Universitat Autònoma de Barcelona

# Virtualization and real-time analysis of pharmaceutical and food products by near infrared spectroscopy

Dong Sun

Tesi doctoral

Programa de Doctorat de Química

Directores: Dr. Manel Alcalà Bernàrdez

Prof. Dr. Marcelo Blanco Romia

Departament de Química

Facultat de Ciències

2017

### 9.3.1.2 Spectral band selection

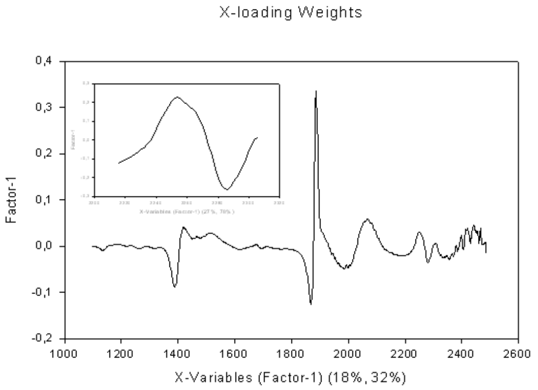
At first, models were calculated with the whole spectral range which was pretreated from 1100 to 2498 nm. The X loading weights were taken as the reference to determine the important spectral bands. Models in Table 9.2 were constructed with bands which had high X loading weights, and the performance of them was better than models in Table 9.1.

***Table 9.2 Performance of calibration model built with selected spectral band. The best pretreatment for every parameter has been marked with the * symbol.***

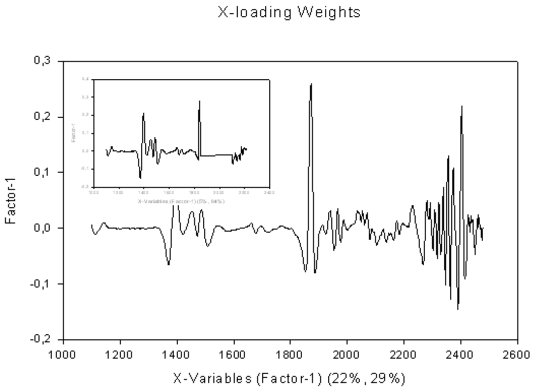
| **Parameter** | **Spectral band** | **Pretreatment** | **Calibration** | | | |
|---|---|---|---|---|---|---|
| | | | Slope | Correlation | RMSEC | Factor |
| Fructose in juice | 2212-2310nm | *SG and 1st D, order 2, 5 points | 0.9735 | 0.9867 | 0.4105 (g/L) | 7 |
| | | SG and 2nd D, order 2, 5 points | 0.9698 | 0.9848 | 0.4385 (g/L) | 7 |
| | | MSC and 1st D, gap size 5, segment size 2 | 0.9566 | 0.9781 | 0.5258 (g/L) | 7 |
| Glucose in juice | 2212-2310nm | SG and 1st D, order 2, 5 points | 0.9486 | 0.974 | 0.5874 (g/L) | 7 |
| | | *SG and 2nd D, order 2, 5 points | 0.9631 | 0.9814 | 0.4978 (g/L) | 7 |
| | | MSC and 1st D, gap size 5, segment size 2 | 0.9610 | 0.9803 | 0.5117 (g/L) | 7 |
| Fructose in puree | 1100-1850nm, 2094-2228nm | SG and 1st D order 2, 11 points | 0.9383 | 0.9687 | 0.1507 (g/100g) | 3 |
| | | *SG and 2nd D order 2, 11 points | 0.9405 | 0.9698 | 0.1481 (g/100g) | 3 |
| | | MSC and 2nd D, gap size 3, segment size 3 | 0.9156 | 0.9569 | 0.1763 (g/100g) | 3 |
| Glucose in puree | 1100-1850nm,2094-2228nm | SG and 1st D order 2, 11 points | 0.9614 | 0.9805 | 0.1271 (g/100g) | 4 |
| | | *SG and 2nd D order 2, 11 points | 0.9649 | 0.9823 | 0.1201 (g/100g) | 4 |
| | | MSC and 2nd D, gap size 3, segment size 3 | 0.9501 | 0.9747 | 0.1445 (g/100g) | 4 |
| Brix in | 1100- | SG and 1st D order 2, | 0.9497 | 0.9745 | 0.3148 | 4 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| puree | 1834nm,2120-2320nm | 11 points |  |  | (°Bx) |  |
|  |  | *SG and 2nd D order 2, 11 points | 0.9567 | 0.9781 | 0.2921 (°Bx) | 4 |
|  |  | MSC and 2nd D, gap size 3, segment size 3 | 0.9558 | 0.9777 | 0.2948 (°Bx) | 4 |
| Dry matter in puree | 2200-2340nm | *SG and 1st D order 2, 11 points | 0.9499 | 0.9746 | 0.3503 (g/100g) | 3 |
|  |  | SG and 2nd D order 2, 11 points | 0.9181 | 0.9582 | 0.448 (g/100g) | 3 |
|  |  | MSC and 2nd D, gap size 3, segment size 3 | 0.9478 | 0.9736 | 0.3576 (g/100g) | 3 |

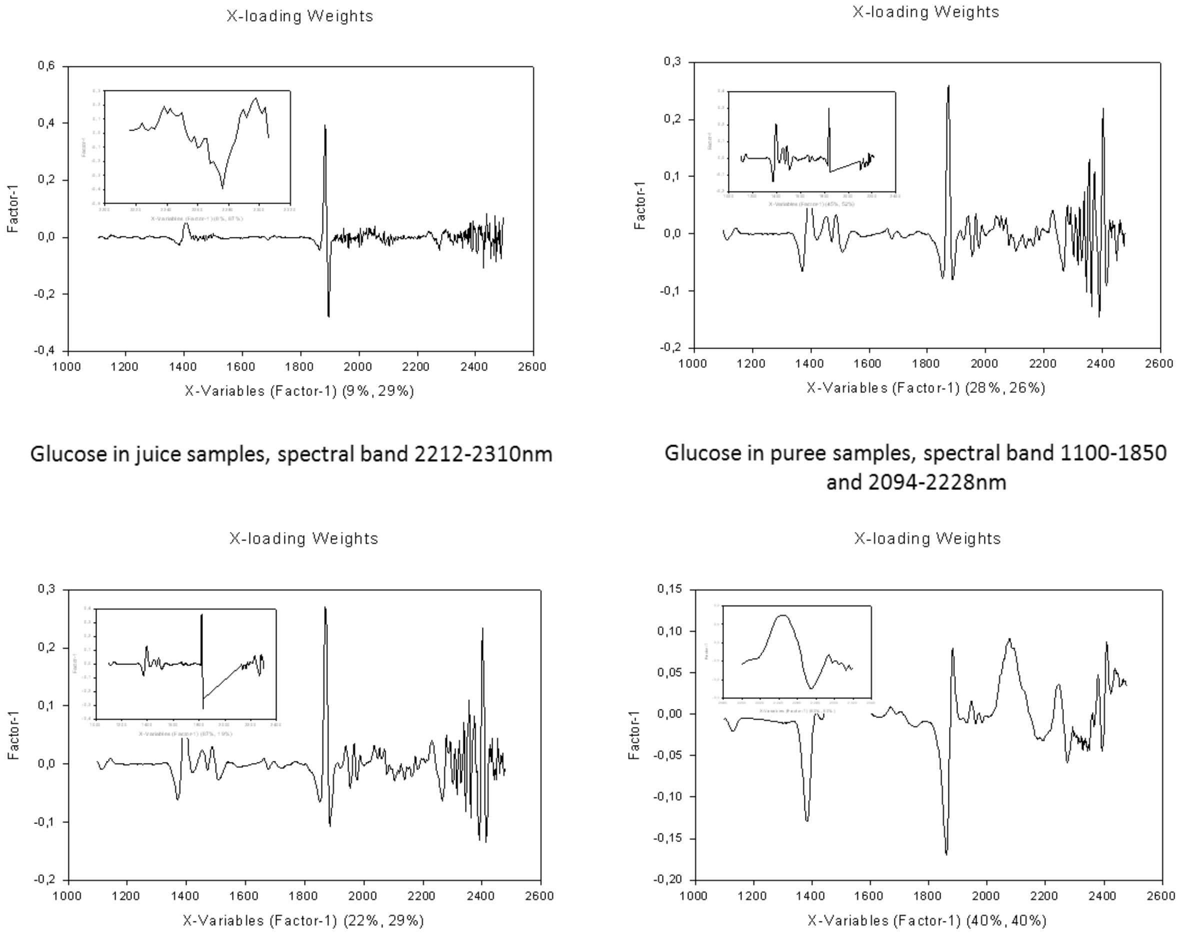
In Fig.9.4, we can see important bands which were selected for each parameter. For fructose and glucose content in juice samples, the peaks of water and the noise tail of spectra were removed. Spectral band from 2212 to 2310 nm was selected to calculate models. Besides, the spectral band of dry matter content was from 2200-2340 nm, which removed peaks of water and the noise tail neither. This was because the water absorbance peaks and the noise had a strong influence on their calibration models, which made the value of slop, correlation drop down and RMSEC go up.



Fructose in juice samples, spectral band 2212-2310nm



Fructose in puree samples, spectral band 1100-1850 and 2094-2228nm



Glucose in juice samples, spectral band 2212-2310nm

Glucose in puree samples, spectral band 1100-1850 and 2094-2228nm

Brix in puree samples, spectral band 1100-1834 and 2120-2320nm

Dry matter in puree samples, spectral band 2200-2340nm

***Fig.9.4 The X loading weights of calibrations within the selected band and whole range.***

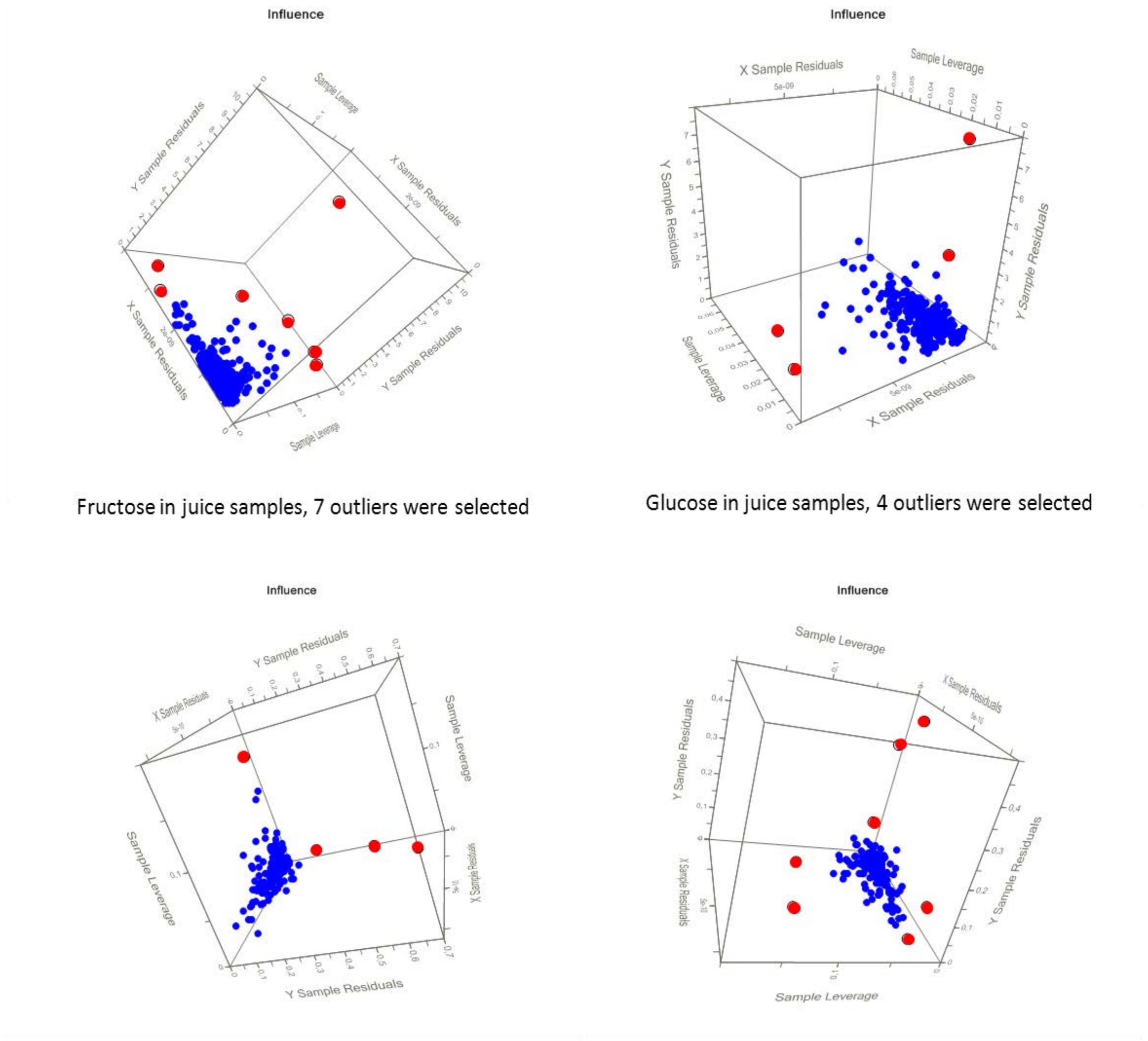
However, for soluble solids, fructose and glucose content in puree samples, we only excluded the noise tail and the second main peak of water which around 1830-2100 nm. Band around 1360-1500 nm can be taken into consideration when we analyze the puree, because the concentration of water in puree was much lower than in juice. Instead of decreasing the quality of calibration, these peaks added some useful information into models, for example the signal of the overtone of O-H in fructose and glucose molecular structure.

Compared to models from Table 9.1, slope of all models in Table 9.2 has reached the value more than 0.90, correlation was over 0.95 and the RMSEC has decreased for each PLS model. After the spectral band selection procedure, the quality of calibrations has been improved. However, these models were still not the most optimal choices because there were some outliers in the calibrations.

### 9.3.1.3 Outliers

Outliers may be caused by the spectra measurement or may indicate experimental error. In order to improve the quality of calibration model, samples which have high residual and leverage value must be excluded from the model. Quantitative models of fructose, glucose were recalculated without outlier samples. But for the soluble solids and dry matter content, the sample residuals did not decrease the quality of calibration models, so the whole sample set was adopted.

Fig.9.5 is a sample influence plot. It is a 3D plot of the residual X- and Y-variances vs. leverages. The sample which has an abnormally high value of them could be an outlier for the calibration model. Red points in Fig. 9.5 which were located far away from the main cluster had a high value of X- and Y-variances residuals or samples leverage. Therefore, they were outlier samples. In total, there were seven outliers for fructose in juice samples, four outlies for glucose in juice samples, four outliers for fructose in puree samples and seven outliers for glucose in puree samples. Those samples have been removed from calibration set of each parameter.



Fructose in juice samples, 7 outliers were selected

Glucose in juice samples, 4 outliers were selected

Fructose in puree samples, 4 outliers were selected

Glucose in puree samples, 7 outliers were selected

***Fig. 9.5 Outliers in the calibrations of fructose and glucose.***

#### 9.3.1.4 Calibration and validation models

After steps mentioned above, final calibrations were constructed by PLS. In order to ensure the quality of models, they need to be validated. 60 samples were selected as validation set for every model. These samples spanned a range which could cover the concentration range of samples in calibration.

The performance of quantitative models was evaluated by slope, correlation, $r^2$, RMSEC and RMSEP. The high value of slope, correlation resulted in a low number of RMSEC. But if the value of slope is too high there may be a risk of overfitting in the calibration, so the validation set is necessary for avoiding this situation. With the same LVs number as calibration, validation was calculated. The slope, $r^2$ of validation sample sets need reach a high value, normally over 0.95. What´s more, the RMSEP is an important parameter to check residual between predicted data and reference data.

In Table 9.3, we can see calibrations and validations of 4 chemical parameters of tomato samples. In models for determining fructose and glucose content, the slop and correlation of calibrations reached 0.95 and the RMSEC were lower than 0.5 g/L and 0.2 g/100g. The LVs were lower than 10. And in their validations, the values of RSEP were lower than 6%. In the model of soluble solids, the slop and correlation of calibration were 0.9567 and 0.9781. And the RMSEP of validation was 0.2365 °Bx. At last, in the model of dry matter content, the correlation of calibration was 0.9746 and the slop was 0.9499. Although the slop of dry matter was a little bit low, the RMSEC was 0.3503 g/100g and the RMSEP of validation was 0.2722 g/100g, which showed us a good predictive capability.

***Table 9.3 Final models of all 6 parameter of tomato samples***

| **Parameter** | **Calibration** | | | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Samples No. | Slop | Correlation | RMSEC | Factor | Explained Y | Samples No. | Slop | RSEP (%) | RMSEP |
| Fructose in juice | 249 | 0.9824 | 0.9912 | 0.3300 (g/L) | 7 | 98.24 | 60 | 0.9857 | 5.5284 | 0.4418 (g/L) |
| Glucose in juice | 252 | 0.9699 | 0.9848 | 0.4448 (g/L) | 7 | 96.99 | 60 | 0.9957 | 5.6007 | 0.4497 (g/L) |
| Fructose in puree | 204 | 0.9561 | 0.9778 | 0.1252 (g/100g) | 3 | 95.61 | 60 | 0.9734 | 3.8152 | 0.0809 (g/100g) |
| Glucose in puree | 201 | 0.9649 | 0.9823 | 0.1201 (g/100g) | 4 | 96.49 | 60 | 0.9783 | 3.9359 | 0.0821 (g/100g) |
| Brix in puree | 202 | 0.9567 | 0.9781 | 0.2921 (°Bx) | 4 | 95.67 | 60 | 0.9704 | 3.8169 | 0.2365 (°Bx) |
| Dry matter in puree | 202 | 0.9499 | 0.9746 | 0.3503 (g/100g) | 3 | 94.99 | 60 | 0.9673 | 3.5828 | 0.2722 (g/100g) |

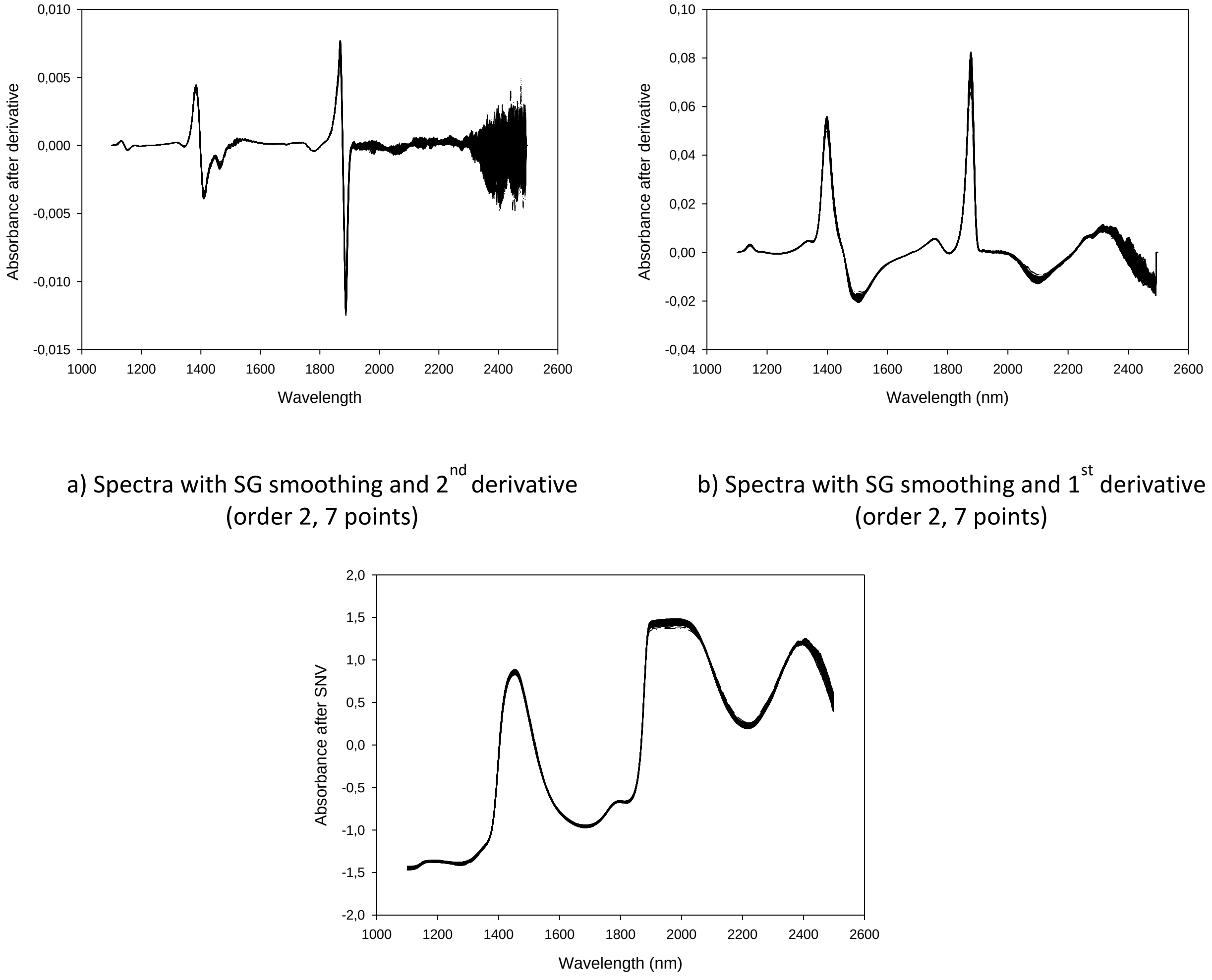
High values of slop, correlation, explained Y and $r^2$ resulted in low RMSEC and RMSEP. It indicates that these PLS models were suitable to quantify the fructose, glucose, soluble solids and dry mater content in tomato juice or puree samples.

### 9.3.2 Sensory parameters

Besides chemical parameters, the determination of sensory parameters was carried out by PLS models too. These parameters included sweetness, taste intensity, aroma intensity, mealiness, acidity, crunchiness, skin perception, explosiveness and juiciness. Reference data of these parameters were obtained by human sense. Samples were divided in to calibration set and validation set by PCA. Calibration sets of sweetness, taste intensity, aroma intensity, mealiness, acidity, crunchiness and skin perception had 45 samples, and their validation sets included 10 samples. The calibration sets of explosiveness and juiciness had 20 samples and 6 samples in validation sets. The sample range of validation sets has covered the similar range of calibration set.

#### 9.3.2.1 Spectra pretreatment

As the same reasons as the former process in chemical parameters, raw spectra need to be pretreated. The raw spectra of puree were pretreated by SNV or Savitzky-Golay smoothing with derivative. More pretreatments have been applied but the results of their application were not better than results of these two algorithms. Fig. 9.6 is the spectra after pretreated, we can find that water absorption peaks were still the main signal in raw spectra and there was noise from 2320 nm to the end.



a) Spectra with SG smoothing and 2$^{nd}$ derivative (order 2, 7 points)

b) Spectra with SG smoothing and 1$^{st}$ derivative (order 2, 7 points)

c) Spectra with SNV

***Fig. 9.6 Pretreated sense spectra***

As the first step, calibration models were calculated with spectra pretreated. The spectral range was from 1100-2498 nm. In Table 9.4 is the performance of these models. For sweetness and aroma intensity, better results of calibration have been obtained with SNV than derivative. Because the value of slop was higher and the RMSEC was lower. Then for taste intensity, mealiness, acidity, crunchiness, skin perception and explosiveness, the 2$^{nd}$ derivative gave a better result than the other one. The juiciness was special, the 1$^{st}$ derivative was better. Because it has removed the linear baseline drift and the constant baseline drift.

*Table 9.4 Calibration models of sense with two pretreatment methods. The better pretreatment for every parameter has been marked with the * symbol.*

| Name | Pretreatment | Calibration | | |
|---|---|---|---|---|
| | | Slop | RMSEC<br>(0-10) | Factor |
| Sweetness | 2D 7point | 0.7317 | 1.0701 | 3 |
| | *SNV | 0.7827 | 0.9634 | 3 |
| Tast intensity | *2D 7point | 0.883 | 0.4062 | 7 |
| | SNV | 0.7242 | 0.6233 | 7 |
| Aroma intensity | 2D 7point | 0.7514 | 0.9812 | 3 |
| | *SNV | 0.777 | 0.9293 | 3 |
| Mealiness | *2D 7point | 0.732 | 0.6993 | 3 |
| | SNV | 0.6874 | 0.7552 | 3 |
| Acidity | *2D 7point | 0.5799 | 0.8568 | 4 |
| | SNV | 0.4649 | 0.937 | 4 |
| Crunchiness | *2D 7point | 0.8803 | 0.7405 | 5 |
| | SNV | 0.5562 | 1.0401 | 5 |
| Skin perception | *2D 7point | 0.5595 | 0.6527 | 3 |
| | SNV | 0.1922 | 0.8839 | 3 |
| Explosiveness | *2D 7point | 0.9357 | 0.0662 | 5 |
| | SNV | 0.7833 | 0.1215 | 5 |
| Juiciness | *1D 7points | 0.9629 | 0.0512 | 6 |
| | SNV | 0.8609 | 0.0991 | 6 |

#### 9.3.2.2 Band selection

After the selection of the best pretreatment, PLS regression models of all sensory parameters were calculated within the spectral range from 2120-2320 nm. As discussed before, in this range, the noise range from 2320 to 2498 nm was avoided and absorption peaks were mainly caused by the overtone of C-H from puree. For this reason, the O-H peaks of water did not have a huge influence to the calibration models.

The RMSEP of validation was compared with the one which was calculated within the whole wavelength range. Table 9.5 shows the result of this comparison. Sweetness, aroma intensity, skin perception and juiciness had better models within whole wavelength range because the RMSEP was lower. And models of taste intensity, mealiness, acidity, crunchiness, explosiveness had a lower RMSEP within selected wavelength band.

***Table 9.5 Comparison of models calculated with in different range***

| Name | Band range | Pretreatment | Validation RMSEP (0-10) | RMSEP in whole range (0-10) |
|---|---|---|---|---|
| Sweetness | 2120-2320nm | SNV | 0.9023 | 0.6092 |
| Tast intensity | 2120-2320nm | 2D 7point | 0.7317 | 1.229 |
| Aroma intensity | 2120-2320nm | SNV | 0.9138 | 0.7668 |
| Mealiness | 2120-2320nm | 2D 7point | 0.5226 | 0.8232 |
| Acidity | 2120-2320nm | 2D 7point | 1.0445 | 1.2964 |
| Crunchiness | 2120-2320nm | 2D 7point | 0.8417 | 2.1184 |
| Skin perception | 2120-2320nm | 2D 7point | 0.8025 | 0.7875 |
| Explosiveness | 2120-2320nm | 2D 7point | 0.2063 | 0.3209 |
| Juiciness | 2120-2320nm | 1D 7point | 0.3512 | 0.0528 |

#### 9.3.2.3 Calibration and validation

Finally, models were calculated with the pretreatment and band commented in the previous paragraphs. Nine quantitative PLS models were calculated to determine sensory parameters. The t test was carried out to validate the residual between the predicted value and reference value. The results show that there was no significant system error for the residual. In Table 9.6, we can see the performance of these models.

***Table 9.6 Calibrations for sense parameters***

| Name | Model | Factors | Band range | Pretreatment | Slop of calibration | RMSEC (0-10) | RMSEP (0-10) | T test |
|---|---|---|---|---|---|---|---|---|
| Sweetness | PLS1 | 3 | 1100-2498nm | SNV | 0.7827 | 0.9634 | 0.6092 | $t_{experiment}$ 0.62＜2.26 |
| Taste intensity | PLS1 | 7 | 2120-2320nm | 2D 7point | 0.6921 | 0.6590 | 0.7317 | $t_{experiment}$ 0.04＜2.26 |
| Aroma intensity | PLS1 | 3 | 1100-2498nm | SNV | 0.7649 | 0.9543 | 0.7729 | $t_{experiment}$ 0.69＜2.26 |
| Mealiness | PLS1 | 3 | 2120-2320nm | 2D 7point | 0.7594 | 0.6609 | 0.5226 | $t_{experiment}$ 0.42＜2.26 |
| Acidity | PLS1 | 4 | 2120-2320nm | 2D 7point | 0.4681 | 0.9641 | 1.0445 | $t_{experiment}$ 0.35＜2.26 |
| Crunchiness | PLS1 | 5 | 2120-2320nm | 2D 7point | 0.6174 | 0.9656 | 0.8417 | $t_{experiment}$0.09＜2.26 |
| Skin perception | PLS1 | 3 | 1100-2498nm | 2D 7point | 0.5595 | 0.6527 | 0.7875 | $t_{experiment}$0.50＜2.26 |
| Explosiveness | PLS1 | 5 | 2120-2320nm | 2D 7point | 0.9571 | 0.0541 | 0.2063 | $t_{experiment}$1.78＜2.57 |
| Juiciness | PLS1 | 6 | 1100-2498nm | 1D 7point | 0.9629 | 0.0512 | 0.0528 | $t_{experiment}$1.58＜2.57 |

Models of explosiveness and juiciness gave us the best performance among all the sensory parameters. The slop of calibration was higher than 0.95 and the RMSEP was quite low. Models of sweetness, aroma intensity and mealiness had a slope of calibration higher than 0.75 and RMSEP lower than 0.8. This was a good result for sensory parameters. For taste intensity and crunchiness, the slope of calibration was higher than 0.6 and the RMSEP lower than 0.85. The slope was not high but the RMSEP was low compared to their sample range, so they are good models too. At last, for skin perception and acidity, the slope of calibration was higher than 0.45 and RMSEP lower than 1.05. The RMSEP could be accepted by the factory but the slope still needs to be improved.

In general, the calibration range was wider than the validation range. The number of LVs was lower than 7 which avoided the risk of overfitting.  With the selected band and pretreatments, the RMSE of calibration were lower than 1.0. The lowest RMSEP of validations in Table 9.6 was 0.2063 and the highest was 1.0445, which was low enough for the determination of sensory parameters in factory.

## 9.4 Conclusion

NIRS is a simple, fast and environmental friendly technology. In order to make a full use of this tool, chemometric methods were applied to find out the NIRS solutions which could quantify both chemical and sensory parameters. Through the workflow mentioned in this study, PLS models were built to analyze the CQAs of tomato products.  The results of validation show us that residuals between the predicted value and reference values were low. Calibrations of chemical parameters have reached the thresholds of quantitative model which means they were accurate and robust. Calibrations of sensory parameters were lower than the standard but this was caused by the inaccuracies of human sense and their performance was accepted by the producer. The feature that all the CQAs can be predicted with the unique spectrum of the same sample offers us a way to combine 13 inspections into 1 analysis. For the sensory parameters, the evaluation standards which are depended on the experience of skilled people have been introduced into the calibrations. These models demonstrated the ability of NIRS to record and perform the experience of human experts.

**Acknowledgement**

Thanks to Prof. Joan Casals Missioc and his team members from Escola Universitària Salesiana de Sarrià (EUSS).

# Reference

[1] Food and agriculture organization of United Nations, FAOSTAT database 2003-2013

[2] Gould W A. Tomato production, processing and technology [M]. Elsevier,2013.

[3] S. S. Nielsen, Food Analysis [M]. Springer. 2010. [4] Rosas J G, Blanco M, González J M, et al. Real-time determination of critical quality attributes using near-infrared spectroscopy: A contribution for Process Analytical Technology (PAT) [J]. Talanta, 2012, 97: 163-170.

[5] Alcalà M, Blanco M, Bautista M, et al. On-line monitoring of a granulation process by NIR spectroscopy[J]. Journal of pharmaceutical sciences, 2010, 99(1): 336-345.

[6] Alcalà M, Ropero J, Vázquez R, et al. Deconvolution of chemical and physical information from intact tablets NIR spectra: Two-and three-way multivariate calibration strategies for drug quantitation[J]. Journal of pharmaceutical sciences, 2009, 98(8): 2747-2758.

[7] Armenta S, Garrigues S, De la Guardia M. Green analytical chemistry[J]. TrAC Trends in Analytical Chemistry, 2008, 27(6): 497-511.

[8] Casals, J., Pascual, L., Canizares, J., Cebolla-Cornejo, J., Casanas, F., Nuez, F., 2011. The risks of success in quality vegetable markets: possible genetic erosion in Marmande tomatoes (Solanum lycopersicum L.) and consumer dissatisfaction. Sci. Hortic. 130, 78-84.

[9] ISO, 1998. Sensory analysis - General guidance for the design of test rooms (Ref. ISO 8589). International Organization for Standardization, Genova, Italy.

[10] Hongsoongnern, P., Chambers, E., 2008. A lexicon for texture and flavor characteristics of fresh and processed tomatoes. J. Sens. Stud. 23, 583-599

[11] S., 2010. Sensory Quality of Fresh French and Dutch Market Tomatoes: A Preference Mapping Study with Italian Consumers. J. Food Sci. 75, S55-S67.

[12] Tikunov, Y., Lommen, A., de Vos, C.H.R., Verhoeven, H.A., Bino, R.J., Hall, R.D., Bovy, A.G., 2005. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol. 139, 1125-1137.

# Chapter 10

# Conclusion

# 10. Conclusion

In this thesis, the NIRS has been applied as a tool which helps the realization of industry 4.0 in the food and pharmaceutical industries. The NIR spectrometers could virtualize the physical and chemical information of the process product in real time. CPPs and CQAs are learned by the multivariate models so that computers could analyze all these data and control the manufacturing line automatically. By this way, the manufacturing process becomes smart. The operations or decisions, which used to depend on the personal experience, can be carried out by the computer. The smart manufacturing line could learn and improve the production process according to the customized requirements, so the production process is flexible and can be changed.

The conclusion of the thesis could be summarized as following:

The NIRS is a method which plays an important role in the industry 4.0 concept. The analyzing time of NIRS in this thesis were less than 1 minute without destructing samples, which has proved that NIRS could make RTM, RTA and RTR come true in the pharmaceutical industry. Instead of only some represents, all products could be analyzed so the evaluation of quality is more close to the true situation.

The virtualization ability of NIRS could transfer the personal experiences into scientific data which could control production process automatically and make the industry smart. The physical and chemical information could be virtualized by NIRS into the multivariate models, so CQAs and CPPs are learned by computers. The machine learning function could be realized by this ability.

Solid drugs quality can be controlled and assured by the control charts which are plotted with the NIRS data. The QC and QA departments could have a further understanding of the manufacturing process which is more detailed and scientific. The quality consistency could be accurately achieved with the quantitative models calculated with NIR spectra of process products.

The chemical and sensory parameters are able to be analyzed at the same time. This combination makes the food quality assurance become a simple process. Specially, the sensory analysis of tomato offers a simple and precise way to manufacture products which are customized for the requirements of consumers.

The MicroNIR has performed well in the studies of pharmaceutical products. It is portable and has a small size which makes it suitable for customer use and the on line analysis. The instrument noise is low and the peak accuracy is high so the spectra quality is good enough for the application in the pharmaceutical industry. Comparing the bench top instruments, the MicroNIR is more robust and cheaper.

Experiment design of the pharmaceutical studies has expanded the concentration range of APIs and excipients in the calibration set. Robust quantitative PLS1 models were calculated and their validations proved that these models had reliable prediction ability. The experiment design is efficient, and the validation methods have verified it.

The whole analysis process is human and environment friendly. All the experiments were carried out without any poisonous solution, so operators were safe when they acquire the spectra of samples. Meanwhile, the nondestructive analysis didn´t produce any waste from the process, so no pollution was created for the environment.