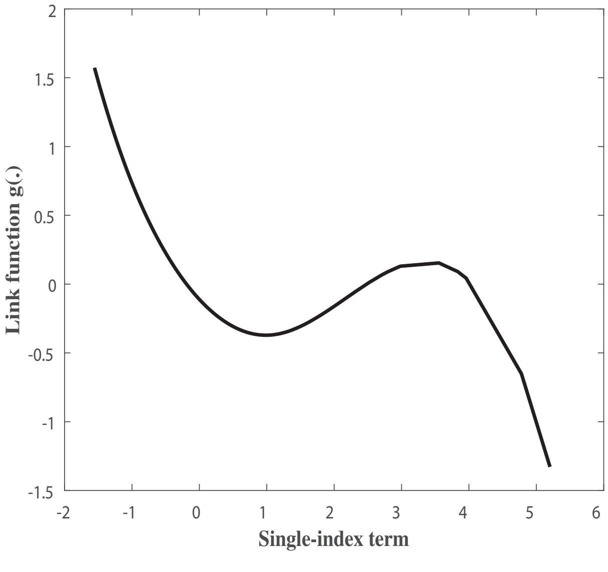

Fig. 3 – Graph of the link function about single-index term at median level.

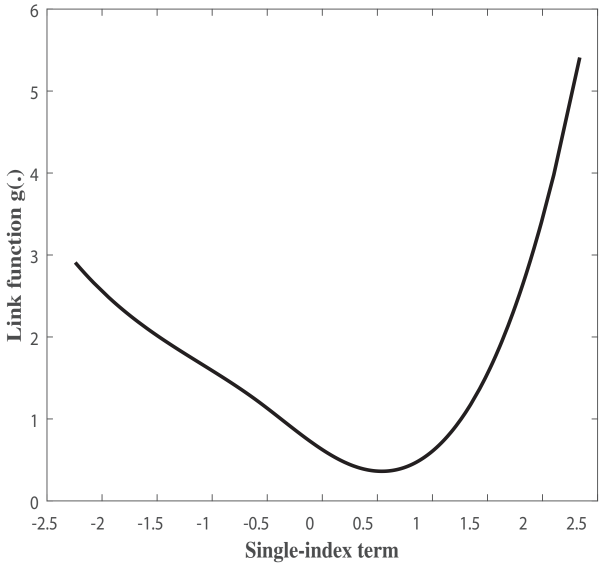

Fig. 4 – Graph of the link function about single-index term at 0.8-quantile.

2.2. Regression analysis at median energy consumption level

At median energy consumption level, it can be seen in Table 3 that $Stand_2$, $Stand_3$ are positively contributing to *Energy_Consmp*, and $Stand_3$ exerts slightly larger effect. The result meets the expectation that the higher quality of treated water demands for longer hydraulic retention time, and *Energy_Consmp* increases consequently. One possible reason for the insignificance of $Stand_1$ is that the samples of $Stand_1$ are far less than $Stand_2$ and $Stand_3$ in the dataset so that the impact of it is not evident.

In another branch, *COD* plays a prominent role to affect *Energy_Consmp*, and as is shown in Fig. 3, the effect of it is basically negative except for a flat part during the decline. The reasonability of the result could be confirmed through the analyses in Table 2 that *COD* shows high correlation with most of the influencing factors and could represent principal information of them, furthermore, *COD* is negatively correlated with *Energy_Consmp* on the basis of Spearman rank correlation. Compared with the analyses in Table 2, the proposed approach depicts a more detailed and more clear trend for the effect of *COD* on *Energy_Consmp*. In addition, the result also implies that the *COD*-rich wastewater (*e.g.*, starch mill effluent, dyeing mill effluent, industrial effluent and bean products effluent) tends to cut down *Energy_Consmp*, or at least it would not aggravate the burden of energy consumption.

2.3. Regression analysis at higher energy consumption level

At the quantile $\tau = 0.8$, the higher level of energy consumption most likely results from the factors $Degree_2$ and *TN*. Detailedly, $Degree_2$ is apt to increase the *Energy_Consmp*, with the indication that the relatively high temperature may hinder the reaction of microorganism and lead to the rise of *Energy_Consmp*. Hence, keeping appropriate water temperature is crucial in wastewater treatment.

For another significant factor, *TN* is similar to *COD* as reflected in Table 2 that it covers much information of other influencing factors. However, different from the explicitly negative correlation in Table 2, the effect of *TN* on *Energy_Consmp* is actually complex rather than simply negative. Specifically, in light of the $\hat{g}(V^{\mathrm{T}}\boldsymbol{\alpha}_r)$ in Fig. 4, *Energy_Consmp* decreases with *TN* at first, then it begins to increase after *TN* amounting to a certain value, which illustrates that the *TN*-rich wastewater (*e.g.*, ammonia plant effluent, pesticide wastewater) cannot guarantee the reduction of *Energy_Consmp*, on the contrary, *TN* needs to be controlled within a rational range to avoid the case of energy intensive. The outcome also embodies the comprehensive perspective of the proposed approach to analyze the effect of influencing factors.

Note that *pH* and *Scale* perform inactively in all the three energy consumption levels. On one hand, it is consistent with the report in Table 2 that both of them appears uncorrelated with *Energy_Consmp* and would not be responsible for the variation of *Energy_Consmp*. On the other hand, a more likely reason, the dispersion of *pH* and *Scale* in the dataset is quite small so that the samples cannot represent the population, and *Energy_Consmp* is consequently not sensitive to them.

Concluding from the above analyses, the normal temperature of influent wastewater is important to the lower level of energy consumption. And for median level of energy consumption, it is proposed to increase the *COD*-rich wastewater, while for higher level of energy consumption, the *TN*-rich wastewater is preferably to be controlled.

3. Conclusions

A Bayesian semi-parametric quantile regression approach was proposed in the paper to address the energy consumption modeling in biochemical wastewater treatment. The energy consumption model was fitted using a semi-parametric PLSIM, and the nonparametric link function was approximated *via* free knot spline. By means of the binary indicator variables, the contributing factors that truly affect the energy consumption were identified, and with the ALD distributed random errors, the models for lower, median and higher levels of energy consumption were formulated respectively.

Based on the results, *Chrom* and $Degree_1$ acted as the determinative factors for the lower level of energy consumption, and the normal temperature of wastewater contributed to the reduction of energy consumption, while *Chrom*-rich wastewater also appeared helpful. At median consumption level, the higher quality of treated water led to the increase of energy consumption, while *COD*-rich wastewater tended to cut down it. The *TN*-rich wastewater and the relatively high temperature of wastewater were the major causes of higher level of energy consumption, and it was necessary to control the *TN*-rich wastewater and lower the water temperature to avoid excessive energy consumption.

It is noteworthy that the proposed approach was illustrated *via* a specific treatment plant, and therefore the analyses and suggestions in Section 2 mainly targeted at that plant and cannot be referred as a general conclusion. Similarly, the approach could be applied to other interested treatment plants, and the proposals for energy saving can be summarized after identifying the significant factors and their effect on energy consumption. Conclusively, the proposed approach facilitated systematic and robust energy consumption models to guide the energy saving in biochemical wastewater treatment.

Acknowledgment

The research was supported by the National Natural Science Foundation of China (Nos. 51478025, 11701023, 71420107025).

REFERENCES

Araki, T., Ikeda, K., Akaho, S., 2015. An efficient sampling algorithm with adaptations for Bayesian variable selection. Neural Netw. 61, 22–31.

Bai, T.X., 2012. The Energy Consumption of Municipal Sewage Treatment Plants and Energy Saving Analysis. (Master Degree thesis). Chang'an University, Xi'an, China.

Boente, G., Rodriguez, D., 2012. Robust estimates in generalized partially linear single-index models. TEST 21 (2), 386–411.

Carroll, R.J., Fan, J.Q., Gijbels, I., Wand, M.P., 1997. Generalized partially linear single-index models. J. Am. Stat. Assoc. 92 (438), 477–489.

Denison, D.G.T., Mallick, B.K., Smith, A.F.M., 1998. Automatic Bayesian curve fitting. J. R. Stat. Soc. B 60 (2), 333–350.

Gao, D.W., Li, Z., Guan, J.X., Liang, H., 2017. Seasonal variations in the concentration and removal of nonylphenol ethoxylates from the wastewater of a sewage treatment plant. J. Environ. Sci. 54 (4), 217–223.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732.

Han, C.F., Liu, J.X., Liang, H.W., Guo, X.S., Li, L., 2013. An innovative integrated system utilizing solar energy as power for the treatment of decentralized wastewater. J. Environ. Sci. 25 (2), 274–279.

Han, H.G., Lu, Z., Qiao, J.F., 2016. An energy consumption model of wastewater treatment process based on adaptive regressive kernel function. CIESC J. 67 (3), 947–953.

Huang, X.Q., Han, H.G., Qiao, J.F., 2013. Energy consumption model for wastewater treatment process control. Water Sci. Technol. 67 (3), 667–674.

Jiang, Y., Fu, W., Mao, L.H., Ren, F.M., Yang, L., Xiang, J., Liang, R., Hao, H.M., Wang, Z., 2014. Influence factors analysis of urban sewage treatment plant on energy consumption. J. Beijing Jiaotong Univ. 38 (1), 33–37.

Jin, W.J., Yang, D.D., 2012. Methods for analyzing energy consumption in wastewater treatment plants: review. Environ. Prot. Technol. 18 (2), 18–21.

Jin, C.Q., Wang, C.W., Zeng, S.Y., He, M., 2009. Characteristics analysis and energy saving research of energy consumption in wastewater treatment plant. Water Wastewater Eng. 35 (s1), 270–274.

Jin, W.J., Yang, P.W., Cong, B.B., Li, D.W., 2014. Research on energy consumption prediction model for biochemical pool based on neural network. Environ. Eng. 32 (s), 961–963.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Koenker, R., Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. 94 (448), 1296–1310.

Kusiak, A., Zeng, Y.H., Zhang, Z.J., 2013. Modeling and analysis of pumps in a wastewater treatment plant: a data-mining approach. Eng. Appl. Artif. Intell. 26 (7), 1643–1651.

Li, W., 2010. Research on Evaluation System of Energy Consumption for Municipal Wastewater Treatment Plant. (PhD thesis). Xi'an University of Architecture and Technology, Xi'an, China.

Liang, R., 2014. Urban sewage treatment plant energy consumption influence factors research. (Master Degree thesis). Beijing Jiaotong University, Beijing, China.

Liang, H., Liu, X., Li, R.Z., Tsai, C.L., 2010. Estimation and testing for partially linear single-index models. Ann. Stat. 38 (6), 3811–3836.

Lindstrom, M.J., 2002. Bayesian estimation of free-knot splines using reversible jumps. Comput. Stat. Data Anal. 41 (2), 255–269.

O'Hara, R.B., Sillanpaa, M.J., 2009. A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. 4 (1), 85–117.

Poon, W.Y., Wang, H.B., 2013. Bayesian analysis of generalized partially linear single-index models. Comput. Stat. Data Anal. 68, 251–261.

Ren, F.M., Mao, L.H., Fu, W., Yang, L., Meng, Y., Wang, Z., Liang, R., Xiang, J., Hao, H.M., 2015. Study of influent factors on energy consumption of municipal wastewater treatment plant operation in China. Water Wastewater Eng. 41 (1), 42–47.

Wang, X., Li, M.Y., Liu, J.X., Qu, J.H., 2016. Occurrence, distribution, and potential influencing factors of sewage sludge components derived from nine full-scale wastewater treatment plants of Beijing, China. J. Environ. Sci. 45, 233–239.

Yang, L.B., Zeng, S.Y., Ju, Y.P., He, M., Chen, J.N., 2008. Statistical analysis and quantitive identification of energy consumption in municipal wastewater treatment plant. Water Wastewater Eng. 34 (10), 42–45.

Yi, X.N., Fan, Y.H., Hu, C.F., Li, C.G., 2009. Study on bio-ecological combination process with high efficiency and low energy consumption for municipal wastewater treatment. International Conference on Bioinformatics and Biomedical Engineering. Beijing, China. June 11–13.

Yu, K.M., 2002. Quantile regression using RJMCMC algorithm. Comput. Stat. Data Anal. 40 (2), 303–315.

Yu, K.M., Moyeed, R.A., 2001. Bayesian quantile regression. Stat. Probab. Lett. 54 (4), 437–447.