CHEMICAL ENGINEERING TRANSACTIONS
VOL. 32, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at: www.aidic.it/cet
Chief Editors: Sauro Pierucci, Jiří J. Klemeš

**ISBN** 978-88-95608-23-5; **ISSN** 1974-9791
DOI: 10.3303/CET1332209

# Prediction of the By-products Formation in the Adiabatic Industrial Benzene Nitration Process

Anabela G. Nogueira[a,b,*], Dulce C.M. Silva[a], Marco S. Reis[b], Cristina M.S.G. Baptista[b]

[a] CUF - Químicos Industriais, Quinta da Indústria, Beduído,3860-680 Estarreja, Portugal
[b]CIEPQPF– Department of Chemical Engineering, University of Coimbra, Pólo II, Rua Sílvio Lima, 3030-790 Coimbra, Portugal
anabela.nogueira@cuf-qi.pt

Side reactions are undesirable in most industrial processes, as they decrease process yield and selectivity. For this reason, mononitrobenzene's manufacturers set nitrophenols minimization as a critical goal, along with the MNB production targets. The mechanism of these side reactions in benzene nitration is still under debate and, so far, none of the alternatives has achieved general consensus in the scientific community. As an alternative, industrial data may provide valuable information on the contribution of inlet process variables and operating conditions upon the formation of nitrophenolic compounds in the adiabatic nitration process. In this work, Partial Least Squares regression was applied to data collected from a mononitrobenzene industrial production plant. This methodology allowed concluding that nitration temperature and mixed acid volumetric flow rate as the most influential variables in nitrophenols formation. The models developed enable proper estimates of DNP and TNP concentrations in the industrial process, although their explanation power is lower than those previously obtained by Quadros et al. (2005), in a pilot plant, and by Portugal et al. (2009) in their extended models.

## 1. Introduction

The industrial adiabatic process for the manufacture of mononitrobenzene (MNB), carried out by reacting benzene (B) with nitric acid (NA) using sulfuric acid (SA) as catalyst, produces undesirable by-products, such as 2,4-dinitrophenol, trinitrophenol and dinitrobenzene, lowering process yield and selectivity. Several methods for nitrophenols (NP) treatment are patented, which fall under one of the following categories: chemical/physical processes including extraction (Hanson et al., 1976) and precipitation (Adams and Barker, 1990) and thermal processes, where high pressure and high temperature are employed (Larbig, 1980). Nitrophenols removal increases the overall process cost due to the high investment on equipment and operation costs associated with abatement procedures. Furthermore, nitrophenols elimination or disposal must comply with increasingly strict environmental regulations. In order to mitigate such negative impacts, it is essential to optimize the MNB's production process in such a way that nitrophenolic by-products formation is minimized while simultaneously achieving the MNB production targets. The knowledge of the NP's formation mechanism would be very helpful in this regard, but the complexity of mass transfer phenomena and the lack of kinetic data are hindering the development of such mechanistic models. In this context, statistical methods offer a promising and valuable alternative to infer models for the prediction of NP's formation, taking advantage of available data from normal industrial operation or from experiments, requiring no much phenomenological knowledge or thermodynamic and kinetic data. Regression analysis is a flexible and powerful approach for inferring relationships between process variables from historical databases. Different modeling approaches were applied to the benzene nitration process by Quadros et al. (2005). Applying multiple linear regression (MLR) to data collected from experiments in a pilot plant reactor, the authors identified nitration temperature (*T*), inlet molar feed ratio of benzene and nitric acid $\left(F_{m,B}/F_{m,NA}\right)$ and interfacial area (a) as the most influencing variables. This

Please cite this article as: Nogueira A., Silva D., Reis M., Baptista C., 2013, Prediction of the by-products formation in the adiabatic industrial benzene nitration process, Chemical Engineering Transactions, 32, 1249-1254 DOI: 10.3303/CET1332209

model was tested for an industrial set of nitration reactors working in series, and found valid for the first reaction, where operating conditions are similar to those for which the model was developed. In order to extend the range of operating conditions where these models can be used, Portugal et al. (2009) derived a new version of this prediction model, that includes a different set of input variables. Nevertheless, when applying to industrial conditions, both models present a very limited ability to predict NP's formation. Therefore, the aim of this work is to build a predictive model for the benzene nitration by-products based on real industrial data set.

## 2. Statistical Analysis

Industrial variables, typically present a large amount of correlation, as a consequence of the conservation laws, control loops, redundant instrumentation, etc. In this conditions, MLR models such as those used by Quadros et al. (2005) become highly unstable and alternative methodologies must be used. Several methods, such as Principal Components Regression (PCR) and Partial Least Squares Regression (PLS), are capable of overcoming this problem by taking the variables multicollinearity into account (Martens and Næs, 1989).

PLS regression is a multivariate algorithm that converts correlated variables into linear combinations with predictive power regarding the response. These linear combinations are called latent variables (LVs). The PLS algorithm computes the LVs with maximal covariance with the response variables. After mathematical treatment, the PLS model can be recast into a standard linear regression model, such as MLR. Further details on the PLS method are provided by Geladi and Kowalski (1986) and Martens and Næs (1989). The maximum number of LVs that can be used in PLS corresponds to the number of original regressors, in which case the PLS models is equivalent to MLR, presenting the same limitations. Therefore, those LVs that do not contribute significantly to the prediction ability of the model should be discarded. Several methods were developed to determine the optimum number of LVs (Geladi and Kowalski, 1986). Cross-validation is one of these methods, where the number of LVs to adopt is the one minimizing the PRESS statistic (Prediction Residual Sum of Squares) or, equivalently, the one maximizing the predictive coefficient of determination $R^2_{pred}$ Eq (1):

$$R^2_{pred} = 1 - \frac{PRESS}{SS_{Total}} = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{i=1}^{n}(y_k - \bar{y})^2} \quad (1)$$

where, $y_k$ is the response variable value, $\hat{y}_k$ the response variable value obtained by the model for observation $k$, and $\bar{y}$ is the mean of the response values.

Once estimated, every model should be validated by testing its performance on an independent and never seen before test data set. According to Esbensen et al. (2002) this is the best validation method as the new data set is supposed to represent future observations. The performance of the model will then be evaluated using Normalized Root Mean Squares Error of Prediction (NRMSEP) that measures, in percentage, the dispersion between predicted values and observed ones Eq (2) relatively to the operation range.

$$NRMSEP = \frac{\sqrt{\sum_{k=1}^{n}(y_k - \hat{y}_k)\Big/n}}{\max(y) - \min(y)} \times 100 \quad (2)$$

In this work, the database available is sufficiently rich to allow for the use of two different sets: the training set consists of data from the year 2011, while the test set is comprised of data collected in 2012.

## 3. Data collection

All measured variables available in the industrial plant, concerning to reaction section, were collected, analyzed and evaluated. More specifically, the input variables or regressors, are: temperature along the reaction section ($T_1$, $T_2$ and $T_3$); mixed acid inlet temperature and volumetric flow rate ($T_{MA}$ and $F_{v,MA}$, respectively); nitric and sulfuric acid weight fractions in mixed acid at reactor inlet ($\%\ NA_i$, and $\%\ SA_i$); benzene and nitric acid mass flow rate ($F_{w,B}$ and $F_{w,NA}$); inlet molar feed ratio between benzene and nitric

acid ( $F_{m,B}/F_{m,NA}$ ) and residence time ($\theta$). Temperatures and flow rates were collected using the plant DCS system, while the samples collected in the plant were analyzed in the quality control laboratory in order to determine composition. The response variables are 2,4-dinitrophenol (DNP) and trinitrophenol (TNP) content in crude mononitrobenzene.

## 4. Results and discussion

Table 1 shows some summary statistics for the regressors and response variables in raw data set. The coefficient of variance (CV) value is defined as the ratio of SD to the mean, and is a useful quantity to assess variability irrespectively of the units used. From Table 1 one can see that DNP and TNP concentrations have higher CV values, followed by the flow rates, denoting a high relative variability in these variables.

*Table 1: Coefficient of variance for the raw data set.*

| | $T_1$ | $T_2$ | $T_3$ | $T_{MA}$ | $F_{w,NA}$ | $F_{w,B}$ | $F_{v,MA}$ | %$NA_i$ | %$SA_i$ | $\theta$ | $\frac{F_{m,B}}{F_{m,NA}}$ | DNP | TNP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CV | 2.9 | 1.4 | 1.5 | 1.4 | 16.2 | 15.7 | 12.5 | 5.7 | 1.1 | 14.0 | 2.6 | 18.5 | 27.3 |

In the exploratory data analysis the presence of outliers was also analyzed. Outliers are extreme values that do not conform with the dominant variability pattern. In a multivariate context, they can be easily overlooked and influence the results of the whole regression analysis. Therefore, before accepting a model as final, it is crucial to carefully analyze the presence of these observations, through a residual and influential analysis, in order to identify outliers and evaluate their role in the model. Outliers should only be removed from the dataset if a sound and explainable reason for the unusual behaviour exists, or if they clearly distort the accepted behaviour of the process (Montgomery et al., 2007).

### 4.1 Model Development

The statistical analysis of data was conducted with *Minitab (version 16).* Preliminary Gage R&R studies were performed to validate the measurement systems in use. The statistical significance of the DNP and TNP formation model as a whole was assessed through an Analysis of Variance statistical test for regression (ANOVA). The test statistics is described in Eq (3), which should be higher than the tabulated critical f-value ( $f_{crit}$ ) determined based on degrees of freedom (DF), at a significance level ($\alpha$) of 0.05.

$$f = \frac{SS_R/DF_R}{SS_{Res}/(n - DF_R - 1)} = \frac{s_R^2}{s_{Res}^2} > f_{crit} \tag{3}$$

The ANOVA table also provides information on the models' p-value, that represents a measure of the significance of the model (the lower the p-value, the more significant is the model; a model is declared significant if its p-value is lower than the adopted significance level, $\alpha$, which in our case was set to 0.05).
All variables were autoscaled by subtracting their mean value and dividing by their standard deviation. Therefore, the predicted models obtained in this work directly reveal which variables play a more significant role through the direct comparison of the magnitudes of variables' coefficients. These normalized models are equivalent to their counterparts that do not make use of normalized variables and thus, both versions (scaled and un-scaled variables) present the same $R^2$.

**Predicting 2,4-Dinitrophenol formation**

A total of 8 LVs were selected by cross-validation, in the PLS model to explain the DNP formation, leading to $R^2$ value of 0.395 and a $R^2_{pred}$ value of 0.369. These values are low and denote a weak relation between regressors and the response variable. The limited range of the responses, which is a desirable feature in industry, does not help the development of regression models on the other hand. Some important input variables are also kept approximately constant under normal operation conditions, making it difficult to capture their impact on the response. However, the model obtained is significant (Table 2): the F-value is higher than critical f-value ($f_{crit}$ < 1.93) and the p-value is also lower than $\alpha$, proving the adequacy of rejecting the null hypothesis. Therefore, it should be analyzed in detail.

*Table 2: Analysis of variance for the model estimated for predicting the formation of DNP (training set).*

| | DF | SS | $s^2$ | F-Value | P-Value |
|---|---|---|---|---|---|
| **Regression (R)** | 8 | 45.1 | 5.6 | 37.7 | <0.001 |
| **Residual Error (Res)** | 461 | 68.9 | 0.1 | | |
| **Total** | 469 | 114.1 | | | |

Model validation was conducted through residual analysis, where all assumptions underlying the regression model were confirmed. Therefore, the final model, composed by standardized coefficients, for predicting the DNP formation is given by Eq (4).

$$DNP^* = 1.04 \times T^*_1 - 0.47 \times T^*_3 - 0.17 \times T^*_{MA} + 0.70 \times F^*_{v,MA} + 0.10 \times \frac{F_{m,B}}{F_{m,NA}}^* + 0.64 \times \theta^* + 0.27 \times \%NA_i^* - 0.34 \times \%SA_i^* \quad (4)$$

In Eq (4), one can see that $T^*_1$, the autoscaled temperature at the beginning of reaction section, is an influencing variable on DNP formation as well as mixed acid volumetric flow rate and residence time, and it is known that these three variables contribute to main reaction extent. The relationship between $T_1$ and DNP formation was already seen in preliminary exploratory data analysis and nitration temperature is also an influential variable in Quadros et al. (2005) model and was later confirmed by Portugal et al. (2009). However, the signal of the coefficient of regression corresponding to the mixed acid volumetric flow rate shows an opposite influence on DNP formation comparing with that observed in exploratory data analysis. This can be a result of the particular correlation structure of data in industry, caused for instance by the control systems implemented. The effects for other variables are easily explained, such as for residence time (increasing the time in which the mixture passes through reaction section also increases the DNP formation).

**Predicting Trinitrophenol Formation**

The PLS model for predicting the TNP formation is based on 8 latent variables, selected by cross-validation. Model quality is assessed by $R^2$ and $R^2_{pred}$ parameters, which now take the values of 0.675 and 0.662, respectively. These values are much higher than the ones obtained for DNP model, implying a better relationship between response variable and regressors. This statement is corroborated by the larger variability explained by regressors ($SS_R$) when compared with variability caused by unexplained reasons ($SS_{Res}$), registered in Table 3. The p-value and f-value reported in Table 3 also confirm that the TNP model is statistically significant ($f_{crit}$ < 1.93).

*Table 3: Analysis of variance for the model estimated for predicting the formation of TNP (training set).*

| | DF | SS | $s^2$ | F-Value | P-Value |
|---|---|---|---|---|---|
| **Regression (R)** | 8 | 66.6 | 8.3 | 121.4 | <0.001 |
| **Residual Error (Res)** | 467 | 32.0 | 0.1 | | |
| **Total** | 475 | 98.6 | | | |

Model adequacy was checked by residual analysis that showed the non-existence of any violation of regression assumptions. The mathematical expression, using autoscaled variables and standardized coefficients, for predicting the TNP formation is given by Eq (5).

$$TNP^* = -0.81 \times T^*_1 + 0.72 \times T^*_3 + 0.27 \times T^*_{MA} - 0.87 \times F^*_{v,MA} - 0.08 \times \frac{F_{m,B}}{F_{m,NA}}^* - 0.21 \times \theta^* + 0.03 \times \%NA_i^* - 0.28 \times \%SA_i^* \quad (5)$$

Mixed acid volumetric flow rate is again one of the most influencing variables on TNP formation, followed by $T_1$, temperature in the beginning of reaction section. Contrary to what was obtained in DNP model, these variables are inversely proportional to TNP content in crude MNB and this was not expected as it is accepted that DNP is a precursor to TNP formation. It is important to highlight the model includes temperature $T_3$ as one of the most influential variables. This can be envisaged as a confirmation that NPs formation reactions are consecutive and $T_3$ must be kept low when the minimization of TNP content on MNB is a goal. Most operating parameters in Eq (5) were already included in the model of Quadros et al.

(2005) and also used by Portugal et al. (2009), although their contribution might not always have been the same.

Figure 1 shows the quality of DNP and TNP model, where dashed lines represent error bounds related to the measurement systems obtained from Gage R&R Studies previously performed. Figure 1a reveals the low explanatory capacity due to low variability of data in the industrial data set. On the other hand, Figure 1b displays the good predictive performance of the TNP model using the training set.

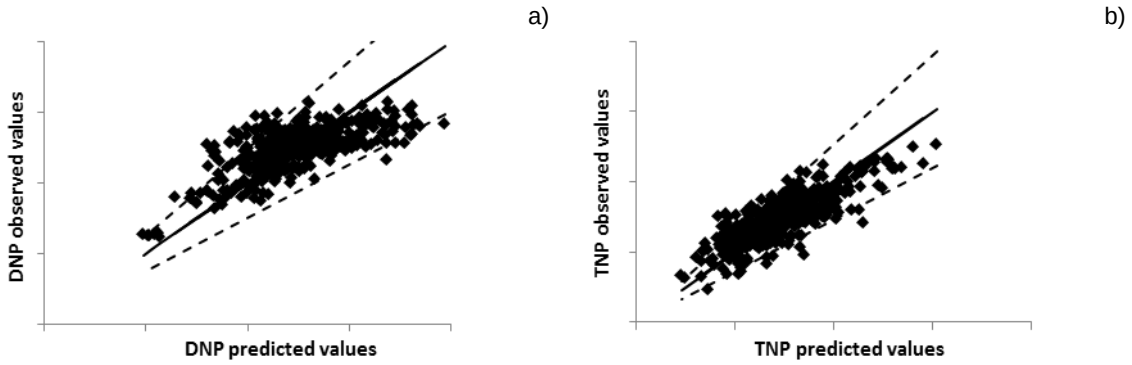


*Figure 1: Observed values versus predicted by models for training set a) DNP and b) TNP.*

**4.2 Model Validation**

Every model developed should be validated by evaluating its predictive performance in future observations. NRMSEP is the parameter used to evaluate model performance in a test set validation, allowing a comparison between different models. Table 4 shows $R^2$ as well as NRMSEP values for the test set, indicating, as expected, a worse performance of DNP model over TNP model. The prediction ability of the TNP model is confirmed, as the values of the performance indicators are in line with those obtained for the training set, from which it was estimated (Figure 1).

*Table 4: Coefficient of determination and normalized NRMSEP values for test set.*

| | $R^2$ | NRMSEP (%) |
|---|---|---|
| **DNP** | -1.039 | 10.12 |
| **TNP** | 0.626 | 7.04 |

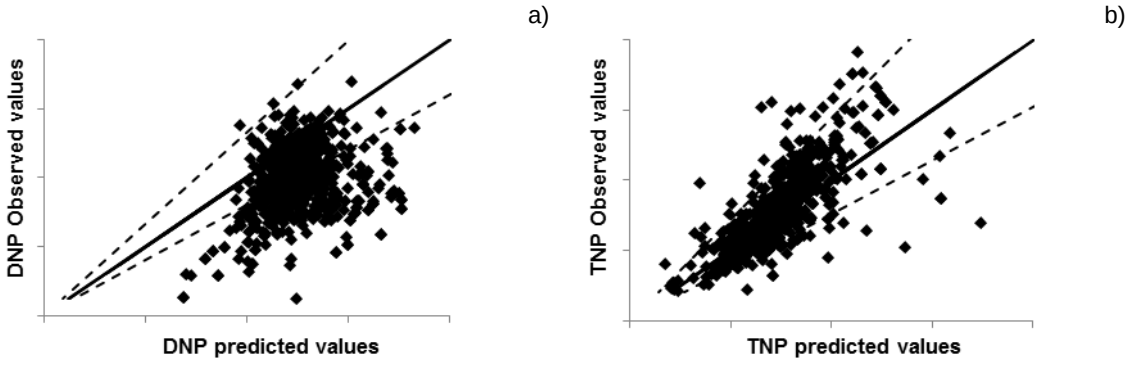


*Figure 2: Observed values versus predicted by models for test set a) DNP and b) TNP*

In fact, comparing Figure 1 and Figure 2 the importance of the validation step using an independent data set is obvious. In Figure 1a, where DNP observed values were related with DNP predicted values for training set, the predictive ability of the model seems to be confirmed, although not as good as expected. The predictive ability of DNP model decreases significantly when assessed using a test set (Table 4 and Figure 2a).

## 5. Conclusions

The application of the PLS methodology to an industrial data set is hoped to highlight the influence of operating conditions upon the performance of the process and contribute to its optimization. Nevertheless this goal is not easy achieving. Plant operation is ruled by production goals and final product specifications, which leave little opportunity to explore the influence of changes in operating conditions which might enrich the data set. On the other hand, the information a statistical model provides is much dependent on data set variety and quality.

This work allowed confirming reaction temperature as the most influential variables in nitrophenols formation in the benzene nitration process. Mixed acid volumetric flow rate is another relevant variable which was not considered in previous models, although the molar feed ratio between benzene and nitric acid ($F_{m,B}/F_{m,NA}$) had been included. Surprisingly, the error measures associated with the two models are different and lower for the TNP model, which presents a good prediction power and utility potential for industrial applications. The correlation coefficients obtained are lower than those previously achieved by Quadros et al. (2005) using pilot plant data which provides a dataset with increased variety. For the DNP model, these values are 0.395 and 0.369, respectively, whereas for the TNP model, they are 0.675 and 0.662, respectively. This difference can confirm the lower measurement variability in TNP collected data, leading to a better predictive performance of the model, when compared to the one developed for DNP. As these models are to be used under industrial conditions, reliability is an important issue, which was taken into account during the development of the models.

**Acknowledgment**

Financial support from Fundação para a Ciência e Tecnologia (FCT) for Ph.D. Grant SFRH/BDE/33907/2009 and from CUF-Químicos Industriais S.A., is gratefully acknowledged.

**Nomenclature**

| | |
|---|---|
| $F_m$ | Molar flow rate (mol/h) |
| $F_w$ | Mass flow rate (t/h) |
| $F_v$ | Volumetric flow rate ($m^3$/h) |
| $n$ | Number of observations |
| $s^2$ | Variance |
| $T$ | Temperature (ºC) |

| | |
|---|---|
| *Subscript* | |
| 1, 2, 3 | along reaction section |
| *Superscript* | |
| * | autoscaled variables |
| *Greeks and symbols* | |
| θ | Residence time (min) |
| % | Weight composition (wt/wt %) |

**References**

Adams E.G., Barker R.B., 1990, Process for extracting and disposing of nitrophenolic by-products, US patent application 4,986,917.

Esbensen K.H., Guyot D., Westad F., Houmøller L.P., 2002, Multivariate Data Analysis - in Practice: An Introduction to Multivariate Data Analysis and Experimental Design, Camo, Denmark.

Geladi P., Kowalski B.R., 1986, Partial Least-Squares Regression: A Tutorial, Analytica Chimica Acta, 185, 1-17.

Hanson C., Kaghazchi T., Pratt M. W T., 1976, Side Reactions During Aromatic Nitration. Industrial and Laboratory Nitrations. American Chemical Society,Washington, U.S.A.

Larbig W., 1980, Process for working up effluents containing nitro-hydroxy-aromatic compounds, U.S. patent application 4,230,567.

Martens H., Næs T., 1989, Multivariate Calibration, Wiley, Chichester, UK.

Montgomery D.C., Peck E.A., Vining G.G., 2007, Introduction to Linear Regression Analysis, 4th ed, John Wiley & Sons, USA.

Portugal P.A.G., Reis M.S., Baptista C.M.S.G., 2009, Extending model prediction ability for the formation of nitrophenols in benzene nitration, Chemical Engineering Transactions 17, 117-122. DOI:10.3303/CET0917020

Quadros P.A., Reis M.S., Baptista C.M.S.G., 2005, Different modelling approaches for a heterogeneous liquid-liquid reaction process, Industrial & Engineering Chemistry Research 44, 9414-9221.