Near-infrared mass median particle size determination of lactose monohydrate, evaluating several chemometric approaches

P. Frake,*[ac] I. Gill,[b] C. N. Luscombe,[a] D. R. Rudd,[a] J. Waterhouse[b] and U. A. Jayasooriya[c]

[a] *GlaxoWellcome Research and Development, Park Road, Ware, Hertfordshire, UK SG12 0DP*
[b] *GlaxoWellcome Operations, Priory Street, Ware, Hertfordshire, UK SG12 0DJ*
[c] *School of Chemical Sciences, University of East Anglia, Norwich, Norfolk, UK NR4 7TG*

Received 2nd April 1998, Accepted 11th August 1998

The influence of particle size on near-infra red (NIR) spectra is typically considered a 'nuisance factor' which many scatter correction methods attempt to eliminate, *e.g.*, multiplicative scatter correction. However, particle size is a key issue in the formulation of many pharmaceutical products and has a profound effect on the behaviour of both raw materials and drug substances during formulation. NIR has already been demonstrated as a potential alternative particle sizing technique to current accepted methodology. This investigation assessed several chemometric approaches that model this information, using lactose monohydrate as the raw material. A variety of modelling techniques were applied to both zero order and second derivative spectra namely multiple linear regression, partial least squares, principal component regression and artificial neural networks. One further data transformation evaluated was polar coordinates, although no statistical data were generated. Typically, cross-validation root mean square errors of calibration and cross-validation root mean square errors of prediction of approximately 5 μm were calculated for all of the modelling techniques. These values are comparable to those associated with the reference technique (laser diffractometry). Correlation coefficients of approximately 0.98 for all techniques were also calculated. The predictive abilities for models generated using second derivative spectra were found to be comparable to those obtained using zero order spectra.

Introduction

The determination of particle size by near infra-red (NIR) diffuse reflectance spectroscopy, applied to the testing of pharmaceutical products and raw materials, has been alluded to in many review articles.[1–7] However, only a few papers have been published dedicated to the determination of particle size.

Ciurczak and co-workers[8] correlated particle size with NIR data for four pharmaceutical materials. On plotting the absorbance at a given wavelength *versus* inverse mean particle size, or two absorbances against each other, acceptable linearity was achieved. Ilari *et al.*,[9] determined mean particle sizes for NaCl, powdered glass and sorbitol. Calibration models were generated using partial least squares (PLS) based upon the residual data calculated following scatter correction treatment. Blanco and co-workers[10] investigated the effect of the sample cell with respect to reproducible NIR particle size determinations for piracetam. Calibration models in this example were generated using multiple linear regression (MLR). Recently O'Neil *et al.*,[11] demonstrated that by fitting a fourth order polynomial to the baseline of NIR spectra, acceptable correlation between NIR data and particle size was possible. These examples therefore demonstrate that there is more than one way to model particle size data with NIR spectra.

Typically, the physical properties of materials such as particle size and the effect on diffuse reflectance spectra are secondary to the analyst's requirements. As a result, many methods exist which attempt to remove this effect.[12–14]

Within the pharmaceutical industry, the determination of particle size is a key issue in the formulation of many products. Fundamental issues such as blending characteristics, bulk density, bioavailability and vital flow properties are all governed by the particle size (and distribution of particle size).

At present, the most commonly used technique in-house for the determination of particle sizing is laser diffractometry (Malvern Mastersizer). The disadvantages of laser diffractometry are that the determination of particle size is based on the equivalent sphere diameter (the majority of particles are not spherical) and that sample preparation often requires additional solvents in order to suspend the test material. To some extent NIR will have the same errors as are associated with laser diffractometry with respect to particle shape. Laser diffractometric measurements are based on the Mie theory for light scattering from a single particle. This theory requires the assumption that all particles are of a simple shape, *i.e.*, spherical, for the basis of its calculations. The Mie theory has since been expanded to include scatter from a number of particles, (this includes theories proposed by Kubelka and Munk). NIR diffuse reflectance measurements, in turn, use Kubelka and Munk light scattering theories, and therefore the two techniques, laser diffractometry and NIR diffuse reflectance measurements, can be considered to have the same basis of origin. However, it is possible for NIR models to predict with greater precision than the reference technique.[15] This is possible as the use of a sufficient number of training samples will average out most of the random error associated with the reference technique.

The advantages to be gained from the use of NIR include the avoidance of sample preparation and the potential to be used on-(or at-) line, generating results in real time.[16]

Using laser diffractometry as the reference technique, several chemometric modelling techniques were applied to correlate NIR spectral data with particle size. Comparison of some of the methods previously employed in the literature was made, *i.e.*, MLR and PLS.[17,18] In addition to these, artificial neural networks[19] (ANN) and principal component regression (PCR)[17] were applied to both zero order and second derivative spectra.

In a similar manner to that presented by Van Der Vlies *et al.*,[20,21] a polar qualification system was also evaluated. In this data transformation, the NIR spectrum is plotted around a centre of gravity, producing only two variables (***x*** and ***y*** vectors for the centre of gravity). Using this approach it was observed that the centre of gravity was indeed displaced according to the mass median particle size (mmps). Following this data transformation, the distance between the centre of gravity and an arbitrarily selected point on the polar plot was used for the purposes of linear regression. As the centre of gravity appeared to be displaced owing to mmps, and this effect manifests itself across the entire NIR spectrum, it was felt that this approach was acceptable. It was considered that this approach had some potential; however, we did not undertake a detailed evaluation.

Experimental

Materials

Lactose monohydrate BP/USNF/NCZ grades (including 40# and 170#) were obtained from Borculo Whey Products (Saltrey, UK), Dairy Crest (Thames Ditten, Surrey, UK) and Zimmermann Hobbs (Bletchley, Milton Keynes, UK).

Laser Diffractometry

A Malvern Mastersizer MSX03LA was obtained from Malvern Instruments (Malvern, UK). Duplicate determinations were performed on test material suspended in organic solvent during sample measurement.

Near-infrared spectrophotometry

A Model 6500 spectrophotometer equipped with a rapid content scanner accessory was used (Foss NIR Systems, Silver Springs, MD, USA).

Prior to the generation of any calibration model, it was necessary to eliminate as many sources of experimental error as possible, allowing reproducible sample measurement. Early experiments used sieved fractions of lactose monohydrate placed in glass vials. This approach failed on two counts: first, the reproducibility of sample measurement was very poor owing to the inconsistent moulding effects observed at the base of every glass vial, and second, the use of sieve analysis as the reference technique proved to be unsuccessful.

Poor correlation between the two techniques prompted the further analysis of the sieved fractions by laser diffractometry. In short, the sieve reference values, *i.e.*, mesh size, bore no resemblance to the data generated by laser diffractometry for the same samples. The cause of this can be attributed to the mechanics involved in sieving *i.e.*, long thin particles can easily pass through a mesh, whereas when rotating in a suspension undergoing laser diffractometry, the particle will be measured as being much larger. In addition, electrostatic forces can cause particles to agglomerate. In a suspension these particles will separate and as a result will be measured as smaller particles by laser diffractometry.

One final observation during these early experiments was the possibility of generating false low absorbance readings. This was proved possible by subjecting the sample to a small degree of agitation. This caused small particles (fines) to collect at the bottom of the sample container, thereby biasing the distribution, hence producing low absorbance readings.

To overcome these effects, glass vials were replaced by an IR grade silica cuvette. To minimise errors associated with sample preparation, bulk material was transferred directly to the cuvette with no further processing.

Once these preliminary investigations had been completed, it was considered that true particle size effects would be observed (see Fig. 1) and the use of laser diffractometry as the reference technique would demonstrate the desired correlation.

Thirty-seven batches of lactose monohydrate encompassing an mmps range of 20–110 μm (commercially available material requiring no pre-processing) were scanned by NIR in triplicate. Triplicate spectra for each batch were averaged prior to any model generation (no other spectral pre-treatment was made).The same batches were then analysed by laser diffractometry.

Of the 37 batches, 30 were arbitrarily selected from across the mmps range and used in the generation of each respective calibration model. The remaining seven batches were used to test each model as an independent test set.

The acceptance criteria applied to NIR data were standard errors for both calibration and prediction in the order of 5 μm. This value is considered comparable to the error associated with laser diffractometry. In-house, it has been estimated that the error associated about a size determination is approximately 2.5%. However, this value increases exponentially as the mmps decreases owing to the mechanics of laser determinations. Therefore, a value of about 5 μm calculated across the entire range analysed is considered acceptable.

Software

NSAS software version 3.30 (NIR Systems), Neudesk version 2.11 (NCS Manufacturing Intelligence), and the Unscrambler version 6.1 (Camo) were used.

Chemometrics

The acquisition of an NIR spectrum can produce several hundred wavelengths (variables), some of which will be highly correlated with the parameter of interest, and many which are less correlated. Typically, in order to define a relationship between a measurable quantity with features in the NIR spectrum, it is necessary to reduce the number of variables to a manageable size, potentially removing a considerable amount of noise. This can be achieved in a number of ways.

In the event of a simple, single variable correlating with the parameter of interest (in this example, mmps), MLR will isolate

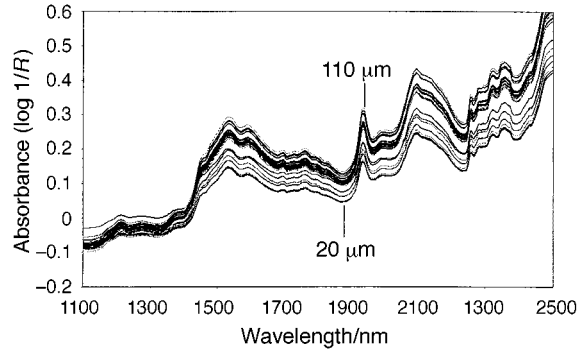

Fig. 1 NIR spectra [log (1/*R*)] of lactose monohydrate with mass median particle size range 20–110 μm.

this relationship, *i.e.*, a single wavelength response correlating with mmps.

However, physical effects, such as particle size, which will manifest itself across the entire spectrum, heavily influence an NIR spectrum. Consequently, there may be a number of variables correlated with mmps. For MLR, this potentially causes co-linearity problems (whereby more than one variable displays the desired relationship, *i.e.*, one wavelength can be substituted for another, producing a comparable result, and the linear equation therefore cannot be solved). Consequently, the model's ability to predict the parameter of interest decreases. Data reduction techniques which are unaffected by co-linearity, include PCR and PLS.

PCR replaces the initial input variables, *i.e.*, 700 wavelengths, with latent variables (principal components) that account for the variance within the dataset. The first principal component is the one that describes the most variance, with each subsequent component describing progressively less variance. Each principal component is calculated orthogonally to the preceding component and is therefore not correlated to any of the preceding components. This removes any problems associated with co-linearity. A subset of all the components can then be regressed against the response in a manner analogous to MLR.

PLS can be considered analogous to PCR, and often the results generated by the two methods are very similar. Once again the original variables are reduced to latent variables; however, in this case, the latent variables are constructed so as to account for the maximum amount of covariance between the original variables and the response. As a result, the first latent variable (factor) is the most correlated with the response, *i.e.* it reduces the residual sum of squares error. Each subsequent factor further reduces this error.

In both of these techniques there is the potential for overfitting, *i.e.*, too many factors or components will only add noise to the model, and consequently the model's ability to predict a new sample diminishes. The use of cross-validation to determine the correct number of factors/components is one recommended way to overcome this.

Typically PCR has a tendency to model the physical effects in the first few components as it is these physical effects which contribute significantly to the spectral response. As the PLS model has been calibrated to predict the mmps of the test material, the components and factors will ultimately produce similar models.

The last technique evaluated was ANN. This approach differs from the previous treatments in that there is no assumption as to a particular form of the model to be used. In a manner analogous to biological systems, the network is trained by example to learn the nature of the relationship between a series of input parameters and an output. One advantage of using this approach is the ability of the artificial neural network to model any non-linearity in the data set.

Results and discussion

Multiple linear regression

Of the five data treatment techniques evaluated for zero order spectra, MLR produced the least effective model. Using NSAS, a correlation coefficient (r) of 0.991 was achieved with three wavelengths. Although this was able to model the calibration set satisfactorily, it was unable to predict the test set as accurately. A standard error of calibration (SEC) of the order of 3.9 μm and a standard error of prediction (SEP) of 7.8 μm were calculated. As less wavelengths were selected, the correlation, as expected, decreased, introducing a higher SEC.

Partial least squares

With the current data set, Unscrambler calculated a correlation coefficient of 0.986. This used three factors and appeared to model both the calibration set and test set satisfactorily with SEC and SEP values of approximately 4 μm.

Principal component regression

The same data set using the PCR function within Unscrambler produced a calibration model with a correlation coefficient of 0.983, using three principal components. SEC and SEP values of the order of 4 μm (almost identical with the values calculated when using PLS) were calculated.

Artificial neural networks

In a manner analogous to PCR, the first three PCs from each of the 30 training batches were used to train the ANN, to learn the underlying relationship between the input spectra and mmps. The network architecture was optimised with respect to the number of hidden layer neurons, and three neurons were found to produce an SEC and SEP of 4.6 and 5.3 μm, respectively. A correlation coefficient of 0.988 was calculated, again showing good agreement with the other techniques evaluated. A comparison of predictive abilities and statistical summary for all data treatments can be seen in Table 1. In addition, Fig. 2 displays the predictive ability of each model on the independent test set. As demonstrated by Ciurczak and co-workers[8] derivatisation does not remove particle size effects. Consequently, a further comparison was made between models based on zero order spectra and those built using second derivative data.

Table 1 Statistical evaluation of predictive ability for each model

		Calibration (n = 30)				
Order	Model	No. of factors	r^2	cvRMSEC[a]	cvRMSEP[a]	Test (n = 7): SEP
Zero order	MLR	3	0.991	3.9		7.8
	PCR	3	0.983	4.5	5.1	4.1
	PLS	3	0.986	4.3	4.7	4.2
	ANN		0.988	4.6		5.3
Second derivative	MLR	1	0.987	4.7		3.5
	PCR	3	0.990	3.5	3.9	3.7
	PLS	3	0.990	3.4	3.9	4.0
	ANN		0.987	4.8		4.8

[a] cvRMSEC = cross-validation root mean square error of calibration; cvRMSEP = cross-validation root mean square error of prediction.

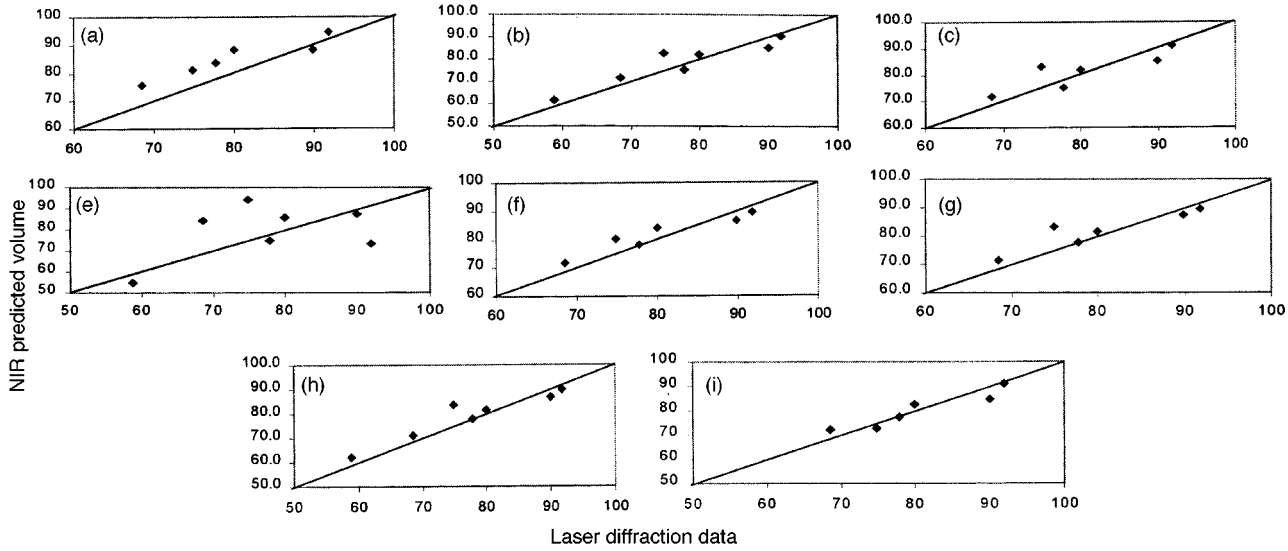

Fig. 2 Graphical analysis of the predictive ability of each model on the independent test set: (a) MLR (zero order); (b) PCR (zero order); (c) PLS (zero order); (d) ANN (zero order); (e) (MLR (second derivative); (f) PCR (second derivative; (g) PLS (second derivative); (h) ANN (second derivative).

The statistical summaries for each algorithm, based on second derivative data, are given in Table 1. The predictive abilities of these second derivative models on the independent test set are shown in Fig. 2.

It appears that derivatisation does not significantly change the ability of each data treatment technique to model particle size information. This confirms previous observations that derivatisation does not remove particle size effects. It is also worth noting that for the derivatised data the automatic wavelength selection (NSAS software) calculated the MLR model using the values at 1100 nm. As the first and last few data points tend to zero following derivatisation, this is considered an unacceptable region on which to base the model. Further evaluation of this particular model would require assessment at other wavelengths.

Conclusions

This exercise has demonstrated that a number of data treatment techniques can be successfully applied to NIR data for the prediction of mmps. Data generated by PCR, PLS and ANN showed good agreement. This was not surprising, as the latent variables associated with PCR and PLS in this particular example will be modelling the same effect. The ANN, in order to predict the underlying trend had also used the first three principal components to model the data. MLR was found not to be as reliable, probably owing to the problems associated with co-linearity.

When evaluating models based upon second derivative spectra, it was found that no significant improvement in the predictive abilities upon the independent test set were observed.

With respect to sample preparation, it was found that the minimum the sample was handled the better. It was found to be reasonably easy to bias the particle size distribution by any means of tapping or vibration. As part of this exercise, an investigation between the particle size distribution and the NIR spectra was undertaken. This information has yet to be extracted; however, it appears that samples with very different distributions but similar mmps do not significantly alter any model's ability to predict the mmps.

Acknowledgements

The authors acknowledge the additional contributions made by members of the Physical Properties Group, GlaxoWellcome.

References

1 E. W. Ciurczak, *Appl. Spectrosc. Rev.*, 1987, **23**(1/2), 147.
2 J. J. Drennen and R. A. Lodder, *J. Pharm. Sci.* 1990, **79**(7), 622.
3 E. W. Ciurczak, *Pract. Spectrosc.*, 1992, **13**, 549.
4 E. W. Ciurczak, *Chemtech.*, 1992, **22**(61), 374.
5 E. W. Ciurczak and J. K. Drennen, *Spectrosc.*, 1992, **7**(6), 12.
6 W. Plugge and C. Van der Vlies, *J. Pharm. Biomed. Anal.*, 1993, **11**(6), 435.
7 M. Blanco, J. Coello, H. Iturriaga, S. Maspoch and C. De La Pezuela, *Talanta*, 1993, **40**(11), 1671.
8 E. W. Ciurczak, R. P. Torlini and M. P. Demkowicz, *Spectrosc.*, 1986, **1**(7), 36.
9 J. L. Ilari, H. Martens and T. Isaksson, *Appl. Spectrosc.*, 1988, **42**(5), 722.
10 M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, F. Gonzalez and R. Pous, *Near Infra-red Spectroscopy* (*Bridging the Gap Between Data Analysis and NIR Applications*), ed. K. I. Hildrum, T. Isaksson, T. Naes and A. Tandberg, Ellis Horwood, Chichester, 1992, pp. 401–406.
11 A. J. O'Neil, R. D. Lee, R. A. Watt and A. C. Moffat, *J. Pharmacy Pharmacol.*, 1997, **49**(4), 19.
12 L. S. Aucott, P. H. Garthwaite and S. T. Buckland, *Analyst*, 1988, **113**, 1849.
13 C. R. Bull, *Analyst*, 1991, **116**, 781.
14 B. G. Osborne, T. Fearn and P. H. Hindle, *Practical NIR Spectroscopy, with Application in Food and Beverage Analysis*, Longman Scientific, Harlow, 2nd edn., 1993.
15 R. DiFoggio, *Appl. Spectrosc.*, 1995, **49**(1), 67.
16 P. Frake, D. Greenhalgh, S. M. Grierson, J. M. Hempenstall and D. R. Rudd, *Int. J. Pharm.*, 1997, **151**, 75.
17 H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester, 1989.
18 M. A. Sharaf, D. L. Illman and B. R. Kowalski, *Chemometrics*, Wiley, Chichester, 1986.
19 J. R. Long, V. G. Gregoriou and P. G. Gemperline, *Anal Chem.*, 1990, **62**, 1791.
20 C. Van der Vlies, K. J. Kaffka and W. Plugge, *Pharm. Tech.*, 1995, **7**(4), 43.
21 C. Van der Vlies, *Eur. Pharm. Rev.*, (February), 1996.

Paper 8/02532K