CHEMICAL ENGINEERING TRANSACTIONS Volume 17, 2009
Editor Sauro Pierucci
 ISBN 978-88-95608-01-3 ISSN 1974-9791
DOI: 10.3303/CET0917020

Extending model prediction ability for the formation of nitrophenols in benzene nitration

Paula A. G. Portugal[1,2*] Marco S. Reis[1], Cristina M. S. G. Baptista[1]
1CIEPQPF, Chemical Engineering Department, University of Coimbra, Pólo II, Rua Sílvio Lima, 3030–790 Coimbra, Portugal.
2Chemical Engineering and Environment Department, High School of Technology of Tomar, Polytechnics Institute of Tomar, Quinta do Contador, Estrada da Serra, 2300-313 Tomar, Portugal.

In the industrial adiabatic benzene nitration process, the by-products dinitrophenol (DNP) and trinitrophenol (TNP) have a great contribution to the environmental footprint of industrial plants. Therefore, nitrophenols (NPs) reduction is a relevant target in process operations. In this context, Quadros et al. (2005) studied the benzene nitration in a pilot plant and developed both mechanistic as well as empirical multivariate linear regression (MLR) models for the formation of mononitrobenzene (MNB) and NPs. Nevertheless, those MLR models are not able to predict the global performance in a set of continuous adiabatic reactors in series. To overcome this, in the present work new models were developed by a different approach which includes in the set of regressors the CSTR outlet state variables. In order to develop the MLR models, the sequential forward stepwise regression method was applied. The extended models thus obtained, confirm that the selectivity and conversion of these liquid-liquid reactions can be optimized, as they depend significantly on reaction temperature, residence time and interfacial area as well as on the concentrations of some of the chemical compounds involved in this process.

1. Introduction

Benzene nitration with mixed acid is an important industrial chemical process involving liquid-liquid reactions, usually carried out adiabatically in a series of CSTRs, whose degree of conversion to MNB and selectivity are a function of a wide range of parameters. There are several by-products that can be formed depending on the process variables set-points. The most important by-products formed are the nitrophenols (NPs), namely dinitrophenol (DNP) and trinitrophenol (TNP). Their concentration in the outlet stream can be in the range of 3000 to 5000 ppm, and these compounds cannot be disposed without a specific treatment. Therefore, improving process selectivity is a high priority goal. Addressing this problem, Quadros et al. (2005) studied benzene nitration in a pilot plant under operating conditions in the range of those used in industrial

* Corresponding author. Tel + 351-239-798793. E-mail:pagp@eq.uc.pt

Please cite this article as: Portugal P., Reis M. and Baptista C., (2009), Extending model prediction ability for the formation of nitrophenols in benzene nitration, Chemical Engineering Transactions, 17, 117-122 DOI: 10.3303/CET0917020

practice. They developed both mechanistic and empirical (multivariate linear regression, MLR) models for the formation of MNB and NPs. As explanatory variables for the empirical MLR models, several operating parameters were used, including some of the reactor inlet conditions: *a*; T_o; NA_i; SA_i; θ, $(F_B/F_N)_i$. The models obtained exhibit good prediction ability. In the industrial adiabatic process in CSTRs a high conversion is expected in the first reactor (Alexanderson et al. 1978), where Quadros et al. (2005) models can be accurately used. However, the inlet conditions of the subsequent reactors are expected to be quite distant from those in the first one, which means that the models are no longer valid for those reactors. In order to extend the prediction ability of the MLR models, using the same data, new models are here developed, by replacing the inlet variables in the set of the explanatory regressors, with outlet variables. By doing so, the set of regressors becomes the following one: *a*; T_o; NA_o; SA_o; θ, $(F_B/F_N)_o$. Additionally, DNP was considered as a regressor in TNP models since Burns and Ramshaw, 1999 support that TNP is produced by nitration of DNP.

2. Building the MLR models

The models were established by resorting to the stepwise regression methodology (SR), implemented in Statistica 6.1 (StatSoft). Variable transformations, such as log x, x^2 and $\sqrt{x}$, were considered as additional regressors, along with the original untransformed variables (x), in order to model the process non-linearity, but using the formalism of linear regression (models are linear regarding to parameters, but non-linear regarding to variables). The SR methodology (Montegomery and Runger, 2007) consists of incrementing, step by step, the number of explanatory regressors, starting by a one-variable model using the regressor variable that has the highest correlation with the response variable y. Then, at each step, another regressor is added, if it is considered significant, according to the partial F-test, and the set of previously selected variables are analyzed in order to check whether any of them should be disregarded. Regression analysis requires the residuals in the estimated regression model to be normally and independently distributed with mean zero and constant variance, σ^2, which must be verified through residual analysis, namely using the normal probability plot of residuals and the plot of residuals against y or $\hat{y}$, allowing the verifications of residuals homoscedasticity and independence. Several statistical tests are relevant in the analysis of the estimated models. The test for significance of the regression model as a whole, is a ANOVA test to determine whether a significant linear relationship exists between the response variable y and a subset of regressor variables considered. Tests on individual regression coefficients, called partial or marginal tests (Montegomery and Runger, 2007), are useful in determining the potential value of each regressor in the model, and they are based on a student-t statistic. Significance is usually assessed by calculating the p-level associated with the statistical tests. The p-level represents the probability of obtaining a deviation at least as large as the one observed, assuming the null hypothesis to be correct. In the case of the tests mentioned so far, the higher the p-level, the less the parameter significance. A p-level of 0.05 is customarily treated as a border-line acceptable error level.

We should point out that SR does not avoid completely collinearities in the regressors selected, something to be done by other means. Multicollinearity cannot be seen as a

modeling error itself, but as a condition regarding collected data. However, it can have serious effects on the estimates of the regression coefficients and on general applicability of the estimated model. A parameter that enables the quantification of multicollinearity is the variance inflation factor (VIF), defined by (Montegomery and Runger, 2007):

$$VIF(\beta_j) = \frac{1}{1 - R_j^2} \qquad j = 1,2,...k \tag{1}$$

where R_j^2 is the coefficient of multiple determination resulting from regressing x_j on the other k-1 regressor variables. The larger the VIF, the more severe the effect of multicollinearity might be. Some authors support that, if any VIF exceeds 10, multicollinearity becomes a real problem. Another clue for the existence of multicollinearity occurs when the F-test for significance of regression is significant, but tests on the individual regression coefficients result in not significant scores for some variables.

Several criteria may be used for evaluating and comparing the different regression models obtained (Chatterjee et al, 1991). A commonly used criterion is based on the value of the multiple correlation coefficient, R^2, or the adjusted R-squared, $Radj^2$. Both are measures of the model fitting quality and have unity as maximum value (for perfect models), but only $Radj^2$ introduces penalties linked to the number of variables considered in the model, a way to guard against over-fitting. In fact, $Radj^2$ will often stabilize and actually begin to decrease as the number of non-significant variables in the model increases. Usually, a model that maximizes $Radj^2$ is considered to be a good candidate for the best regression equation. The root mean square error (RMSE), equation 2, and the root mean square relative error (RMSRE), equation 3, were also measures considered for assessing and comparing the MLR models.

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \Big/ (n - p)} \tag{2}$$

$$RMRSE = \sqrt{\sum_{i=1}^{n} \left(\left(y_i - \hat{y}_i \right) \Big/ y_i \right)^2 \Big/ (n - p)} \times 100 \tag{3}$$

As to prediction assessment, an approach consists in splitting the observations into two sets: a training set (TRS) and a test set (TSS). The regression parameters are recalculated using only the TRS observations (e.g., by random selection of 80% of the original observations). All the parameters that enable to check the model adequacy are recalculated and compared with the complete model. The TRS model is subsequently used to predict y values, $\hat{y}$, for the TSS observations. The quality of the prevision may be done by calculating the RMSE and the RMSRE for the estimated TSS values and compare with their counterparts for the TRS model and complete model (RMSEP and RMSREP), that differ from RMSE and RMSRE only in the denominator, which is now

simply n. A good prevision ability, coherent in all cases, indicates that the estimated model is stable and robust, useful features in practice.

When using multiple regression one also occasionally finds that some subset of observations is unusually influential. Sometimes these influential observations, often called outliers, are relatively far away from the vicinity where the bulk of data were collected. They can be related to wrong measures or not. In this situation, the first step is to identify them, and then decide whether they should be, or not, rejected. Residual analysis helps in this activity. A model can be considered correct and precise, at a 95% confidence level, if the standardized residuals are randomly distributed about zero and confined approximately between the ±2 range. Other measures, such as the Cook's distance and the Mahalanobis distance, also indicate if an observation is influential. The values should be of about the same magnitude for all observations.

3. Results and conclusions

The procedure described above was carried out using the set of data of Quadros et al., 2005 but this time with the outlet operating conditions replacing the inlet ones as variables (Table 1). As a result, several models where estimated but, after careful analysis, only five happened to be adequate (equations 4 to 8).

DNP $\ln(DNP) = -9{,}31683 + 3{,}07843\ln(T) - 0{,}01527(F_B/F_N)_o + 0{,}1561\ln(a) - 0{,}04344 NA_o$ (4)

TNP1 $\ln(TNP) = -103{,}492 + 0{,}103T - 1{,}17\times10^{-5}a + 0{,}116(NA_o)^2 + 0{,}725\ln(NA_o) + 23{,}045\ln(SA_o) + 0{,}476\ln\theta$ (5)

TNP2 $\ln(TNP) = -94{,}4123 + 0{,}108T + 0{,}9978(NA_o) - 0{,}2236\ln(F_B/F_N)_o + 20{,}4036\ln(SA_o) + 0{,}6249\ln\theta$ (6)

TNP3 $\ln(TNP) = -11{,}66 + 4\times10^{-4}T^2 - 4{,}6\times10^{-3}\sqrt{a} + 0{,}0988(NA_o)^2 + 0{,}6711\ln(NAo) + 2{,}4\times10^{-3}(SA_o)^2 + 0{,}5762\ln\theta + 0{,}0493\sqrt{DNP}$ (7)

TNP4 $\ln(TNP) = -9{,}72948 + 3{,}9\times10^{-4}T^2 - 0{,}01059\sqrt{a} + 0{,}74229\ln(NA_o) + 2{,}26\times10^{-3}(SA_o)^2 + 0{,}62699\ln\theta + 0{,}03595\sqrt{DNP}$ (8)

Table 2 presents some of the several statistical parameters calculated for models adequacy checks. It is very clear that all the models show good data fitting ability, with excellent scores for $Radj^2$. An adequate prediction ability can also be inferred from the agreement between RMSRE values

Table 1 – Range of the variables used in the MLR models.

Variable	Quadros et al. (2005)	Present work
T	81 – 135	81 – 135
$(F_B/F_N)_i$	0.93 – 1.5	not used
$(F_B/F_N)_o$	not used	0.3 – 16
a	580 – 58 000	580 – 58 000
θ	1.9 – 6.1	1.9 – 6.1
NA_i	2.6 – 6.4	not used
SA_i	57 – 69	not used
NA_o	not used	0.06 – 4.5
SA_o	not used	58 – 72
DNF	138 – 1500	138 – 1500
TNF	0 – 1211	0 – 1211

In Figure 1 we can also confirm the good prediction ability of the DNP model, with most of the points lying very close to the line ŷ=y, and the same results can be extended to the TNP models.

The four models obtained for TNP are statistically valid and therefore are all shown here. Model TNP2 (equation 6) has the advantage of only using operating conditions that are measured in the process, therefore not requiring the estimation of other parameters.

The DNP and TNP model results for the outlet stream of a CSTR were compared to those obtained with the models in Quadros et al., 2005 showing good agreement. The advantages of the models presented here stand out when the models are used to calculate DNP and TNP weight fractions in continuous reactors in series. In fact, only the models in equations 4 to 8 are adequate as the concentration values are, in this case, inside the ranges of data used to estimate the models.

The present work constitutes a contribution to increasing the understanding of the MNB production process in industrial operating conditions, while extending the DNP and TNP concentration prediction ability. Among several applications, the models developed can be combined with other models for estimating the output variables (now set as regressors), allowing the prediction of by-products formation in CSTR´s in series.

Table 2 –Statistical parameters for Quadros et al., 2005 models and the extended MLR models

	Quadros et al, 2005		Extended MLR models				
	DNP	TNP	DNP	TNP1	TNP2	TNP3	TNP4
N	139	130	147	132	135	134	135
$Radj^2$	0,9738	0,9569	0,9533	0,9773	0,9757	0,9786	0,9661
RMSRE (%) (complete)	6,65	--	1,61	3,47	3,77	3,30	4,56
RMSRE (%) (TRS)	6,76	16,53	1,59	3,49	3,62	3,23	4,45
RMSRE (%) (TSS)	5,37	13,1	1,73	3,89	4,28	4,44	4,96

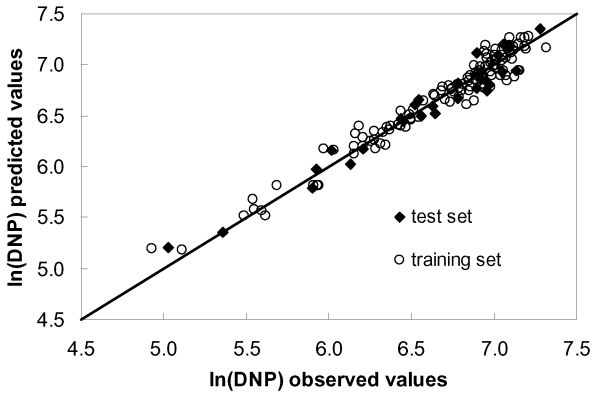

Figure 1 – Predicted versus observed values for the ln(DNP) model: results for the training and test data sets.

Notations

a	interfacial area between the organic and acid phases (m^2/m^3)
DNP	DNP concentration in the outlet organic stream (ppm)
F_B/F_N	ratio of the inlet molar flow rates of benzene and nitric acid
k	number of regressors in the model
n	number of observations (experiments)
NA	nitric acid inlet weight fraction in the acid phase (wt %)
p	number of parameters in the model
SA	sulfuric acid inlet weight fraction in the acid phase (wt %)
T	temperature in the reactor (ºC)
TNP	TNP concentration in the outlet organic stream (ppm)

Subscripts		*Greek*	
i	reactor inlet	θ	residence time (min)
o	reactor outlet	β	model parameter

References

Alexanderson, V., Trecek, J. B. and Vanderwaart, C. M., 1978, Continuous adiabatic process for the mononitration of benzene, US Patent 4,091,042.

Burns, J. R. and Ramshaw, C., 1999, Development of a micro reactor for chemical production, Trans IChemE, 77, Part A, 206-211.

Chatterjee, S. and Price, B., 1991, Regression analysis by example. John Wiley & Sons, 2nd ed., New York.

Montgomery, D. and Runger, G., 2007, Applied Statistics and Probability for Engineers. 4th edition John Wiley and Sons.

Quadros, P. A., Reis M. S. and Baptista, C. M. S. G., 2005, Different modelling approaches for a heterogeneous liquid-liquid reaction process, Ind. & Eng. Chem. Res., 44, 9414-9421.