OPTIMIZATION REGRESSION MOLDELS USING BIPLS FOR DETERMINATION OF PROTEIN CONTENT IN COMMERCIAL WHEAT FLOUR

Henrique Rigobello Guedes (UNISC)
hrigobello@terra.com.br
Rafael Guedes de Azevedo (UNISC)
rafael_gazevedo@yahoo.com.br
Marco Flôres Ferrão (UNISC)
ferrao@unisc.br
Geraldo Lopes Crossetti (UNISC)
geraldoc@unisc.br
Celso Ulysses Davanzo (UNICAMP)
celso@iqm.unicamp.br

The use of infrared reflection techniques, in particular Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS), has allowed the development of fast and non-destructive methodologies of supply analyses in food industries, faciilitating the diverse routine analyses from industries. Therein, in this work the multivariate biPLS algorithm calibration has been applied to generate an optimized model to predict the protein content in commercial wheat flour samples, using DRIFTS spectra. The obtained results were compared to the global model, applying the algorithm PLS and to the models generated by the selection algorithm iPLS .The best prediction model to determine the protein content in commercial wheat flour was obtained applying the algorithm biPLS, using the combination of 182 wavenumbers spectral signals resulting in 0.4665% RMSEP.

Keywords: optimization, biPLS, protein content, infrared spectroscopy.

1. Introduction

Wheat is a basic component in human nourishment. Its flour is largely used to make bread, pastries and cookies. The quality of the produced grain determines its industry usage, and to do so a lot of quality tests are performed, which determine the levels of protein, ash and gluten (EMBRAPA, 2007). Economically, wheat is a very important cultivation in Brazil, and there is a growing population demand for its cereal products, estimated in 11.2 millions of tons in 2005, although the country's production (4.5 millions of tons in 2003) attends only part of this demand (AGRIANUAL, 2003). The country has climate, soil, genetic material and available technology to cultivate more than 10 millions of cereal hectares.

Most of food industries, mainly the ones which focus on the production of pastries and cookies, work with low stocks of its diverse raw materials, and wheat flour is one of them. And for this a lot of routine tests are performed. Many of the official tests, which are used, don't allow an immediate control of these flour lots, being merely formal checks of raw material, because the flour goes to the packing industries before the test results are available. For this reason food industries work with research centers to find out alternatives to provide the diagnose of its raw materials in real time, guaranteeing the process and final product quality, lowering the losses of inadequate handling of determined supplies in its production lines and obtaining faster test results (FERRÃO, 2000).

The protein content is mentioned by a lot of authors as one of the main indicators of final quality use of the flour (OSBORNE e FEARN, 1983; OSBORNE *et al*, 1982; SCHILLER, 1984; FERRÃO *et al*, 2004). The protein content is related to the capacity of constituting dough because when wheat flour and water are mixed we have as result the formation of dough constituted of a gluten protein net linked to starch granules that retain the carbon dioxide produced during the fermentation process. This makes the bread retain the gas produced and increase its size. The protein content is determined using the Method 2055 of the Association of Official Analytical Chemists in 1984 (AOAC, 1990). However, it is a methodology that requires a lot of time, around 10 hours according to the National Center of Technological Agro-industrial Food Research at EMBRAPA (Brazilian Enterprise of Agro-cattle breeding Research). Another drawback of this method is the use of copper sulfate, selenium, sulfuric acid and boric acid, which are highly toxic to the environment and require great care being handled because they cause irritations, nauseas and even death (CISQ/IBILCE, 2007).

Interested in developing and optimizing a fast and non-destructive analytical methodology able to estimate the protein content present in commercial wheat flour samples, this article presents results referent to the application of multivariate regression methods using PLS, *i*PLS and *bi*PLS for the data obtained by Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS).

2. Multivariate analysis

Chemometric techniques for multivariate analysis have become common approaches for a fast analysis of complex samples in their spectral analyses and get more efficient with the choice of spectral range, minimizing prediction errors. The benefits of selecting a spectral range include the stability of the calibration model in relation to co-linearity as well as the interpretation of the relations between the model and the sample composition.

There are several ways of selecting a spectral wavelengths: objective criteria, such as

determinant of calibration matrix and evaluation of the root mean square error of validation and algorithms that indicate the spectral group that will be able to bring the best results, being the most used: genetic algorithms and nowadays, the method of Partial Least-Squares – PLS, method of Interval Partial Least-Squares – *i*PLS and Backward Interval Partial Least-Squares – *bi*PLS, which are used in this article.

In PLS the multivariate calibration is set using the information of the whole spectrum to build the regression model, related to the property you are interested in. Because of this it is called method full-spectrum (PASCHOAL *et al*, 2003;BORIN e POPPI, 2005; ROCHA *et al*, 2006). In this work PLS was used with MATLAB, which integrates mathematical computation, visualization and efficient language in a flexible environment for technical computation.

The basis of the method partial least-squares is the decomposition of the matrix **X**, in the sum of various matrixes **M** that present dimensionality one, and that are added with a residual matrix (which correspond to the non-modeled part of **X**), according to the following equation:

$$\mathbf{X} = \mathbf{M_1} + \mathbf{M_2} + ... + \mathbf{M_a} + \mathbf{E}$$

where *a* corresponds to the number of latent variables (principal components or factors) selected to truncate the equality, and **E** corresponds to the residual matrix, related to the naumber of latent variables chosen.

The matrixes **M** constitute the so called latent variables, and are constituted by the product of two vectors, **t** (the scores) e **p** (the loadings) according to the expression below:

$$\mathbf{X} = \mathbf{t_1}.\mathbf{p_1}^t + \mathbf{t_2}.\mathbf{p_2}^t + ... + \mathbf{t_v}.\mathbf{p_a}^t + \mathbf{E}$$

The dimensionality of the original space is equal to the number of columns in **X**, that is, the number of original variables expressed by *m*. In the new model, the dimensionality is described by the number of matrixes $\mathbf{M}_i$ necessary to describe **X**. Thus, if it is possible to describe a matrix **X** that has several variables, with a small number of these matrixes $\mathbf{M}_i$ there will be a decrease in dimensionality, without significant loss of information.

In the modelling by partial least-squares, the matrix of independent variables **X**, as well as the dependent variables **Y** are represented by scores and loadings, according to the expressions below:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^t + \mathbf{F}$$

The relation between the two data matrixes **X** and **Y** can be obtained correlating the scores in each block, in order to obtain a linear relation described in the following expression:

$$\mathbf{U} = \mathbf{b}\mathbf{T} + \mathbf{e}$$

where **U** is matrix containing the properties (dependent variables) of all these samples; **b** is a vector containing the model parameters; **T** is the answer matrix (as well as a set of spectra) for a series of calibration samples and **e** is a vector that represents the spectrum noise and the model errors (KONZEN *et al*., 2003).

On the other hand, the technique *i*PLS is an extension of PLS and divides the set of data into a number of intervals, calculates the model PLS for each interval and presents the data in a graph. The method is planned to give an overview of data and might be useful to interpret which spectrum signals are more representative in the building of a good calibration model.

Finally, the optimization by *bi*PLS removes spectral ranges without relevance, in which the

models PLS are initially calculated with each of the intervals left aside. For example, by choosing 20 intervals, each PLS model is build with 19 intervals, leaving aside one interval each time it is calculated. The first interval to be eliminated will the one that when left aside results in the worst model, that is, the one that provides the highest RMSEV (Root Mean Square Error of Validation). This procedure is continued until there is a remaining interval, as represented in the pseudocode below:

While the number of intervals > 1 do
Build the PLS models leaving one interval each time
Verify each PLS model generates the biggest RMSEV
Eliminate the interval (number of intervals – 1)
End while

The definition of the number of intervals in *bi*PLS is a very important task because if the number of intervals is too small, the spectral ranges will be wide and consequently the effects of lowest peaks might be lost. On the other hand, if the number of intervals is too big, the results will be in a local scale and a bigger computational time will be necessary.

3. Materials and methods

3.1 Samples

100 Samples of wheat flour were collected by Filler S.A., a Brazilian company from Santa Cruz do Sul in the state of Rio Grande do Sul. The flour samples were purchased from different mills and were a blend of Triticum aestivum L. wheat that had protein content between 8.85 and 13.23%.

3.2. Reference method

The determination of protein content of wheat flours used as reference value was based on official published method. The method number 2055 (1984) of the Association of Official Analytical Chemists (AOAC, 1990) was used. For each sample, the analyses were made in triplicate and the mean value was taken as the reference value for both parameters.

3.3. Spectroscopic measurements and multivariate regression routines

Each flour sample was analyzed by Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) ranging from 3800 to 600 cm^{-1} with 32 scans. The spectra were read in absorbance duplicate for each sample, normalized and processed by autoscaling in the software MATLAB, using iPLSToolbox (NORGAARD *et al*., 2000), by processing of partial least-squares with crossvalidation with the exclusion of one spectrum each turn. The statistical paramaters used to select the obtained models were the coefficient of calibration correlation (R^2_{cal}), as closest to 1 as possible, and the smallest root mean square errors of validation (RMSEV) and of prediction (RMSEP).

4. Results and discussions

The application of algorithms PLS, *i*PLS e *bi*PLS resulted in differentiated models related to the correlation coefficient (R^2_{cal}), RMSEV and to RMSEP as Table 1 shows.

Algorithm	NV[a]	LV[b]	R^2_{cal}	RMSEV (%)	RMSEP (%)
PLS	520	5	0.661	0.4573	0.5547
*i*PLS	26	2	0.453	0.4792	0.5489

*bi*PLS	182	4	0.617	0.3088	0.4665

[a]NV = number of wavenumbers selected
[b]LV= number of latent variables

Table 1 – Results of multivariate regression PLS, *i*PLS e *bi*PLS models for the prediction of the protein content.

The algorithm PLS applied to all the spectroscopic signal generated a low quality model, according to the prediction values versus reference values for external samples presented in Figure 1. When all the spectrum is selected, a 0.5547% root mean square error of validation was found and a 0.661 determination coefficient for 5 latent variables.

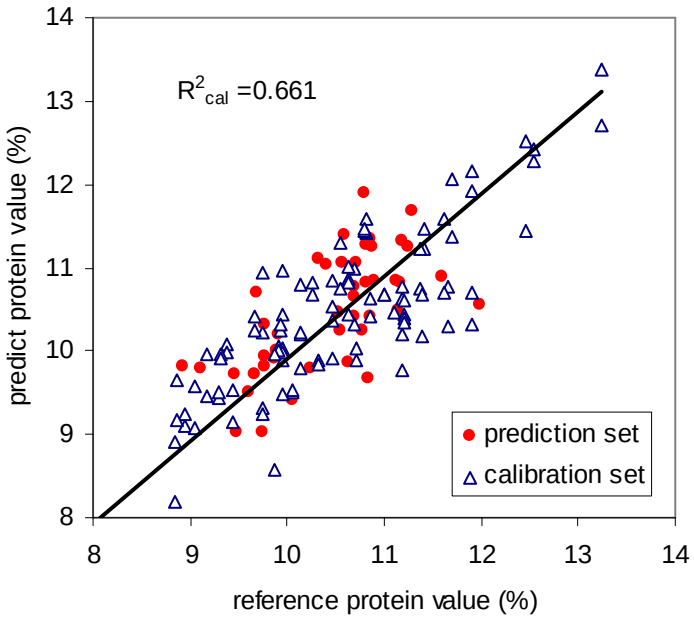

Figure 1 – PLS Result: graph of correlation of the global model for 5 latent variables.

Due to the poor performance of the PLS using the whole spectral range, the spectrum was divided into 10, 20 e 40 sub-interval equidistant to the application of *i*PLS, being the 20-interval model the most representative. In Figure 2, the bars indicate the prediction error obtained by validation for each selected interval and the dotted line indicate the error for the global model. The figure shows that for 5 latent variables the global model is better than for any of the local models.

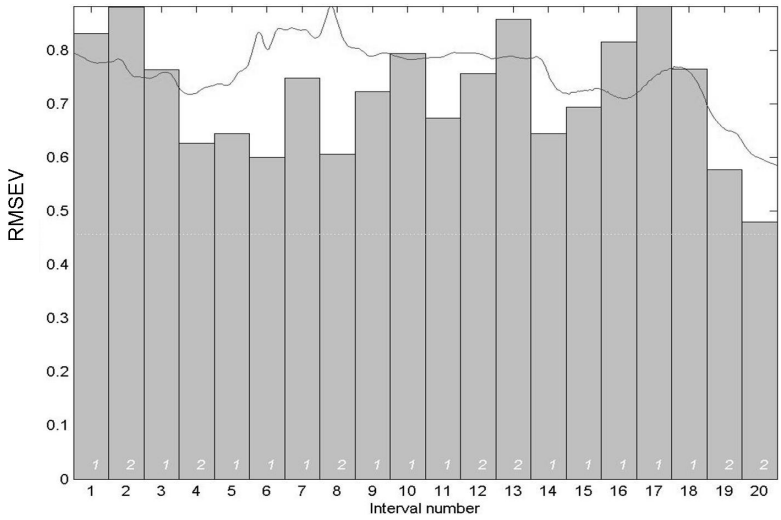

Figure 2 - Mean square error of crossvalidation obtained by crossvalidation for each interval.

The best model obtained with the application of *i*PLS technique (Figures 3 and 4) presented a performance similar to the PLS model (full spectra), presenting a 0.5489% root mean square error of validation and a 0.453 determination coefficient, for 2 latent variables. In spite of the worst correlation for calibration samples (R^2), the model *i*PLS is the simplest because it uses 1/20 of the wavelength numbers from the PLS, being less vulnerable.

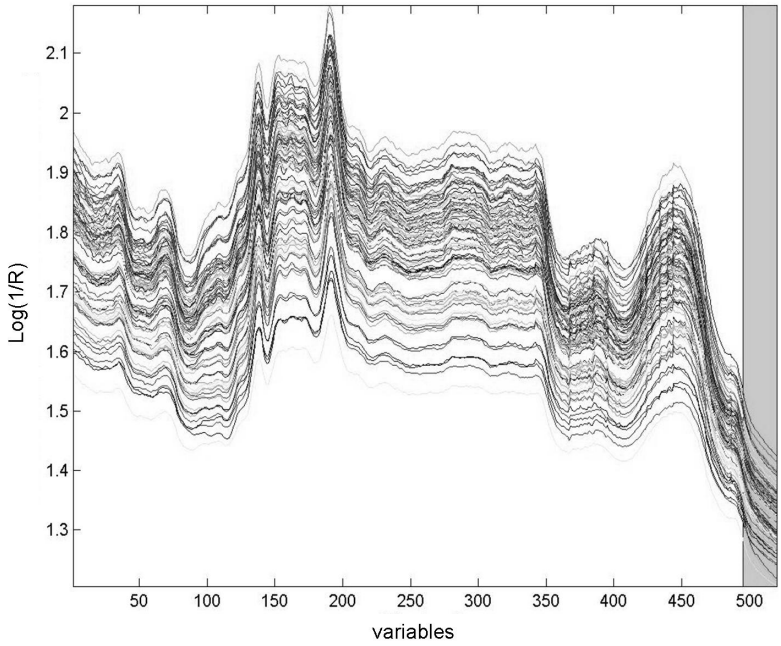

Figure 3 – Spectral wavelenght range used for the *i*PLS modelling.

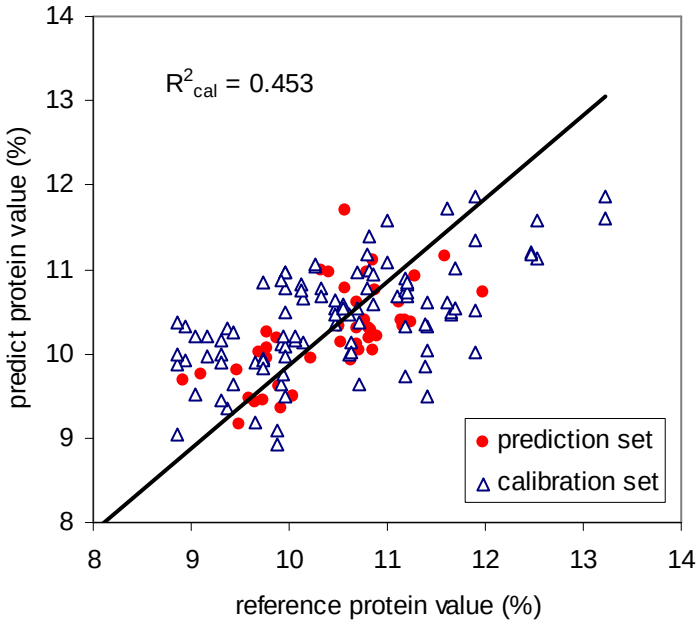

Figure 4 – *i*PLS result: correlation graph for the 20th interval using 2 latent variables.

The best prediction model for the protein content finding was obtained combining 6 intervals, having presented a 0.4665% RMSEP (Figures 5 and 6). For this model the technique used was *bi*PLS, with 4 latent variables, considering the same 20 intervals of *i*PLS. Comparing to the PLS model (full spectra) there was a 16% reduction in the root mean square error of validation.

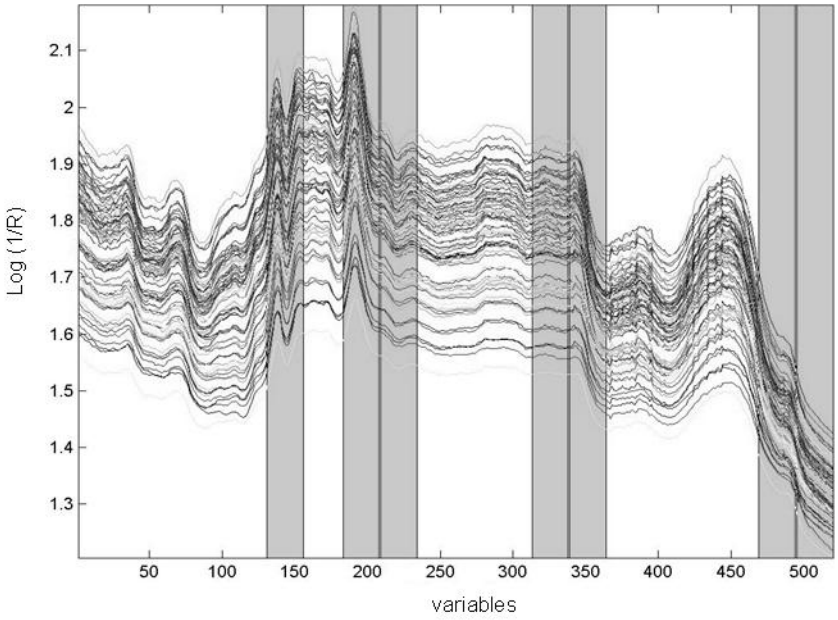

Figure 5 – Spectral wavelenght range used for the *bi*PLS modelling.

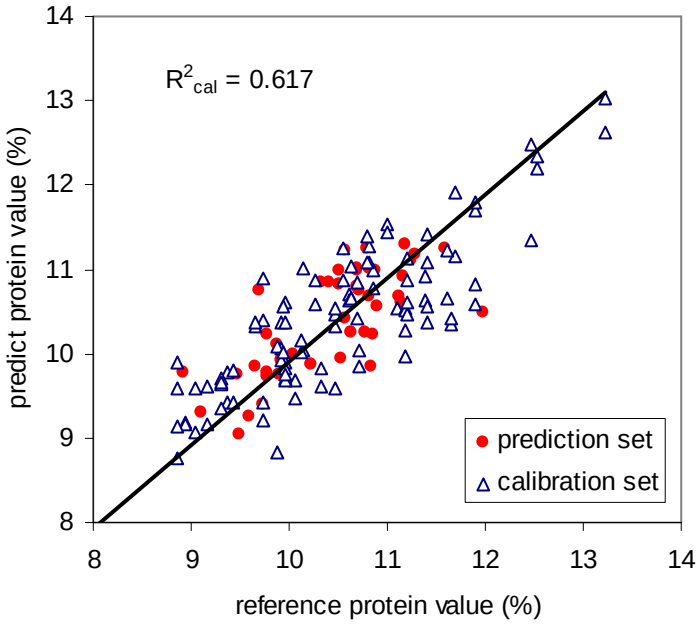

Figure 6 – *bi*PLS result: correlation graph using 4 latent variables.

Conclusions

This article applied the deterministic algorithms of multivariate calibration PLS, *i*PLS and *bi*PLS for Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) of wheat flour samples aiming at the content determination. The obtained results show the importance of the spectral wavelenghts choice in order to obtain optimized models, as well as the multivariate calibration algorithm to be employed.

The comparison between the three algorithms has showed that *bi*PLS led to the most robust models, that is, with the smallest root mean square error of validation, demonstrating that the parameter to be estimated, in this case the protein content, is dependent of several signals representend in the infrared spectrum. From the mathematical point of view, the combination of the different spectral ranges (intervals) has promoted a synergetic effect that can be observed at the increase of the prediction capacity when *bi*PLS is used.

At last, having reached all the proposed objectives in the beginning of this article, the protein content in wheat flour samples can be estimated by Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS), modeled by the algorithm *bi*PLS, resulting in a fast and clean methodology, once it doesn't use chemical reagents to acquire the spectra and doesn't destroy the employed samples.

References

AGRIANUAL. São Paulo:FNP Consultoria e Comércio. 2003. p.479-487.

AOAC, Official Methods of Analysis of the Association of Official Analytical Chemists, Association of Official Analytical Chemists, Arlington, 1990.

BORIN, A. & POPPI, R.J. *Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil*. Vibrational Spectrocopy, n. 37, p. 27-32, 2005.

CISQ/IBILCE Universidade do Estado de São Paulo-UNESP. Disponível em: http://www.qca.ibilce.unesp.br/prevencao/produtos/msds.html Acesso em: 18 de abril. 2007.

EMBRAPA, acessado no dia 06 de abril de 2007 às 10hs, link de acesso: http://www.cnpso.embrapa.br/index.php?op_page=43&cod_pai=66

FERRÃO, M.F. *Aplicação de técnicas espectroscópicas de reflexão no infravermelho no controle de qualidade de farinha de trigo.* Ph.D Thesis, Instituto de Química, Universidade Estadual de Campinas, Campinas, 2000.

FERRÃO, M.F.; CARVALHO, C.W.; MÜLLER, E.I. & DAVANZO,C.U. *Determinação simultânea dos teores de cinza e proteína em farinha de trigo empregando NIRR-PLS e DRIFT-PLS*. Ciência e Tecnologia de Alimentos, v.23, p.333-340, 2004.

KONZEN, P.H.A.; FURTADO, J.C.; CARVALHO, C.W.; FERRÃO, M.F.; MOLZ, R.F.; BASSANI, I.A. & HÜNING, S.L. *Otimização de métodos de controle de qualidade de fármacos usando algoritmos genéticos e busca tabu.* Pesquisa Operacional, v. 23, p.189-207, 2003.

NORGAARD, L.; SAUDALAND, A. WAGNER, J.; NIELSEN, J.P.; MUNCK, L. & ENGELSEN, S.B. *Interval Partial Least-Squares Regression (iPLS): a comparative chemometric study with an example from Near-Infrared Spectroscopy*. Applied Spectroscopy, v. 54, p. 413-419, 2000.

OSBORNE, B.G. & FEARN, T. *Collaborative evaluation of universal calibration for measurement of protein and moisture in flour by near-infrared reflectance*. Journal of Food Technology, v.18, p.453-460, 1983.

OSBORNE, B.G.; DOUGLAS, S. & FEARN, T. *The application of near-infrared reflectance analysis to rapid flour testing.* Journal of Food Technology, v.17, p.355-363, 1982.

PASCHOAL, J.; BARBOZA, F.D. & POPPI, R.J. *Analysis of contaminants in lubrificant oil by near infrared spectroscopy and interval partial least-squares*. Journal of Near Infrared Spectroscopy, v.11, p.211-218, 2003.

ROCHA, W.F.C.; FERRÃO, M.F. & POPPI, R.J. *Modelos de regressão empregando iPLS e DRIFTS para determinação de Carbamazepina em medicamento.* Tecno-Lógica, v.10, p.89-101, 2006.

SCHILLER, G.W. *Bakery flour specifications*. Cereal Foods World, v.29, n.10, p.647-651, Oct. 1984.