# 다변량 통계분석법을 이용한 PET 중합공정 중 직접 에스테르화 반응기의 거동 및 생산제품 예측

# Multivariate Statistical Analysis Approach to Predict the Reactor Properties and the Product Quality of a Direct Esterification Reactor for PET Synthesis

김 성 영, 이 범 석*, 정 창 복, 최 수 형
(Sung Young Kim, Bomsock Lee, Chang Bock Chung, and Soo Hyoung Choi)

**Abstract :** The multivariate statistical analysis methods, using both multiple linear regression(MLR) and partial least square(PLS), have been applied to predict the reactor properties and the product quality of a direct esterification reactor for polyethylene terephthalate(PET) synthesis. On the basis of the set of data including the flow rate of water vapor, the flow rate of EG vapor, the concentration of acid end groups of a product and other operating conditions such as temperature, pressure, reaction times and feed monomer mole ratio, two multi-variable analysis methods have been applied. Their regression and prediction abilities also have been compared. The prediction results are critically compared with the actual plant data and the other mathematical model based results in reliability. This paper shows that PLS method approach can be used for the reasonably accurate prediction of a product quality of a direct esterification reactor in PET synthesis process.

**Keywords :** direct esterification, PET, multivariate analysis, PLS, MLR

## I. 서론

PET(polyethylene terephthalate)는 합성섬유, 필름, 음료수병, 성형 플라스틱 등의 제조에 가장 널리 사용되는 고분자 물질이다. PET에 대한 계속되는 수요로 인하여 PET 제조경비의 절감과 생산제품의 품질 향상에 관한 연구가 계속되어지고 있다. PET 생산라인에 대한 연구는 크게 두 가지로 나눠지고 있는데, 그 하나는 관련된 반응들에서 사용되는 촉매들의 효율을 높이기 위한 연구이며, 또 하나는 제품 생산성을 높이기 위해 반응조건들을 최적화하기 위한 모델링에 관한 연구이다[1].

PET 제조에 대한 대부분의 기술은 특허나 현장의 고유기술로 남아있으며, 실제로 현장에서 사용되고 있는 PET 제조공정은 테레프탈산(Terephthalic acid, TPA) 혹은 디메틸테레프탈레이트(Dimethyl Terephthalate, DMT) 및 에틸렌글리콜(Ethylene Glycol, EG)을 원료로 사용하고 있다. 현재 공업적으로 가장 많이 사용되고 있는 PET 연속식 제조공정은 크게 에스테르화 반응공정(esterification)과 중축합 반응공정(polycondensation)으로 나뉘어 질 수 있으며, 이중 에스테르화 반응공정은 DMT와 EG를 원료로 사용하는 에스테르화 교환반응(transesterification)방법과 TPA와 EG를 원료로 사용하는 직접 에스테르화 반응(direct esterification)방법이 사용되는데, 최근에 사용되는 PET 제조공정은 대부분 직접 에스테르화 반응방법을 사용하고 있다[2].

TPA와 EG를 원료로 사용하는 직접 에스테르화 반응기의 주요한 반응조건은 반응기의 온도, 압력, 반응시간과 투입되는 TPA와 EG의 몰비 등이다. 또한 직접에스테르화 반응기의 반응율은 반응기의 거동을 나타내는 유출수의 유량 및 EG의 함량, 그리고 반응기의 생산품 내에 존재하는 TPA 농도를 측정함으로써 알 수 있다[3]. 직접 에스테르화 반응기의 반응조건들의 변화에 따른 반응기의 거동과 에스테르화 반응율에 대한 연구들은 80년대 중반부터 본격적으로 연구되어져 왔다. PET 합성반응 중 직접에스테르화 반응기에 대한 수학적 모델링은 Ravindranath[4], Yamada[5], Kang[6] 등에 의해 수행되어져 왔으며, 최근에는 에스테르화 반응기의 수학적 모델을 기반으로 에스테르화 반응율을 높이기 위한 반응조건에 대한 최적화 연구가 진행되고 있다[1,3].

최근에는 화학반응이 복잡하고 반응상수들에 대한 정보가 부족한 실제 화학반응공정에 대해서는 입출력 데이터를 기반으로 하는 실험적 모델이 사용되고 있다. 실험적 모델을 만드는데 많이 사용되는 다변량 통계분석법으로는 다중 선형 회귀법(MLR, Multiple Linear Regression)이나 부분 최소자승법(PLS, Partial Least Square) 등이 있으며, 이들은 복잡한 화학공정의 실험적 모델링이나 생산제품의 품질예측 방법으로 널리 사용되고 있다. Eriksson 등은 PLS와 MLR 방법을 이용하여 수질 오염데이터를 분석하고 그 결과를 비교하였다[7]. MacGregor 등은 저밀도 폴리에틸렌 제조공정의 이상진단에 PLS 방법이 효과적으로 사용됨을 보여주었다[8]. Suzuki 등은 유기화합물의 점도를 예측하기 위한 방법에 PLS와 MLR을 사용하여 그 결과를 비교하였다[9]. Qin은 순

* 책임저자(Corresponding Author)
논문접수 : 2004. 10. 19., 채택확정 : 2005. 2. 8
김성영, 이범석 : 경희대학교 환경응용화학부
(alla97@hanmail.net/bslee@khu.ac.kr)
정창복 : 전남대학교 응용화학공학부(chungcb@chonnam.ac.kr)
최수형 : 전북대학교 화학공학부(soochoi@chonbuk.ac.kr)
※ 본 논문은 한국과학재단(과제번호 R01-2003-000-10697-0)과 경희대학교에서 지원하여 연구하였음.

환 PLS 방법을 개발하여 많은 입출력 변수가 존재하는 촉매 개질기의 생산제품의 옥탄가를 예측하는데 사용하였다[10]. PLS 방법은 증류공정의 국부적인 조성제어 소프트센서를 개발한 연구에도 사용되었다[10]. Han 등은 TPA 제조공정의 생산제품의 품질제어를 위해 PLS 방법을 사용하였다[11].

본 논문에서는 PET 중합공정에 사용되는 직접 에스테르화 반응기의 반응율에 큰 영향을 미치는 온도, 압력, 반응시간, TPA와 EG의 몰비를 입력데이터로 하고, 반응기의 유출수의 유량과 유출수에 포함된 EG의 유량, 그리고 생산제품 내에 존재하는 acid end group의 농도를 출력데이터로 하는 실험적 모델을 만들기 위해 MLR과 PLS 방법을 적용해 볼 것이며 실험적 모델로 만들어진 예측 결과들을 실제 현장 데이터와 비교할 것이다. 또한 두 예측 결과들은 참고문헌 [3]에서 제시되어져 있는 수학적 모델링으로 예측된 결과와도 비교되어질 것이다.

## II. 다변량 통계분석법

입출력 데이터가 많은 복잡한 화학공정의 경우 다변량 데이터를 분석하고 생산제품의 물성을 예측하는데 다변량 통계분석법이 매우 유용하게 사용될 수 있다. 다변량 통계분석법은 간단히 말해 여러 입출력 데이터를 동시에 분석하는 통계적 기법이다. 화학공정의 모델링은 화학반응식과 물질 및 에너지 수지식을 사용하는 수학적 모델링(mathematical modeling)과 실제 실험 데이터에 의존한 실험적 모델링(empirical modeling)으로 구분될 수 있는데, 이때 다변량 통계분석법으로 만들어지는 화학공정의 모델은 실험적 모델이라고 할 수 있다.

복잡한 화학공정의 경우 사용되는 화학반응식이 매우 많고 반응상수들에 대한 정확한 정보가 부족하면 정확한 모델을 구성할 수 없는 제한이 따른다. 반면에 실험적 모델의 경우 공정에 대한 정확한 이해가 없어도 실제 공정 데이터를 기반으로 모델을 세울 수 있다. 본 논문에서는 다변량 통계분석법의 대표적 두가지 방법인 MLR과 PLS 방법을 사용하여 직접에스테르화 반응기의 실제 공정 데이터를 기반으로 실험적 모델을 만들려고 한다.

### 1. MLR(Multiple Linear Regression)

MLR은 입출력 변수가 각각 한개씩인 단변량 데이터에 적용하는 최소자승법(least square method)을 입출력 변수가 각각 여러개인 다변량 데이터에 적용할 수 있도록 확장한 방법으로서, 여러개의 입력 변수와 출력 변수를 한꺼번에 분석할 수 있는 장점을 가진다. MLR에 관련된 이론은 많은 참고문헌[13]에서 찾을 수 있으며, 이를 정리하면 다음과 같다.

$m$개의 입력 변수 $x_i$ ($i$=1,⋯,$m$)와 $p$개의 출력 변수 $y_j$ ($j$=1,⋯,$p$)와의 선형관계는 다음과 같은 선형 회귀식으로 나타낼 수 있다.

$$y_j = \sum_{i=1}^{m} b_{ij}x_i + e_j \quad (j = 1, \cdots, p) \tag{1}$$

여기서 $b_{ij}$는 입력 변수 $x_i$의 출력변수 $y_j$에 대한 회귀 계수이며 $e_j$는 출력변수의 측정값 $y_j$와 회귀값 $\sum_{i=1}^{m} b_{ij}x_i$ 간의 오차이다. 입력 변수 $x_i$와 출력 변수 $y_j$ 사이의 실험적 모델을 만들기 위해 채택한 실험 횟수를 $n$($k$=1,⋯,$n$)번이라고 한다면 $k$번째 실험한 입력 변수 $x_{ki}$ 값과 출력 변수 $y_{kj}$ 값 사이의 선형 관계식은 다음과 같이 표시될 것이다.

$$y_{kj} = \sum_{i=1}^{m} b_{ij}x_{ki} + e_j \quad (k = 1, \cdots, n,\ j = 1, \cdots, p) \tag{2}$$

$n$번의 실험으로 얻어진 입출력 변수들의 값을 다음과 같이 행렬로 표시되는 출력 변수 데이터 블록($\boldsymbol{Y}$) 및 입력 변수 데이터 블록($\boldsymbol{X}$)으로 표시하면 (2)는 행렬 방정식으로 표시할 수 있게 된다.

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} \text{(n×p 행렬)}$$

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \text{(n×m 행렬)}$$

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{pmatrix} \text{(m×p 행렬)}$$

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{e} \tag{3}$$

여기서 B는 입력 변수 데이터 블록($\boldsymbol{X}$) 및 출력 변수 데이터 블록($\boldsymbol{Y}$)을 선형 관계짓는 계수 행렬을 나타내며, $\boldsymbol{e}$는 오차를 나타내는 열벡터이다. 오차 $\boldsymbol{e}$를 최소($\boldsymbol{e}$≒0)로 만드는 계수 행렬 $\boldsymbol{B}$는 최소 자승법을 이용하여 다음과 같은 행렬 연산에 의하여 구해진다.

$$\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{4}$$

(4)로 구해진 행렬 $\boldsymbol{B}$를 이용하면 MLR 모델을 만드는데 이용되었거나 이용되지 않은 입력 변수 $x_i$ 값에 대한 출력 변수 $y_j$의 예측값을 (1)($e_j$=0) 혹은 (3)($\boldsymbol{e}$=0)을 사용하여 계산할 수 있다. MLR 방법은 다른 다변량 통계분석법과 달리 사용하기가 간편하고 변수들이 갖고 있는 의미를 이해하기 쉬운 장점을 갖고 있으나 다음과 같은 단점들을 가지고 있다. 첫째는 입력 변수들의 실험값들이 서로 선형적으로 독립적이지 못할 때, 즉 행렬 $\boldsymbol{X}$의 열들이 서로 선형적으로 독립적이지 못할 때, (4)에 사용되는 행렬 $\boldsymbol{X}^T\boldsymbol{X}$가 full rank를 가지지 못해 역행렬을 구할 수 없으며 따라서 계수 행렬 $\boldsymbol{B}$를 구할 수 없다는 것이다. 혹은 선형적으로 모두 독립적이라고 할지라도 선형 상관관계가 매우 크면 입출력 변수 $x$, $y$값이 조금만 변해도 계수 행렬 $\boldsymbol{B}$ 값의 변화가 너무 커져서 신뢰도가 떨어진다는 단점이 있다. 두 번째 단점으로는 입력 변수의 개수 $m$에 비하여 채택할 수 있는 실험 횟수 $n$가 더 적은 경우 행렬 $\boldsymbol{B}$의 값이 여러개 나올 수 있으며, $n$이 $m$에 비하여 비슷하거나 그리 크지 않으면 행렬 $\boldsymbol{B}$의 신뢰도가 떨어져 예측 성능이 나빠지는 단점이 있다.

### 2. PLS(Partial Least Square)

PLS기법은 입출력 변수값들을 그대로 사용하는 MLR과

는 달리 새로운 잠재 변수(latent variable)들을 찾아내 회귀에 사용하는 비교적 새로운 다변량 통계분석법이다. PLS 회귀에 사용되는 잠재변수들은 서로 선형적으로 독립이며 입력 잠재변수와 출력 잠재변수 사이에는 매우 높은 상관관계를 갖게 되며 입력 변수 중에 분산(variance)값이 높은 순서대로 잠재변수로 채택된다. PLS에 관련된 이론은 많은 참고문헌[10-14]에서 찾아볼 수 있으며, 이를 정리하면 다음과 같다.

입력 변수 데이터 블록($\boldsymbol{X}$)은 score vector $t_h(h=1,\cdots a)$와 load vector $p_h(h=1,\cdots a)$의 곱으로 나타내어지고, 출력 변수 데이터 블록($\boldsymbol{Y}$)은 score vector $u_h(h=1,\cdots a)$와 load vector $q_h(h=1,\cdots a)$의 곱으로 표시되어진다.

$$\boldsymbol{X}=\sum_{h=1}^{a}\boldsymbol{t}_h\boldsymbol{p}_h^T+\boldsymbol{E} \tag{5}$$

$$\boldsymbol{Y}=\sum_{h=1}^{a}\boldsymbol{u}_h\boldsymbol{q}_h^T+\boldsymbol{F} \tag{6}$$

$$\boldsymbol{u}_h=\sum_{h=1}^{a}b_h\boldsymbol{t}_h+\boldsymbol{r}_h \tag{7}$$

윗 식에서 $a$는 채택될 잠재변수의 개수를 나타내며, $b_h$는 입력 변수의 score vector $t_h(h=1,\cdots a)$와 출력 변수의 score vector $u_h(h=1,\cdots a)$간의 회귀계수이다. PLS를 수행하는 대표적인 NIPALS 알고리즘은 다음과 같다.

행렬 $\boldsymbol{Y}$의 한 열을 선택하여 $u_1$의 초기값으로 놓은 후, 다음 식들을 수행한다.

$$\boldsymbol{w}_1=\frac{\boldsymbol{X}^T\boldsymbol{u}_1}{\|\boldsymbol{X}^T\boldsymbol{u}_1\|} \tag{8}$$

$$\boldsymbol{t}_1=\boldsymbol{X}\boldsymbol{w}_1 \tag{9}$$

$$\boldsymbol{q}_1=\frac{\boldsymbol{Y}^T\boldsymbol{t}_1}{\|\boldsymbol{Y}^T\boldsymbol{t}_1\|} \tag{10}$$

$$\boldsymbol{u}_1=\boldsymbol{Y}\boldsymbol{q}_1 \tag{11}$$

(11)에서 계산된 $u_1$값이 일정 오차범위 내에 수렴하면 다음 식들을 수행하고, 그렇지 않으면 (8)로 되돌아간다.

$$\boldsymbol{p}=\frac{\boldsymbol{X}^T\boldsymbol{t}_1}{\boldsymbol{t}_1^T\boldsymbol{t}_1} \tag{12}$$

$$\boldsymbol{p}_1=\frac{\boldsymbol{p}}{\|\boldsymbol{p}\|} \tag{13}$$

$$\boldsymbol{t}_1=\boldsymbol{t}_1\|\boldsymbol{p}\| \tag{14}$$

위와 같이 $t_1$과 $u_1$이 구해지면 다음 식으로 두 score vector간의 회귀계수를 구한다.

$$b_1=\frac{\boldsymbol{u}_1^T\boldsymbol{t}_1}{\boldsymbol{t}_1^T\boldsymbol{t}_1} \tag{15}$$

첫 번째 잠재변수가 결정되면 다음 식으로 데이터 블록 $\boldsymbol{X}$와 $\boldsymbol{Y}$의 오차행렬 $\boldsymbol{E}$와 $\boldsymbol{F}$를 구한다.

$$\boldsymbol{E}_1=\boldsymbol{X}-\boldsymbol{t}_1\boldsymbol{p}_1^T \tag{16}$$

$$\boldsymbol{F}_1=\boldsymbol{Y}-b_1\boldsymbol{t}_1\boldsymbol{q}_1^T \tag{17}$$

또 다른 잠재변수가 필요한 경우에는 행렬 $\boldsymbol{X}$와 $\boldsymbol{Y}$를 각각 $\boldsymbol{E}_1$와 $\boldsymbol{F}_1$으로 바꾸어서 위 알고리즘을 다시 수행한다. 이때 아래 첨자는 1씩 증가한다. PLS 모델에 사용되는 잠재변수의 개수 $a$는 일반적으로 독립변수의 개수보다는 적으며 보통 cross-validation 방법으로 결정된다[13]. 잠재변수의 개수가 독립변수의 개수와 같아질수록 오차행렬 $\boldsymbol{E}$와 $\boldsymbol{F}$는 점점 0에 가까운 값을 갖게 된다.

(5)-(17)로 계산된 load vector $p_h$, $q_h(h=1,\cdots,a)$와 회귀계수 $b_h(h=1,\cdots,a)$를 사용하여 학습에 사용되었거나 혹은 사용되지 않은 입력변수 데이터 블록 $\boldsymbol{X}$에 대한 출력변수 데이터 블록 $\boldsymbol{Y}$의 예측값을 계산하려면, 먼저 (5)($\boldsymbol{E}\cong 0$)를 만족하는 $t_h(h=1,\cdots,a)$를 다음 식으로 구한다.

$$\boldsymbol{t}_h=\frac{\boldsymbol{X}\boldsymbol{p}_h}{\boldsymbol{p}_h^T\boldsymbol{p}_h}\quad(h=1,\cdots,a) \tag{18}$$

$t_h(h=1,\cdots,a)$가 계산되어지면 $b_h(h=1,\cdots,a)$와 (7)($r_h\cong 0$)을 이용하여 $u_h(h=1,\cdots,a)$를 구한 후, 최종적으로 $q_h(h=1,\cdots,a)$와 (6)을 이용하여 $\boldsymbol{Y}$의 예측값을 구한다.

많은 변수들로 구성된 화학공정에 MLR 알고리즘을 적용하면 회귀 모델을 만들기 위해 사용된 학습 데이터에 대해서는 뛰어난 회귀 성능을 보이지만, 학습에 사용되지 않은 새로운 데이터에 대해서는 제대로된 예측을 수행하지 못한다. 이는 너무 많은 입력 변수들에 대해서 모델이 과적합(over-fitting)되었기 때문이다. PLS는 출력 변수값에 큰 영향을 미치는 몇몇 잠재변수들만을 이용하는 다변량 통계 분석법으로서 MLR보다 학습 데이터에 대한 회귀성능은 떨어질지 몰라도 새로운 데이터에 대해서는 뛰어난 예측 성능을 보인다. 그러나 PLS는 사용되는 잠재변수들의 의미를 이해하기 힘들다는 단점을 갖고 있다.

### 3. 데이터 전처리 과정(data prepocessing)

데이터 전처리 과정이란 MLR이나 PLS 알고리즘을 수행하기 전에 입출력 데이터를 보정하여 계산을 용이하게 하고 모델 수립에 보다 편리한 데이터로 만들어 주는 과정이다. 데이터 전처리 과정은 mean centering과 variance scaling 두 과정으로 나뉘는데, mean centering은 측정된 데이터 값들로부터 평균값을 구하고 각 측정값에서 평균값을 빼주는 과정을 말한다. Variance scaling은 변수들이 서로 다른 단위로 측정될 때 각각의 변수값을 표준편차로 나누어 각각의 변수에 대한 variance가 1이 되도록 하는 과정이다. Mean centering 과정을 수행하는 데이터 전처리 과정을 식으로 나타내면 다음과 같다.

$$\text{mean-centered } x_{ki}=x_{ki}-\bar{x}_i \tag{19}$$

일반적으로 데이터 전처리 과정을 수행하면 그렇지 않은 경우보다 MLR이나 PLS의 회귀 성능이 더 좋아진다[13]. 본 논문에서는 MLR과 PLS 알고리즘을 수행하기전에 mean centering만을 사용하여 데이터를 전처리하였다.

### 4. RMSE(Root Mean Square Error)

MLR이나 PLS를 사용하여 구성된 모델의 성능은 실제값

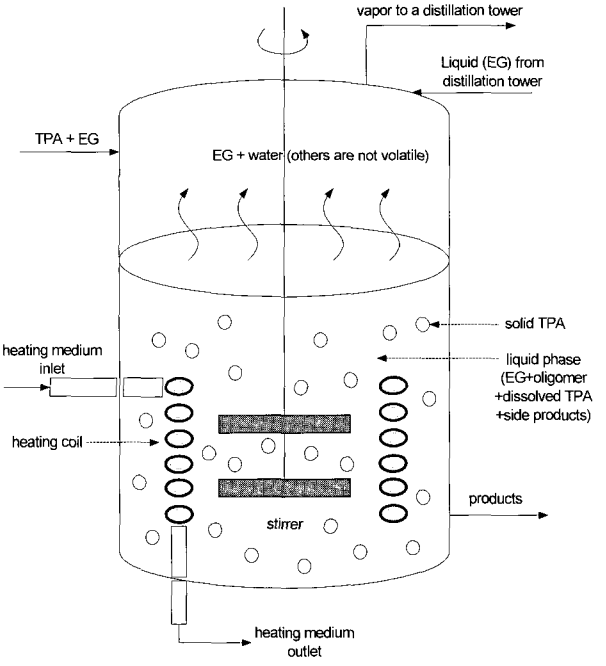



그림 1. PET 중합 공정 중 직접 에스테르화 반응기.
Fig. 1. Direct esterification reactor in PET synthesis.

과 예측값의 차이를 계산함으로써 알 수 있다. 이때 사용되는 수치가 RMSE이며 $j$번째 출력 변수의 RMSE를 수식으로 나타내면 다음과 같다.

$$\mathrm{RMSE}_j = \sqrt{\frac{\sum_{k=1}^{n} (\hat{y_{kj}} - y_{kj})^2}{n}} \tag{20}$$

위 식에서 $\hat{y_{kj}}$는 입력변수 $x_{ki}(i=1,\cdots,m)$를 가지고 MLR이나 PLS를 사용하여 계산된 $j$번째 출력변수의 예측값을 나타낸다. 일반적으로 RMSE는 하나의 출력변수에 대한 모델의 오차를 나타내는 값으로 사용된다.

## III. PET 중합공정 중 직접 에스테르화 반응기

PET 중합공정은 긴 반응시간과 높은 반응온도, 대용량의 다단계 공정시설을 필요로 하는 대표적인 에너지 소비 공정이다[1,3]. 특히, PET 전체 중합공정의 다른 부분에서 사용되는 에너지의 양에 비해 월등히 많은 에너지를 사용하는 직접 에스테르화 반응기 내부에 관한 연구는 많은 참고문헌에서 찾아볼 수 있다[1-6]. PET 직접 에스테르화 반응기를 간단히 표시하면 그림 1과 같다. 직접 에스테르화 반응기에는 TPA와 EG가 원료로 들어가며, 이때 반응율을 결정짓는 반응조건은 반응온도 $T$ [℃], 압력 $P$ [atm], 반응시간 $\tau$ [hr], 원료로 투입되는 TPA와 EG의 몰비율 $a$ [-]등이다[2,3]. 반응기에서 발생하는 water-EG 수증기는 증류탑과 응축기로 구성되어 있는 회수 장치를 통과하여 수증기에 포함되어 있는 EG는 반응기로 다시 회수된다. TPA와 EG의 반응 정도는 반응기에서 유출되는 water-EG 수증기의 양으로서 측정할 수 있다[2]. 직접 에스테르화 반응에서 생성되는 부산물 중 디에틸렌글리콜(DEG)은 PET의 물리적 및 화학적 물성에 가장 큰 영향을 미치는 물질이다. PET에 포함되는 DEG의 양이 증가할 수록 PET의 용융 온도 및 결정화 온도가 낮아지기 때문에 반응기 내부의 DEG 농도를 정확하게 예측하는 것은 매우 중요하다.

## IV. MLR, PLS 모델링 및 분석

PET 중합공정 중 직접 에스테르화 반응기를 대상으로 본 논문에서 사용할 MLR, PLS 모델의 입력 변수는 반응온도 $T$ [℃], 압력 $P$ [atm], 반응시간 $\tau$ [hr], 원료로 투입되는 TPA와 EG의 몰비율 $a$ [-]이며, 출력 변수는 반응기에서 유출되는 수증기의 유량($F^v_{water}$)과 EG vapor의 유량($F^v_{EG}$), 그리고 반응기 내부에 존재하는 acid end group ($C^o_{TPA}$, $C^o_{tTPA}$, $C^o_{bTPA}$)의 농도($C^o_{acid}$)이다. 위에 언급한 입출력 변수들에 대한 실제 공정의 조업 데이터는 참고문헌에서 찾아볼 수 있다[3]. 반응기에서 유출되는 수증기의 유량은 에스테르화 반응 정도에 의해 영향을 받으며, 특히 DEG 생성반응에 크게 의존한다. 따라서 주어진 반응조건을 가지고 유출되는 수증기의 유량을 예측하면 결과적으로 반응기 내부의 DEG 농도를 추정할 수 있다. 이와 비슷하게 반응기의 전체 반응율은 반응기 내부의 acid end group의 농도로 추정할 수 있으며, EG vapor의 유량으로는 에스테르화 반응의 반응율을 추정할 수 있다[3]. 실제 PET 중합공정의 에스테르화 반응기의 조업 데이터를 표 1에 나타내었다. 표 1 중에 데이터 번호 1-13은 MLR, PLS 모델 수립에 필요한 학습데이터로 사용할 것이며, 데이터 번호 14-17은 학습되지 않은 데이터에 대한 두 모델의 예측 성능을 비교하는데 사용될 것이다. MLR, PLS 모델 분석을 하기 전에 각 입출력 변수들의 평균값을 구하여 데이터 전처리 과정을 수행하였다.

표 1의 데이터 1-13으로 주어진 에스테르화 반응기 모델의 입력변수 $T$ [℃], $P$ [atm], $\tau$ [hr], $a$ [-]와 출력변수 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$에 대해 MLR 알고리즘((1)-(4))과 PLS 알고리즘((5)-(18))을 수행하였다. PLS 모델을 수립할 때 잠재변수의 개수는 3개로 정하였다. PLS 잠재변수의 개수가 늘어날 수록 회귀성능은 MLR과 비슷하게 되며, 잠재변수의 개수가 4개가 되면 MLR과 동일한 결과를 얻게 된다. 표 1의 데이터 1-13으로 주어진 학습 데이터에 대한 MLR, PLS 모델의 회귀성능을 비교하기 위해 주어진 입력변수 값에 대한 출력변수의 예측값을 수행한 결과를 표 2에 나타내었다. 표 2에 나타낸 데이터 값들은 두 모델로 예측된 결과값에 대하여 데이터 전처리 과정을 역으로 수행하여 주어진 단위에 맞게 다시 계산한 값들이다. MLR, PLS 각 모델의 회귀성능은 출력변수에 대한 RMSE 값을 가지고 비교해볼 수 있다. 표 2에 나타난 RMSE 값을 비교해 보면 모든 출력 변수 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$에 대해 MLR 모델의 RMSE 값이 PLS 보다 더 작은 것을 알 수 있다. 이는 주어진 학습 데이터의 출력변수에 대한 MLR 모델의 회귀 성능이 PLS

모델의 회귀 성능보다 더 우수하다는 것을 말한다.

그림 2는 MLR 모델을 수립하기 위하여 사용된 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$의 실제 현장 데이터 값과 MLR 모델로 구한 예측값을 비교한 그래프이다.

표 1. 직접 에스테르화 반응기의 현장 조업 데이터[3].

Table 1. The actual plant data of direct esterification reactor[3].

| No. | 입력 변수 | | | | 출력 변수 | | |
|---|---|---|---|---|---|---|---|
| | $T$[℃] | $P$[atm] | $\tau$[hr] | $a$[-] | $F^v_{water}$ | $F^v_{EG}$ | $C^o_{acid}$ |
| 1 | 260 | 1.2383 | 3.3277 | 1.2100 | 0.1421 | 0.1835 | 0.3248 |
| 2 | 259 | 1.2383 | 3.3228 | 1.2422 | 0.1435 | 0.1846 | 0.3237 |
| 3 | 259.9 | 1.2383 | 3.3388 | 1.2518 | 0.1467 | 0.1893 | 0.3082 |
| 4 | 260 | 1.2383 | 3.3472 | 1.2800 | 0.1668 | 0.1902 | 0.2948 |
| 5 | 260.3 | 1.2383 | 3.1729 | 1.3425 | 0.2056 | 0.1890 | 0.2769 |
| 6 | 260 | 1.2383 | 3.2159 | 1.3504 | 0.2117 | 0.1897 | 0.2747 |
| 7 | 259 | 1.1125 | 3.9807 | 1.2288 | 0.1565 | 0.1907 | 0.2898 |
| 8 | 259 | 1.1706 | 3.8836 | 1.2204 | 0.1408 | 0.1842 | 0.3062 |
| 9 | 259 | 1.1899 | 3.4612 | 1.2309 | 0.1426 | 0.1842 | 0.3085 |
| 10 | 259 | 1.1125 | 3.9807 | 1.2288 | 0.1565 | 0.1907 | 0.2898 |
| 11 | 259 | 1.1416 | 3.8014 | 1.2426 | 0.1582 | 0.1868 | 0.2919 |
| 12 | 259 | 1.2090 | 4.0225 | 1.2403 | 0.1519 | 0.1901 | 0.2926 |
| 13 | 259 | 1.2383 | 3.5076 | 1.2422 | 0.1458 | 0.1883 | 0.3084 |
| 14 | 259 | 1.1416 | 3.8357 | 1.2246 | 0.1492 | 0.1915 | 0.2999 |
| 15 | 259 | 1.1899 | 3.6677 | 1.2351 | 0.1435 | 0.1878 | 0.2984 |
| 16 | 259 | 1.1899 | 4.3825 | 1.2403 | 0.1451 | 0.1905 | 0.2829 |
| 17 | 259 | 1.1899 | 3.7829 | 1.2397 | 0.1483 | 0.1835 | 0.2941 |

표 2. 학습 데이터에 대한 실제 현장 데이터와 MLR, PLS 모델의 예측값.

Table 2. The actual plant data and predicted values by MLR, PLS model for trained data.

| No. | $F^v_{water}$ | | | $F^v_{EG}$ | | | $C^o_{acid}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | plant data | MLR | PLS | plant data | MLR | PLS | plant data | MLR | PLS |
| 1 | 0.1421 | 0.1340 | 0.1343 | 0.1835 | 0.1852 | 0.1854 | 0.3248 | 0.3243 | 0.3242 |
| 2 | 0.1435 | 0.1431 | 0.1430 | 0.1846 | 0.1846 | 0.1845 | 0.3237 | 0.3181 | 0.3180 |
| 3 | 0.1467 | 0.1554 | 0.1555 | 0.1893 | 0.1869 | 0.1869 | 0.3082 | 0.3085 | 0.3084 |
| 4 | 0.1668 | 0.1712 | 0.1711 | 0.1902 | 0.1884 | 0.1883 | 0.2948 | 0.2968 | 0.2967 |
| 5 | 0.2056 | 0.2054 | 0.2055 | 0.1890 | 0.1902 | 0.1903 | 0.2769 | 0.2765 | 0.2764 |
| 6 | 0.2117 | 0.2075 | 0.2074 | 0.1897 | 0.1904 | 0.1902 | 0.2747 | 0.2739 | 0.2739 |
| 7 | 0.1565 | 0.1556 | 0.1563 | 0.1907 | 0.1890 | 0.1895 | 0.2898 | 0.2897 | 0.2895 |
| 8 | 0.1408 | 0.1435 | 0.1434 | 0.1842 | 0.1880 | 0.1879 | 0.3062 | 0.3018 | 0.3018 |
| 9 | 0.1426 | 0.1439 | 0.1444 | 0.1842 | 0.1851 | 0.1855 | 0.3085 | 0.3131 | 0.3130 |
| 10 | 0.1565 | 0.1556 | 0.1563 | 0.1907 | 0.1890 | 0.1895 | 0.2898 | 0.2897 | 0.2895 |
| 11 | 0.1582 | 0.1582 | 0.1587 | 0.1868 | 0.1882 | 0.1886 | 0.2919 | 0.2930 | 0.2928 |
| 12 | 0.1519 | 0.1504 | 0.1491 | 0.1901 | 0.1902 | 0.1891 | 0.2926 | 0.2937 | 0.2939 |
| 13 | 0.1458 | 0.1443 | 0.1438 | 0.1883 | 0.1862 | 0.1857 | 0.3084 | 0.3122 | 0.3123 |
| RMSE | | 0.003859 | 0.003900 | | 0.001781 | 0.001844 | | 0.002688 | 0.002702 |

표 3. 학습되지 않은 새로운 데이터에 대한 실제 현장 데이터와 MLR, PLS 모델의 예측값.

Table 3. The actual plant data and predicted values by MLR, PLS model for nontrained data.

| No. | $F^v_{water}$ | | | $F^v_{EG}$ | | | $C^o_{acid}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | plant data | MLR | PLS | plant data | MLR | PLS | plant data | MLR | PLS |
| 14 | 0.1492 | 0.1489 | 0.1494 | 0.1915 | 0.1877 | 0.1881 | 0.2999 | 0.2988 | 0.2987 |
| 15 | 0.1435 | 0.1475 | 0.1475 | 0.1878 | 0.1870 | 0.1869 | 0.2984 | 0.3050 | 0.3049 |
| 16 | 0.1451 | 0.1551 | 0.1534 | 0.1905 | 0.1931 | 0.1915 | 0.2829 | 0.2804 | 0.2807 |
| 17 | 0.1483 | 0.1507 | 0.1504 | 0.1835 | 0.1881 | 0.1878 | 0.2941 | 0.2996 | 0.2996 |
| RMSE | | 0.005519 | 0.004726 | | 0.003279 | 0.002822 | | 0.004507 | 0.004438 |

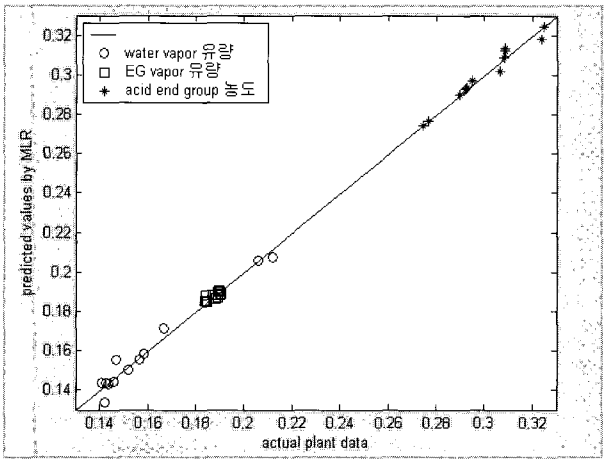



그림 2. 학습 데이터의 실제값과 MLR 모델로 구한 예측값의 비교.

Fig. 2. Comparison between the actual plant data and the predicted values by MLR for trained data.

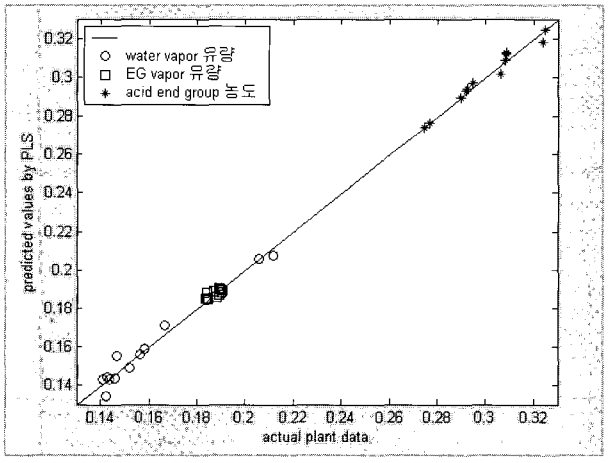



그림 3. 학습 데이터의 실제값과 PLS 모델로 구한 예측값의 비교.

Fig. 3. Comparison between the actual plant data and the predicted values by PLS for trained data.

마찬가지로 그림 3에서는 PLS 모델을 수립하기 위하여 사용된 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$의 실제 현장 데이터 값과 PLS 모델로 구한 예측값을 그래프를 이용하여 비교하였다.

표 2에서 $F^v_{water}$, $F^v_{EG}$에 대한 MLR, PLS 두 모델의 RMSE 값을 비교해볼 때 모델 수립에 사용된 학습 데이터에 대한 전반적인 회귀성능은 MLR 모델이 우수하다고 할 수 있으나, PLS 모델의 경우 잠재변수를 3개만 사용하였다는 점과 두 모델의 RMSE 값의 크기를 고려해보면 PLS 모델이 MLR 모델에 비해 회귀 성능이 매우 떨어진다고 할 수는 없다.

학습 데이터에 대한 회귀 성능과는 달리 학습되지 않은 새로운 데이터에 대한 예측 성능은 MLR 모델보다 PLS 모델이 더 우수하다. 학습 데이터로 구한 MLR 모델과 PLS 모델을 사용하여 학습에 사용되지 않은 표 1의 데이터 14,15

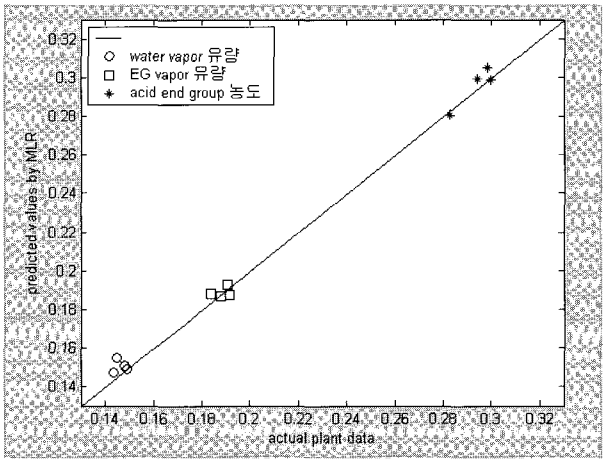



그림 4. 학습되지 않은 데이터의 실제값과 MLR 모델로 구한 예측값의 비교.

Fig. 4. Comparison between the actual plant data and the predicted values by MLR for nontrained data.

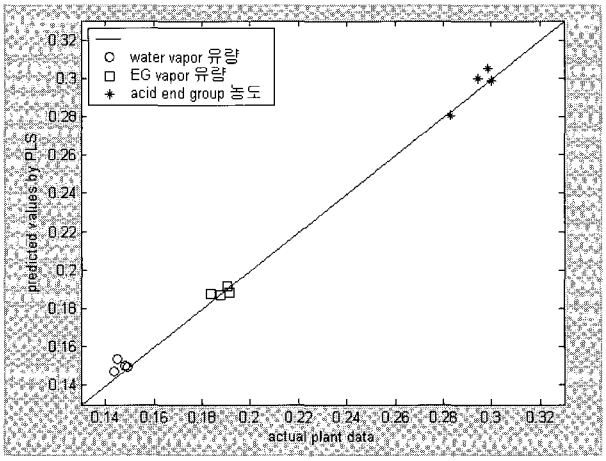



그림 5. 학습되지 않은 데이터의 실제값과 PLS 모델로 구한 예측값의 비교.

Fig. 5. Comparison between the actual plant data and the predicted values by PLS for nontrained data.

표 4. PLS 모델과 수학적 모사 결과[3]의 비교.

Table 4. Comparison between PLS model and mathematical model[3].

| | $F^v_{water}$ | | | $F^v_{EG}$ | | | $C^o_{acid}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| No. | plant data | 참고문헌 [3] | PLS | plant data | 참고문헌 [3] | PLS | plant data | 참고문헌 [3] | PLS |
| 14 | 0.1492 | 0.1475 | 0.1494 | 0.1915 | 0.1887 | 0.1881 | 0.2999 | 0.2985 | 0.2987 |
| 15 | 0.1435 | 0.1449 | 0.1475 | 0.1878 | 0.1879 | 0.1869 | 0.2984 | 0.3034 | 0.3049 |
| 16 | 0.1451 | 0.1475 | 0.1534 | 0.1905 | 0.1903 | 0.1915 | 0.2829 | 0.2876 | 0.2807 |
| 17 | 0.1483 | 0.1500 | 0.1504 | 0.1835 | 0.1887 | 0.1878 | 0.2941 | 0.2976 | 0.2996 |
| RMSE | | 0.001837 | 0.004726 | | 0.002955 | 0.002822 | | 0.003915 | 0.004438 |

에 대한 예측값을 계산하여 표 3에 나타내었다. 표 3에서의 RMSE 값을 비교해보면 학습되지 않은 새로운 데이터에 대한 3개의 출력변수 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$의 예측값은 PLS 모델로 구한 값이 MLR 보다 모두 우수한 것을 알 수 있다. 표 2와 표 3에 나타난 결과를 보면, 주어진 학습 데이터에 대한 회귀 성능은 MLR 모델이 우수하지만, 학습되지 않은 새로운 데이터에 대한 예측능력은 PLS 모델이 더 우수함을 알 수 있다. 이는 표 1로 주어진 입력변수 중 온도의 변화는 다른 변수의 변화에 비해 거의 없는데도 MLR모델은 모든 입력변수에 대해 과적합되었기 때문으로 설명될 수 있다.

그림 4는 학습되지 않은 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$의 실제 현장 데이터 값과 MLR 모델로 구한 예측값을 비교한 그래프이다. 마찬가지로 그림 5에서는 학습되지 않은 $F^v_{water}$, $F^v_{EG}$, $C^o_{acid}$의 실제 현장 데이터 값과 PLS 모델로 구한 예측값을 그래프를 이용하여 비교하였다.

본 논문에서 사용된 PET 중합공정 중 직접 에스테르화 반응기의 실제 현장 데이터는 참고문헌[3]에서 인용한 것으로서, 인용한 참고문헌에서는 직접 에스테르화 반응기에서 일어나는 모든 반응식과 반응식에 사용되는 반응상수들의 값을 이용하여 반응기 내부를 수학적으로 모델링 하였다. 본 논문의 PLS 모델의 예측 결과(표 3)와 참고문헌[3]의 모사 결과를 비교하기 위하여 표 4에 각각의 예측값과 RMSE 값을 나타내었다.

표 4에서 보는 바와 같이 수학적 모델링에 의한 출력변수 $F^v_{water}$의 예측 결과는 PLS 모델의 예측값보다 훨씬 우수한 것을 알 수 있다. 그러나 출력변수 $F^v_{EG}$, $C^o_{acid}$의 경우에는 PLS 모델에 의한 예측값이 수학적 모델링에 의한 예측결과와 비슷하거나 오히려 더 우수하다는 것을 알 수 있다. 따라서 화학반응식이 매우 복잡하고 화학반응식에 사용되는 반응상수를 정확하게 모르는 경우 수학적 모델링이 매우 어렵다는 것을 고려해 볼 때, 복잡한 화학공정의 출력변수를 예측하는 경우에 PLS 모델은 수학적 모델링의 훌륭한 대안이라고 할 수 있다.

## V. 결론

본 논문에서는 PET 중합공정에 사용되는 직접 에스테르화 반응기의 반응조건 중 반응율에 중요한 영향을 미치는 온도, 압력, 반응시간, 원료로 투입되는 TPA와 EG의 몰 비를 입력 변수로 하고, 반응기의 유출수의 유량과 유출수에 포함된 EG의 유량, 그리고 생산제품 내에 존재하는 acid end group의 농도를 출력 변수로 하는 실험적 모델을 만들기 위해 다변량 통계분석법의 대표적인 두 방법인 MLR과 PLS 알고리즘을 적용해 보았다. 참고문헌[3]에 나와 있는 직접 에스테르화 반응기의 실제 현장 데이터를 사용하여 MLR, PLS 두 모델의 회귀성능을 각각 비교해 보았으며, 학습 데이터로 만들어진 두 모델을 가지고 학습에 사용되지 않은 새로운 데이터에 대한 예측성능을 RMSE 값을 가지고 비교해 보았다.

주어진 학습데이터에 대한 회귀성능은 MLR 모델이 PLS 모델 보다 다소 우수하다는 것을 알 수 있었다. 그러나 학습되지 않은 새로운 데이터에 대한 예측 성능은 PLS 모델이 MLR 모델보다 훨씬 우수한 것을 알 수 있었다. 이는 MLR 모델의 경우 모든 입력변수들에 대해 과적합되었기 때문이다. 본 연구에서 PLS 모델을 수립하는데 사용된 잠재변수의 개수는 3개이다.

본 논문의 연구 결과와 참고문헌[3]에서의 연구 결과를 비교하였다. 참고문헌[3]에서의 연구 결과는 직접 에스테르화 반응기에서 일어나는 모든 반응식과 반응식에 사용되는 반응상수들의 값을 이용하여 반응기 내부를 수학적으로 모델링한 결과로서 참고문헌에서의 출력변수 $F^v_{water}$의 예측결과는 PLS 모델 보다 훨씬 우수한 것을 알 수 있었다. 그러나 출력변수 $F^v_{EG}$, $C^o_{acid}$의 경우에는 PLS 모델에 의한 예측값이 수학적 모델링에 의한 예측결과와 비슷하거나 오히려 더 우수하다는 것을 알 수 있었다. 따라서 화학반응식이 매우 복잡하고 화학반응식에 사용되는 반응상수를 정확하게 모르는 경우 수학적 모델링이 매우 어렵다는 것을 고려하면, 복잡한 화학공정의 출력변수를 예측하는 경우에 있어서 PLS 모델은 수학적 모델링의 훌륭한 대안이라고 할 수 있다.

## 참고문헌


[1] H. S. Song, Y. D. Park and J. C. Hyun, "Optimization for the minimum reaction time of PET esterification," *Korean J. of Chem. Eng.*, vol. 13, no. 4, pp. 369-378, 1996.

[2] 김청엽, 조현남, "폴리에틸렌테레프탈레이트(PET) 제조기술," 고분자과학과 기술, 제3권, 제2호, pp. 77-84, 4월, 1992.

[3] J-Y Kim, H-Y Kim and Y-K Yeo, "Identification of kinetics of direct esterification reactions for PET synthesis based on a generic algorithm," *Korean J. of Chem. Eng.*, vol. 18, no. 4, pp. 432-441, 2001.

[4] K. Ravindranath and B. A. Mashelkar, "Modeling of poly(ethylene terephthalate) reactors: a continuous esterification process," *Polym. Eng. Sci.*, vol. 22, no. 10, pp. 610-618, 1982.

[5] T. Yamada, "A mathematical modeling for a continuous esterification process with recycle between terephthalate acid and ethylene glycol," *Journal of Applied Polymer Science,* vol. 45, no. 11, pp. 1919-1936, Aug., 1992.

[6] C. K. Kang, B. C. Lee, D. W. Ihm and D. A. Tremblay, "A simulation study on continuos direct esterification process for poly(ethylene terephthalate) synthesis," *Journal of Applied Polymer Science,* vol. 63, no. 2, Jan., 1997.

[7] L. Eriksson, J. L. M. Hermens, E. Johansson, H. J. M. Verhaar and S. Wold, "Multivariate analysis aquatic toxicity data with PLS," *Aquatic Science,* vol. 57, pp. 217, 1995.

[8] J. F. MacGregor, C. Jaeckle, C. Kiparissides and M. Koutoudi, "Process monitoring and diagnosis by multiblock PLS methods," *AIChE J.*, vol. 40, no. 5, pp. 826-838, 1994.

[9] T. Suzuki, K. Ohtaguchi and K. Koide, "Computer-assisted approach to develop a new prediction method of liquid viscosity of organic compounds," *Computers Chem. Engng,* vol. 20, no. 2, pp. 161-173, 1996.

[10] S. Qin, "Recursive PLS algorithms for adaptive data modeling," *Computers Chem. Engng,* vol. 22, no. 4/5, pp. 503-514, 1998.

[11] 홍선주, 허창구, 한종훈, "PLS 방법을 이용한 증류 공정의 국부적인 조성 추정 소프트센서," 화학공학, 제37권, 제3호, pp. 445-452, 1999.

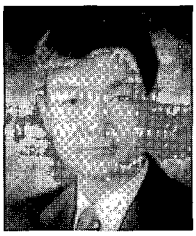
[12] I-S Han, M. Kim, C-H Lee, W. Cha, B-K Ham, J-H Jeong, H. Lee, C. B. Chung, "Application of partial least square methods to a terephthalic acid manufacturing process for product quality control," *Korean J. of Chem. Eng.*, vol. 20, no. 6, pp. 977-984, 2003.

[13] P. Geladi, B. R. Kowalski, *"Partial least-squares regression : a tutorial,"* Analytica Chemica Acta, (1986).

[14] S. de Jong, "SIMPLS : an alternative approach to partial least squares regression," *Chem. Intell. Lab. Sys.*, 18, 251-263 (1993).

**김 성 영**

1979년 6월 8일생. 1999년 경희대학교 환경응용화학부(공학사). 2003년(공학석사). 2003년~현재 경희대학교 환경응용화학부 박사과정. 관심분야는 공정시스템.

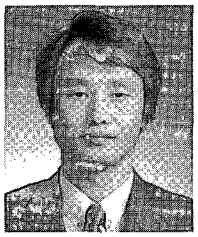


**이 범 석**

1961년 3월 22일생. 1984년 서울대학교 화학공학과(공학사). 1986년 (공학석사). 1992년 Purdue University (Ph.D.). 1993년~현재 경희대학교 환경응용화학부 교수. 관심분야는 공정시스템.

**정 창 복**

1956년 3월 2일생. 1978년 서울대학교 화학공학과(공학사). 1982년 한국과학기술원 화학공학과(공학석사). 1988년 University of Michigan (Ph.D.). 1980년~현재 전남대학교 응용화학공학부 교수. 관심분야는 공정시스템.



**최 수 형**

1961년 1월 10일생. 1984년 서울대학교 화학공학과(공학사). 1986년(공학석사). 1990년 University of Missouri-Rolla (Ph.D.). 1993년~현재 전북대학교 화학공학부 교수. 관심분야는 공정시스템.