과실류의 비파괴 내부품질 인자의 예측 성능 향상을 위한 VIS/NIR 스펙트럼 전처리 기법 개발

Pre-Processing Techniques on VIS/NIR Spectral Data for Non-Destructive Quality Evaluation of Fruits

류동수† 황인근* 노상하†

정회원 정회원 정회원

D. S. Ryu I. G. Hwang S. H. Noh

1. 서론

최근 국민소득의 향상으로 보다 고품질의 과일을 소비하고자 하는 성향이 나타나고 있으며, 이는 고품질의 과일을 생산하기 위한 방법 뿐만 아니라, 고품질의 과실을 선별해 내는 기술의 개발을 요구하고 있다. 또한, 수출을 위해서는 수출대상국가에서 요구하는 규격에 맞추어야 수출이 가능한 실정이다. 따라서 국내 및 국외에서 VIS/NIR 분광분석법을 이용한 과실류의 내부 품질(당도, 산도, 경도 등)을 비파괴적인 방법으로 측정해 내고자 하는 많은 시도가 있었다.[1,2,3,4,5]

과실류의 내부 품질을 비파괴적으로 측정하는 시스템을 개발하고 이의 성능을 향상시키기 위해서는 하드웨어 및 소프트웨어적인 측면으로 나누어 생각할 수 있다. 전자는 스펙트럼을 측정하기 위한 기기장치의 구성을 적절히 설계하는데 있으며, 후자는 스펙트럼 측정시의 구조적인 외란(광산란, 광경로의 변화, 측정장치 자체의 노이즈, 측정환경의 변화 등)을 제거하고 보다 안정적인 모델을 만드는 것이다. 본 연구의 목적은 모델 개발 및 예측성능을 향상시키기 위하여 VIS/NIR 스펙트럼 데이터의 전처리기법을 개발하고, 이를 각 시료 및 스펙트럼 측정방법에 대해 적용하여,

1) 투과 스펙트럼에 대한 전처리 효과를 구명하고, 광산란 및 성분에 독립적인 스펙트럼 성분의 제거효과를 구명하며,

2) 내부품질 인자 예측모델의 성능 향상을 위한 전처리 기법을 개발하는데 있다.

2. 재료 및 방법

가. 공시 재료

공시재료는 1999년산 사과를 대상으로 하였다. 예산 및 김천 지방에서 수확시기를 달리하여 직접 과수원에서 수확한 것을 실험에 사용하였다. 실험 시 상온에서 약 1일간 방치한 후, 품온을 실온으로 맞추어 VIS/NIR 스펙트럼을 측정하고 당도 및 산도를 측정하였다.

나. 내부 품질인자의 측정

과실류의 내부품질인자 중, 당도측정은 디지털 굴절당도계(Model DBX-55, ATAGO, Japan)를 이용하였으며, 산도측정은 자동적정산도측정장치(AUT-301L, Japan)를 이용하였다.

Table 1. Measurements of internal qualities

Data set	Internal qualities	Range	Statistics		
			Mean	Std.	Var.
Apple	Sugar	8~14.9	12.4	1.3913	1.9357
	Acidity	0.1938~0.5975	0.2983	0.0571	0.0033

다. 스펙트럼 측정

스펙트럼 전처리의 효과를 확인하기 위해 온라인 당도판정장치로 제작된 투과광 스펙트럼측정장치[5]를 사용하여 측정하였다.

† 서울대학교 생물자원공학부 농업기계전공

* 농양물산(주) 중앙연구소

Table 2. Measurements of Spectra

Data Set	Apple
Type of Measurement	Transmittance
Wavelength	550~1050nm
Interval	1.8nm
No. of Cal. Data	80
No. of Pred. Data	40

라. 스펙트럼 전처리

VIS/NIR 스펙트럼의 측정은 외란의 영향을 많이 받게 되며, 온라인 측정장치로 스펙트럼을 측정할 경우는 특히 더 많은 요인들이 외란으로 작용하게 된다. 스펙트럼 전처리는 다중 모델의 개발에 있어서 가장 기초적인 단계이며, 스펙트럼에 포함된 외란을 제거하여 보다 안정적인 내부품질인자 예측 모델을 만들기 위하여 이용된다. 스펙트럼 측정시 구조적인 외란으로는 특히 과일과 같은 고형물체의 경우, 표면에서의 광산란이나 측정방식에 기인한 광경로의 차이, 측정센서에서의 노이즈, 주변 측정환경의 변화(온도 등)가 있으며, 이러한 외란은 측정된 스펙트럼의 변이에 주요한 영향을 미치게 된다. 따라서 예측하고자 하는 내부 품질인자의 농도 정보가 포함된 스펙트럼 영역에 이러한 변이가 나타나게 되면, 성분의 농도와 상관이 없는 스펙트럼의 변이는 모델링을 할 때 방해가 되며, 모델의 예측성능을 저하시키는 주요한 원인이 된다.

스펙트럼의 주요 전처리 방법은 평활화(Smoothing), MC, SNV, MSC, OSC 등이 있으며, 각 전처리 방법의 효과는 스펙트럼 측정장치 자체의 노이즈를 제거하는데 평활화를, 1·2차 미분은 광경로의 차이나 측정환경의 변화 등에 기인한 baseline의 이동을 제거하는데 이용되고 있다. 그리고 MSC나 SNV는 스펙트럼의 측정시 광산란의 영향을 제거하는데 이용되고 있다. OSC는 최근에 나온 개념으로 기존의 전처리 방법은 스펙트럼 자체의 변형이나 개선에 노력한 반면에, 이것은 스펙트럼의 변형에 분석 대상 성분의 농도를 고려한 스펙트럼 보정 방법이다.

주요 스펙트럼 전처리의 알고리즘은 다음과 같다.

(1) MSC(Multiplicative Scattering Correction)

MSC는 시료 표면 혹은 시료내부의 불균일성 때문에, 동일 시료에 대한 반복 측정에 대해서도 얻어진 스펙트럼은 차이를 보일 수 있으며, 이는 스펙트럼 데이터 변이의 가장 큰 원인이 된다. 또한 산란정도는 사용된 광원의 종류나 시료표면의 상태, 그리고 시료의 반사지수 등에 영향을 받으며, 이것은 주로 베이스라인의 이동이나 기울기 및 곡률의 변화로 나타난다. 이러한 광산란의 영향은 특히 측정된 스펙트럼의 장파장 영역에서 주로 나타난다. MSC의 기본개념은 모든 스펙트럼을 이상적인 스펙트럼에 의해 보정하는 것이다. 실제로는 이상적인 스펙트럼을 얻을 수 없으므로, 전체 스펙트럼의 평균을 이상 스펙트럼으로 한다. 따라서, 이상적인 스펙트럼은

$$\overline{x_j} = \frac{\sum_{i=1}^{m} x_{i,j}}{m}$$

여기서, x : 스펙트럼 (m×n),

$\overline{x_j}$: j번째 파장의 평균 스펙트럼(1×n)

이 되고, 이 평균 스펙트럼을 이용하여 각 파장에서의 흡광도 데이터에 대해 선형회귀를 취한다. 즉,

$$x_i = a_i\overline{x} + b_i$$

여기서, x_i : 평균에 대한 회귀스펙트럼 (1×n),

a_i : 기울기, b_i : 절편

선형회귀를 하여 구한 a_i및 b_i값을 이용하여 다음 식과 같이 MSC 보정을 한다.

$$x_{i_{MSC}} = \frac{(x_i - b_i)}{a_i}$$

여기서, $x_{i_{MSC}}$: MSC 보정된 스펙트럼 (1×n)

(2) SNV(Standard Normal Variate)

이것은 MSC와 마찬가지로 광산란 보정을 위한 방법으로, 목적은 동일하지만 수학적인 방법은 다르다. SNV는 이상적인 스펙트럼이 필요하지 않으며, 대신에 각 스펙트럼을 전체 스펙트럼의 표준편차로 정규화하여 광산란의 영향을 제거하는 방법이다. 광경로나 광원의 변동에 따른 스펙트럼의 변이등이 SNV에 의해 보정될 수 있으며, SNV보정된 스펙트럼은 무차원이 된다.

$$\overline{a_i} = \frac{\sum_{j=1}^{n} A_{i,j}}{n}$$

$$x_{i_{SNV}} = \frac{(x_i - \overline{a_i})}{\sqrt{\frac{\sum_{j=1}^{n}(x_{i,j} - \overline{a_i})^2}{(n-1)}}}$$

여기서, a_i : i번째 스펙트럼의 모든 파장에 대한 평균 (n×1)

$x_{i_{SNV}}$: SNV 보정된 스펙트럼 (1×n)

(3) OSC(Orthogonal Sigal Correction)

전처리를 하는 이유는 농도 정보와 상관이 없는 구조적인 변이를 제거하기 위한 것이다. 하지만 이러한 전처리는 수학적으로 일종의 필터링의 개념이며, 성분 y는 고려하지 않고 있다. 따라서 스펙트럼에서 구조적인 노이즈를 제거할 뿐만 아니라, PLS 모델링을 할 경우, 스펙트럼의 분산이 큰 부분만 강조하게 되어 성분과 관련된 정보도 동시에 제거하는 경향이 있다. Wold(1998)등이 스펙트럼(x)에서 성분(y)과 상관이 없는 변이만 제거하면 성분과 상관이 높은 스펙트럼 성분만 남는다는 개념을 제안하고, PLS에서 계산되는 스코어(t)를 성분 y와 직교시켜 이를 이용하여 성분에 직교하는 스펙트럼 성분을 계산하여, 이를 원래의 스펙트럼에서 제거하면 성분과 상관이 매우 높은 스펙트럼 성분만을 찾을 수 있다고 보고하였다[6]. 즉 성분 Y와 스펙트럼 X가 서로 직교하기 위해서는 PCA의 기본식 및 PLS의 NIPALS 알고리즘에서,

$$x = tp' + e$$

여기서 x : 스펙트럼 행렬(m×n)

y : 성분 벡터(m×1)

t : 스코어벡터(m×1)

p' : 로딩벡터(n×1),

e : 잔차행렬(m×n)

m : 샘플의 수

n : 파장의 수

t와 y의 내적이 0($t' \cdot y = y \cdot t' = 0$)이 되면 직교하게 된다. 즉,

$$t* = (1 - y(y'y)^{-1}y')t$$

따라서, 이 $t*$를 이용하여 성분 y와 상관이 없는 스펙트럼성분($t* \times p'$, OSC component)을 계산할 수 있다. 이를 원 스펙트럼에서 제거하면 상관이 높은 스펙트럼 성분을 얻을 수 있다.

$$x = x - t* \times p'$$

OSC Component 수는 성분과 직교하는 스펙트럼성분($t* \times p'$)을 몇번 제거할 것인가를 나타낸다.

(4) 전처리 조합

각 전처리의 효과를 보기 위해 Table 3과 같이 사과 및 배의 데이터 셋에 대해 각 전처리 및 그 조합을 처리하였다. 여기서 원스펙트럼에 대하여 기본적으로 스펙트럼의 평활화 및 MC (Mean Centering)을 하였으며, 평활화는 Savitzky-Golay Smoothing을 이용하여, 파장간격 27nm, polynomial order를 1차로 하여 평활화하였다.

스펙트럼 데이터의 전처리 및 PLS를 구현하기 위하여 Matlab (ver. 5.3)을 이용하였다.

Table 3. Correction methods used for the prediction of internal qualities

Method name	Signal correction method
N	No Preprocessing
S	SNV
S1	1st Deriv. + SNV
M	MSC
M1	1st Deriv. + MSC
O	OSC
O1	1st Deriv. + OSC

마. 모델 개발

VIS/NIR 스펙트럼을 이용한 당도 측정시 모델 개발 방법은 PLS모델을 사용하였다. PLS 모델의 특징은 PCR모델과는 달리, 농도정보를 이용함으로써, 고농도의 스펙트럼은 저농도의 스펙트럼보다 큰 가중치를 갖게 하는 방법이다.

최적 PLS Factor의 결정방법은 Cross - validation (leave-one out)에 의한 최소 PRESS 값에서의 Factor수로 하였다.

OSC 전처리가 들어있는 모델의 경우는 OSC Component의 수를 다음과 같이 각 OSC Component를 수행한 후, 제거된 스펙트럼의 분산과 원스펙트럼의 분산의 비를 계산하였으며, 이 값이 일정하게 될 때(≤0.01)의 OSC Component 수를 적정 Components 수로 결정하였다.[7]

$$C_i^2 = \frac{var(x_i)}{var(x)}$$

바. 모델성능 평가 기준

개발된 모델의 평가를 위해 SEC, SEP 및 bias를 다음과 같이 계산하였다.

$$SEC = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m - f - 1}}$$

$$bias = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)}{m}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{m}((y_i - \hat{y}_i) - bias)^2}{m-1}}$$

각 전처리 및 그 조합에 의해 개발된 모델의 성능 평가를 위하여 다음과 같은 기준을 고려하였다.

① SEC 및 SEP가 작을 것
② 예측모델의 bias가 작을 것
③ 개발된 모델의 R^2와 검정시료의 R^2가 클 것
④ 모델에 포함된 PLS factor의 수가 작을 것

3. 결과 및 고찰

가. 각 전처리별 스펙트럼 형태의 변화

실시간용 측정장치인 투과광 스펙트럼 측정장치는 여러 가지 노이즈 성분이 들어갈 가능성이 충분한 것으로 판단된다. 따라서 표에 제시된 전처리 방법 및 그 조합으로 전처리를 수행하였다. 전처리를 수행한 결과, 각 스펙트럼의 형태는 Fig. 1~4와 같이 변화하였다.

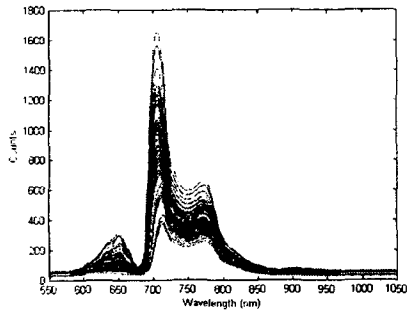

Fig. 1 Raw Spectrum

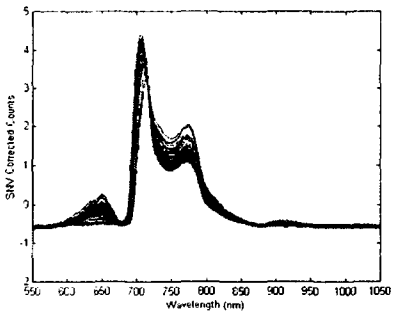

Fig. 2 SNV Corrected Spectrum

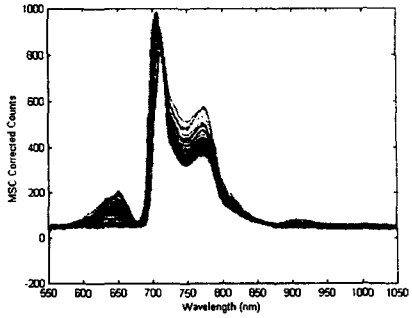

Fig. 3 MSC Corrected Spectrum

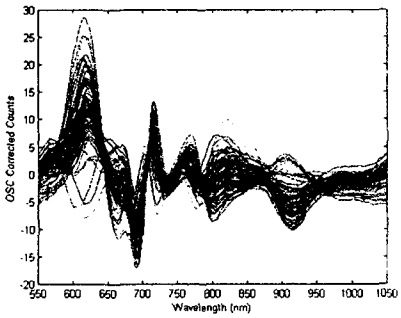

Fig. 4 OSC corrected Spectrum
(OSC Comp. #=5)

MSC나 SNV 보정을 한 경우는 스펙트럼 형태가 유사함을 알 수 있으며, 또한 850nm이상의 영역 데이터는 거의 사라지게 되어 이 영역의 정보를 이용하기가 어렵다는 것을 알 수 있다. OSC의 경우는 각 파장에서의 스펙트럼이 상대적으로 강조되어 이 영역의 데이터가 모델링에 사용될 수 있다.

나. OSC Component의 결정

OSC의 경우, OSC component의 수를 결정하기 위해 C_i^2를 계산하고 이 값이 0.01이하일때의 최소 OSC component의 수를 최적으로 하였다. Fig. 은 원데이터에 대해 평활화만 한 후, C_i^2값의 변화를 나타낸 것이다. 이 경우는 최적 OSC component=5가 되며, 이 값을 이용하여 OSC처리를 하였다.

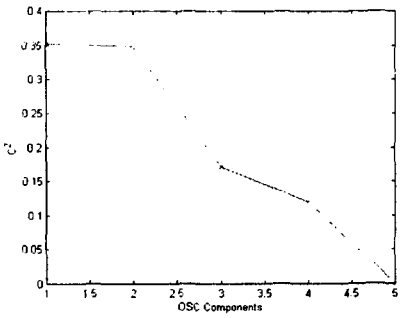

Fig. 5 Determination of No. of OSC components

다. 각 전처리별 모델 개발 결과

온라인용 투과광 스펙트럼 측정장치에서 획득한 스펙트럼 데이터를 이용하여 각 전처리별 품질인자 예측모델을 만들고, 이의 전처리 효과를 분석하였다. Table 4에 각 전처리별 모델의 예측결과를 제시하였다.

Table 4 . Results of Prediction

Methods	Factor	R^2 Cal.	SEC	R^2 P	SEP	bias
N	12	0.8698	0.6008	0.2537	1.657	0.0546
S	12	0.8733	0.5616	0.7491	0.639	0.1465
S1	13	0.8852	0.5385	0.7641	0.617	0.1371
M	15	0.9023	0.4969	0.8561	0.489	0.1713
M1	12	0.8830	0.5397	0.7564	0.631	0.1651
O	13	0.8368	0.6421	0.7445	0.645	0.0260
O1	10	0.7214	0.8205	0.5585	0.885	0.0967

여기서 전처리 방법 O는 OSC #=2, O1은 OSC #=5인 경우이다.

Fig. 6은 각 전처리별로 개발된 모델의 예측결과 비교이다. 전처리 및 그 조합에서 최소 SEP를 갖는 것은 MSC를 수행했을 경우이다. 이때의 SEP=0.489, bias=0.1713이었다.

S, S1, M1, O의 경우는 SEP가 거의 유사하였다. 이중에서 O는 SEP=0.645이었으나, 각 전처리 방법 중 가장 작은 bias(=0.0260)를 보였다. 하지만 Factor수는 13로서 다소 많은 Factor가 사용되고 있다.

Fig. 7은 각 전처리별 PLS Factor수를 비교한 것이다. 그림에서 가장 작은 Factor수는 1차미분과 OSC를 동시에 처리했을 경우로 이때의 PLS Factor수는 10이었으며, SEP=0.885, bias=0.0967이었다. 하지만 R^2가 낮아져 예측성능에 문제가 있는 것으로 판단된다. 또한 가장 낮은 SEP를 보이는 MSC의 경우는 Factor수가 많아져(=15), 성능평가기준에 위배된다.

Table 3 이외의 조합으로, OSC 전처리와 SNV, MSC 혹은 1차미분을 동시에 처리하는 경우는 예측성능이 크게 향상되지는 않았다. 이것은 본 연구에서 사용된 스펙트럼 데이터의 특성에 기인한 것으로 생각된다. Fig. 1의 원스펙트럼을 보면, 각 스펙트럼 사이의 분산이 작으므로, MSC나 SNV만으로도 충분한 외란 제거효과가 있으며, OSC처리만의 결과와 거의 유사할 것임을 알 수 있다.

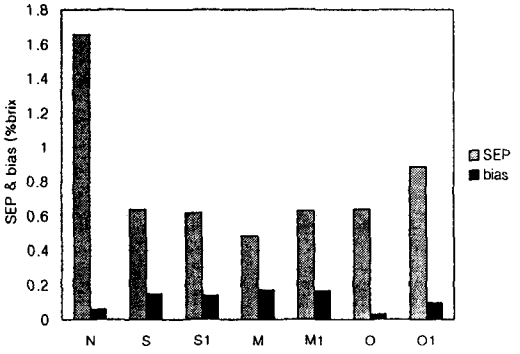

Fig. 6 Comparison of SEP and bias between preprocesiing methods

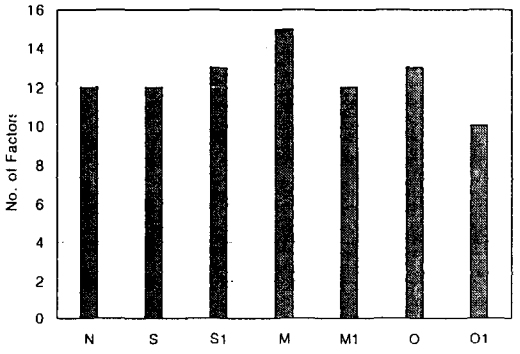

Fig. 7 Comparison of PLS factor between preprocesiing methods

따라서 성능평가 기준에 따라 적합한 모델은 OSC처리만 수행한 것이 모델의 예측 안정성 측면에서 선택될 수 있다. Fig. 8 및 Fig. 9는 OSC Component #=2, OSC전처리를 수행했을 경우의 Press 곡선과 보정 및 검증용 데이터를 표시한 그래프이다.

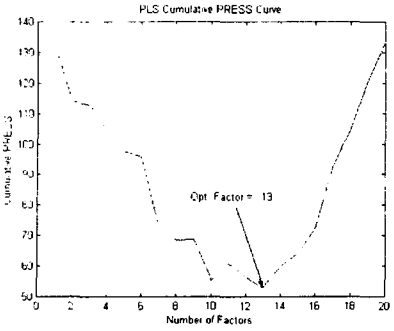

Fig.8 Cumulative press curve for optimal model (O, OSC#=2)

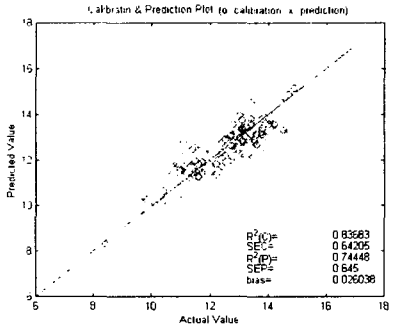

Fig. 9 Plot of predicted vs. actual value of calibration and prediction

4. 요약 및 결론

본 연구는 실시간 투과 스펙트럼장치에서 측정된 VIS/NIR 스펙트럼을 이용하여, 내부품질인자 중 당도 예측모델의 개발에서 스펙트럼 전처리의 효과 및 최적 전처리 방법을 구명하고자 수행되었다. 결과를 요약하면 다음과 같다.

가. 투과광 스펙트럼측정장치[5]를 이용하여 측정한 투과 스펙트럼에 대하여 각 전처리 및 그 조합을 수행한 결과, 최소 SEP를 보이는 전처리는 MSC를 한 경우로 SEP = 0.489 brix, bias = 0.1713이었다.

나. 모델 성능 평가기준을 고려하면, 본 연구에서 분석된 스펙트럼 데이터의 경우, SEP값은 MSC에 비해 다소 높으나, OSC처리를 한 것이 SEP=0.6421, bias=0.0260으로 모델의 안정성 측면에서 유리한 것으로 판단된다.

다. OSC 전처리와 SNV, MSC 혹은 1차미분을 동시에 처리하는 경우는 예측성능이 크게 향상되지는 않았다. 이것은 데이터의 특성에 기인하는 것이며, 본 연구에서 사용된 데이터의 경우, 측정된 스펙트럼의 분산이 작기 때문인 것으로 생각된다. 만약 측정된 스펙트럼의 분산이 매우 큰 샘플의 경우, OSC처리의 효과가 클 것으로 기대된다.

5. 참고문헌

1. 김우기, 1997. 분광반사특성을 이용한 주요 과실의 비파괴 당산도측정. 서울대학교 대학원 석사학위논문.
2. 이강진 외 4인, 1997. 근적외선을 이용한 사과의 당도예측모델 개발과 비교. 한국농업기계학회지 22(1):206-212
3. 최창현 외 2인, 1997. 가시광선/근적외선 분광분석법을 이용한 사과의 당도 및 경도측정. 한국농업기계학회지 22(1):200-205
4. 노상하 외 2인. 1999. 수출용 배의 부가가치 향상을 위한 선별포장시스템 개발에 관한 연구, 농림부 농림기술개발사업 1년차 연구보고서.
5. 황인근, 2000. VIS/NIR 실시간 분광 스펙트럼에 의한 후지 사과의 당산도 선별 시스템 개발, 서울대학교 박사학위논문.
6. S. Wold, H. Antti, F.Lindgren, J.Ohman, Orthogonal signal correction of near-infrared spectra, Chemometrics and Intelligence Laboratory System 44(1998) 175-185
7. J.Sjoblom, O. Svensson, M. Josefson, An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra, Chemometrics and Intelligence Laboratory System 44(1998) 229-244
8. B.M.Wise, PLS_Toolbox 2.0 manual, Eigenvector Reseach,1998
9. PLSplus/IQ manual, Galactic Industries Corp.,1996