Facultat de Química Tarragona

UNIVERSITAT ROVIRA i VIRGILI

Rebeca Ferrer Campos

DEVELOPMENT OF A CHEMOMETRIC MODEL TO PREDICT THE PERCENTAGE IN MASS OF WATER IN TRIOLS BY MEANS OF NIR SPECTROMETRY

BACHELOR'S THESIS

Directed by César Sanz Álvarez

and

Supervised by Dra. Maria Dolores González Candela

IQOXE
Industrias Químicas del Óxido de Etileno

Tarragona

2021

INDEX

1. Summary 2
2. Objective 2
3. Introduction 2
3.1. Polyolethers 3
3.2. Karl Fischer titration 4
3.3. NIR spectroscopy 5
3.2.1. NIR foundations 5
3.2.2. NIR spectrometer 8
3.3. Multivariate calibration: PLS 9
3.4. Validation of the model 12
4. Experimental part 13
4.1. Instrument, measurement parameters and product characteristics 13
4.2. Procedure 14
4.2.1. Karl Fischer titration 14
4.2.2. NIR spectrometry measurements 15
4.2.3. Building of the model 15
4.2.4. Testing of the model 16
5. Results and discussion 17
5.1. Model's optimization 20
5.2. Model's validation 22
5.2.1. Cross validation 22
5.2.2. Test set validation 24
5.2.3. Comparison of validation methods 26
5.3. Testing the model 26
6. Conclusions 28
7. Bibliography 29
8. Annexes 30
8.1. Karl Fischer titrator 30
8.2. NIR spectrometer 31

Resum

Els poliol polièters, també anomenats poliols, són unes de les matèries primes principals per la producció de poliuretà, que és un material versàtil i segur, amb un gran nombre d'aplicacions, que van des de productes industrials fins a productes que s'utilitzen en el dia a dia i que ajuden a fer les nostres vides més practiques i còmodes. El poliuretà és un material plàstic que pot presentar-se en forma rígida o flexible, depenent del poliol emprat en la seva producció. Dins de la industria petroquímica de Tarragona, l'empresa IQOXE, que és una de les principals empreses petroquímiques del sector, produeix aquest tipus de producte.

Aquests poliols han de complir certes especificacions de qualitat per a que el producte final sigui de bona qualitat i una de les especificacions més crítiques és la quantitat d'aigua que conté el poliol. Els mètodes que la ISO (International Organization for Standardization) [1] especifica per aquest anàlisi de quantificació són la mesura per Karl Fischer o un procediment amperometric y coulombimetric automatitzat, però aquest anàlisi es podria fer també per espectroscòpia NIR. Per aquesta raó, en aquest Treball de Fi de Grau es proposa una quantificació d'aigua en poliols Alcupol per espectroscòpia NIR i el desenvolupament de un model quimiomètric PLS (Partial Least Squares) per a la predicció del contingut d'aigua en aquestes mostres de poliol.

1. Summary

Polyol polyethers, also called polyols, are one of the main raw materials to produce polyurethane, which are versatile and safe materials with a wide range of applications (from industrial to daily used products) that help make our lives more practical and comfortable. Polyurethane is a plastic material that can be presented as a rigid or flexible material, depending on the polyol used for its production. IQOXE, one of the main petrochemical companies from Tarragona's petrochemical industrial state, produces this kind of product.

These polyols need to meet certain quality specifications for the final product to have a good quality and one of the most critical specifications is the amount of water they contain. The methods specified by the ISO (International Organization for Standardization) [1] for this quantification analysis are the use of a Karl Fischer measurement or an automated amperometric and coulometric procedures, but it could also be done with a NIR spectrometer. For this, a quantification of water in Alcupol polyols by NIR spectrometry and the building of a chemometric PLS (Partial Least Squares) model for the prediction of water contents in these polyol samples is proposed in this bachelor's thesis.

2. Objective

The objective of this bachelor's thesis is to build and develop a chemometric model able to predict the amount of water in mass percentage (% m/m) of the polyol of interest produced in the company IQOXE.

The model will be validated by both, internal and external calibrations, to compare both validation results, and the results of the prediction will be compared with the ones obtained with the official method for this analysis in this company, which is Karl Fischer titration, to assure the precision of the built model.

3. Introduction

The company IQOXE [2] (*Industrias Químicas del Óxido de Etileno*) is one of the 24 companies that belong to *CL Grupo Industrial* [3], an industrial group that not only is present in the chemical industry, but it is also present in other types of industry, renewable energies, and the consumption field.

IQOXE is one of the many petrochemical companies in the petrochemical industrial estate of Tarragona and it produces ethylene oxide, glycols, and ethylene oxide derivatives. Nowadays, IQOXE is the only ethylene oxide producer in Spain and half of its production is set aside for the manufacturing of glycols, one of the main raw materials used in the production of PET polymers.

As many other companies from the petrochemical sector, IQOXE also produces other products in collaboration with different companies of the petrochemical industrial state. It is a relationship in which both companies produce different compounds that together can be used as reagents for the production of other sellable products. In this case, the product of interest for this bachelor's thesis is the triol, a type of polyol which is produced jointly by IQOXE and Repsol.

The process of production of this polyol and the recipe are confidential information, but a general procedure, some aspects about the process of polyol production and its characteristics can be explained.

3.1. Polyol polyethers

Polyols are organic compounds that contain two or more hydroxyl groups and are obtained from the polymerisation reaction between glycerine, propylene glycol or sugar and ethylene and/or propylene oxide. The general procedure of production is the following:

1) The raw material (glycerine, propylene glycol or sugar) is mixed with an alkaline catalyst, which is usually KOH or NaOH, in a pre-reactor.

2) The mixture is heated and dehydrated under vacuum.

3) The mixture is transferred to the reactor, where an inert atmosphere of nitrogen is created.

4) The ethylene and/or propylene oxide is added gradually, as the reaction is very exothermic, and the mixture is made to be circulating through a cooler.

5) Once there is no more oxide to react, the product is transferred to a post-reactor where the alkaline catalyst is neutralised with an acid.

6) The polyol is mixed with water to solubilise the neutralisation salts and then, the product is dehydrated under vacuum and the polyol is filtered to eliminate the salts.

7) The product is sent to an intermediate tank, where samples are collected for quality analysis. If the results are correct, the polyol is sent to a storage tank to be later transported to the client.

The product is a trifunctional polyolether [4] with a high ethylene oxide content, also called triol. It is a flexible polyol with 32 mg KOH/g of polyol and a viscosity of 1350 cP at 25ºC. It is a clear, colourless liquid with no impurities and a fire point at 256 ºC.

This polyol is highly hygroscopic, so the containers where it is stored must be sealed to protect it against moisture. The optimal storage temperature is between 10 and 35 ºC, as with temperatures below 10 ºC the viscosity increases greatly, making it harder to handle, and at temperatures above 35 ºC the product might degrade, acquiring a yellow-brownish colour. To avoid its oxidation, it is also recommended to storage it under nitrogen.

Triols are mainly used for the production of hyper soft foams and flexible polyurethane foams, although they are also used for rigid foams. Other applications are their use as cell openers in slabstock and in polyurethane foams moulding. Because this polyol is BHT-free (butylated hydroxytoluene-free) and has instead other antioxidant agents with very low migration, it is specially recommended for applications where a minimum '*fogging*' effect is needed, like in the automobile industry. As triols are used to made foam, its applications also expand to household commodities, among other fields.

The quality analyses that this product needs to pass in order to be sellable are the following:

- *Colour:* it is measured performing a comparison of Nessler tubes with standards of platinum-cobalt or with a small portable spectrometer and the maximum is of 50 Hazen units.
- *Hydroxyl number:* it is measured by NIR spectrometry or by performing an acid-base titration, and the hydroxyl index must be of around 32 mg KOH/g of polyol.
- *Acidity:* it is measured by performing an acid-base titration using bromothymol blue and it must not exceed 0.08 mg KOH/g polyol.
- *Water content:* it is measured by a Karl Fischer titration and the maximum permitted is 0.1 % in mass (m/m) of water.

3.2. Karl Fischer titration

The determination of the water content is one of the most used methods in laboratories all around the world, and there are two established methods: the drying methods, which have some disadvantages with respect to the titration methods, that will only determine water when there are not any side reactions. As the name well indicates, Karl Fischer titration [5] corresponds to the second group and has taken a place in many laboratories since it was introduced more than 60 years ago. This titration, also known as KF titration, allows to determine both, free and bound water, work over a wide range of concentration and it gives reproducible and correct results.

When Karl Fischer developed the method, he considered the Bunsen reaction (*Reaction 1*, used for the determination of sulphur dioxide in aqueous solutions) which can also be used to determine water if the sulphur oxide is in excess and the produced acids are neutralised by a base (Fischer used pyridine as he had it on hand).

$$SO_2 + I_2 + 2\,H_2O \rightarrow\ H_2SO_4 + 2\,HI \qquad \textbf{R 1}$$

At first, Fischer presented *Reaction 2* for the use of his reagent in water determination, but, later on, some studies established that the ratio used for the reaction was wrong, as Fischer had made the assumption of having an aqueous Bunsen reaction. Further studies demonstrated that, contrary to what Fischer thought, the reactive component in the KF reagent is not sulphur dioxide but the monomethyl sulphite ion formed from sulphur dioxide and methanol. They also demonstrated that pyridine only acts as a buffer and does not take part in the reaction, so any other suitable base (RN) can replace it. All this gave place to the reformulated and actual KF reaction (*Reaction 3*).

$$2\,H_2O + I_2 + SO_2\ x\ (C_5H_5N)_2 + 2\,C_5H_5N \rightarrow (C_5H_5N)_2\ x\ H_2SO_4 + 2\,C_5H_5N\ x\ HI \qquad \textbf{R 2}$$

$$H_2O + I_2 + [RNH]^+\ SO_3CH_3^- + 2\,RN \rightarrow [RNH]^+\ SO_4CH_3^- + 2\,[RNH]^+\ I^- \qquad \textbf{R 3}$$

The classical KF reagent was an iodine and sulphur dioxide solution in a mixture of methanol and pyridine, but due to the instability of the titer, its odour and its toxicity, in 2002, a research was carried out and it concluded with the replacement of pyridine and methanol by imidazole ($C_3H_4N_2$) and ethanol, respectively, even though nowadays there is a wide range of KF reagents and auxiliary solutions.

3.3. NIR spectroscopy

3.2.1. NIR foundations

The infrared spectral region [6, 7] was discovered by the astronomer Friedrich William Herschel in 1800, but the first analytical applications did not appear until a century and a half later with the creation of the first commercial spectrometers. Infrared spectroscopy measures the absorbance of infrared light by a substance, which depends on the light wavelength.

Infrared light is a type of electromagnetic radiation and when matter is exposed to it, this radiation can be absorbed, reflected, transmitted, scattered or undergo photoluminescence. In the case of IR spectroscopy, it is the absorption of infrared light a substance can absorb that is being measured, which produces the excitation of molecular vibrations and rotations.

The infrared region is divided in three different zones: the far infrared (FIR), where the radiation absorption is produced because of molecular rotations, the mid infrared (MIR), where the absorption is due to fundamental molecular vibrations, and the near infrared (NIR), where absorption is caused by overtones and combination bands of fundamental molecular vibrations.

The movements of vibration and rotation of a molecule provoke changes in the molecule's dipolar moment, and these result in the absorption of radiation. This happens because, when a substance is irradiated, the electromagnetic field of the radiation interacts with the electric field generated by the change in the dipole moment. Then, when the radiation frequency equals the natural vibration frequency of the molecule, there is an energy exchange, making the amplitude of the molecular vibration change and causing the absorption of said radiation.

The atoms of a molecule are always in movement, which can be described by the model of a simple harmonic oscillator. This approximation relates the distance caused by the movement of the atoms of a molecule with its potential energy. The atoms' movement is confined within a potential well, where the minimum will be achieved when the atoms are in equilibrium and the maxima when they are too close or too far from each other (*Figure 1*).

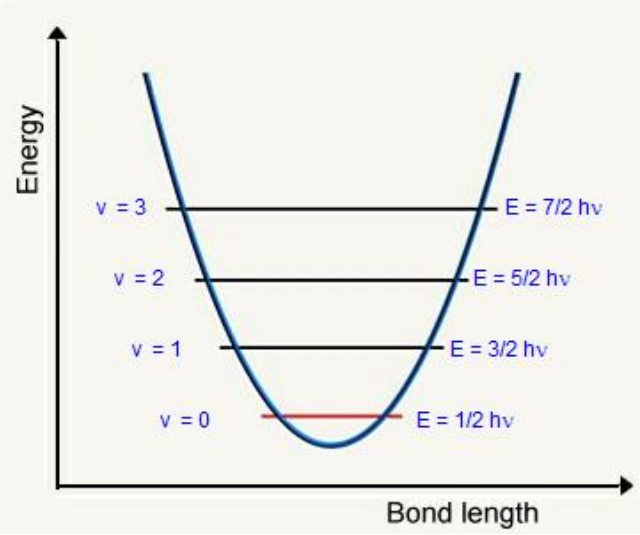


Figure 1. Potential energy in a simple harmonic oscillator [6]

So, the potential energy defined by this approximation is the following:

$$E = \frac{1}{2} \cdot k \cdot x^2 \qquad \textbf{Eq. 1}$$

Where the force constant ***k*** is representing the bond strength and ***x*** the distance between the atoms.

In a system where two masses are joined by an elastic spring, the vibration frequency is described by the following equation:

$$v = \sqrt{\frac{1}{2\pi}\frac{k(m_1+m_2)}{m_1 m_2}} \qquad \textbf{Eq. 2}$$

But these classical mechanics equations do not describe the atomic particles behaviour completely, as they do not take into consideration the quantised nature of this particles. That is where the solution to the Schrödinger equation comes into play, making the vibrational energy be now described as:

$$E_{vib} = \left(n + \frac{1}{2}\right) hv \qquad \textbf{Eq. 3}$$

Where ***n*** is the quantic vibrational number, ***h*** is Planck's constant and ***v*** is the vibrational frequency.

If *Equation 2 and 3* are combined, the vibrational energy of a diatomic molecule can be defined, taking into account its quantised nature:

$$v = \left(n + \frac{1}{2}\right) \frac{h}{2\pi} \sqrt{\frac{k(m_1+m_2)}{m_1 m_2}} \qquad \textbf{Eq. 4}$$

Even after all these considerations, the harmonic oscillator model is still not good enough to describe real molecules, as molecules' behaviour is better represented by an anharmonic oscillator (*Figure 2*) due to the repulsion they generate when an atom approaches another, producing a faster increase of potential energy, and the breakage of the bond (dissociation) when the distance between two atoms increases, producing a decrease of potential energy.

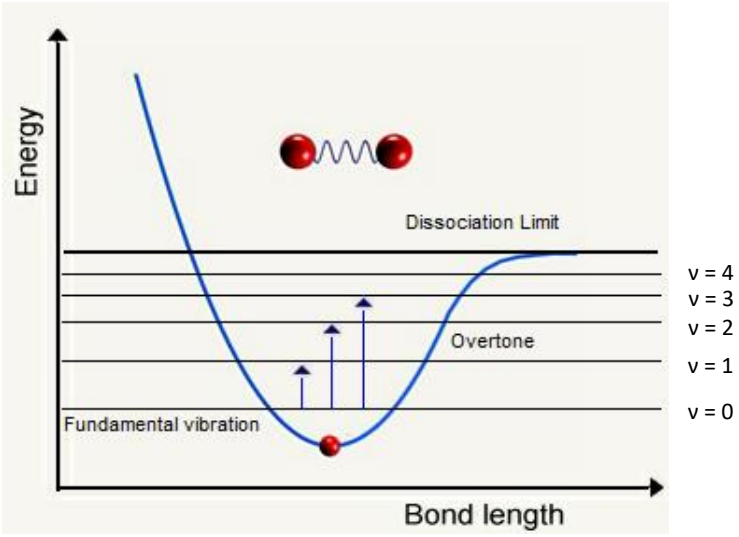


Figure 2. Potential energy in an anharmonic oscillator [6]

Now, to describe this anharmonicity of molecules, *Equation 5* below is used:

$$E_{vib} = \left(n + \frac{1}{2}\right) hv - \left(n + \frac{1}{2}\right)^2 hvy - \left(n + \frac{1}{2}\right)^3 hvyy' \ldots \quad \textbf{Eq. 5}$$

Where ***n*** is the quantic vibrational number, ***h*** is Planck's constant, ***v*** is the vibrational frequency and ***y*** and ***y'*** are anharmonicity constants.

Differently form harmonic oscillator, the anharmonic oscillator energy levels are not equally separated from each other, meaning, not only fundamental bands can be observed (Δn = ±1), but other transitions (Δn = ±2, ±3, …), called overtones, can also be seen. These last transitions can be seen in NIR and appear at larger wavelengths. The absorption in the NIR region will happen when the NIR radiation energy is equal to the energy difference between 2 vibrational levels and a change in the dipolar moment occurs. So, in the NIR region, fundamental bands do not appear, only overtones (Δn > ±1), generally just the first and second ones, and combination bands, produced in polyatomic molecules when there are simultaneous energy changes in two or more vibration modes, making them interact with each other.

NIR bands tend to be wide and not very well defined due to overtones and combination bands overlapping and they show a lower intensity than bands in other IR regions. The intensity of the bands in NIR is related to the anharmonicity of the bonds: the more anharmonic a bond is, the higher the probability to see overtones and combination bands. In *Figure 3*, the most common absorption bands are displayed.

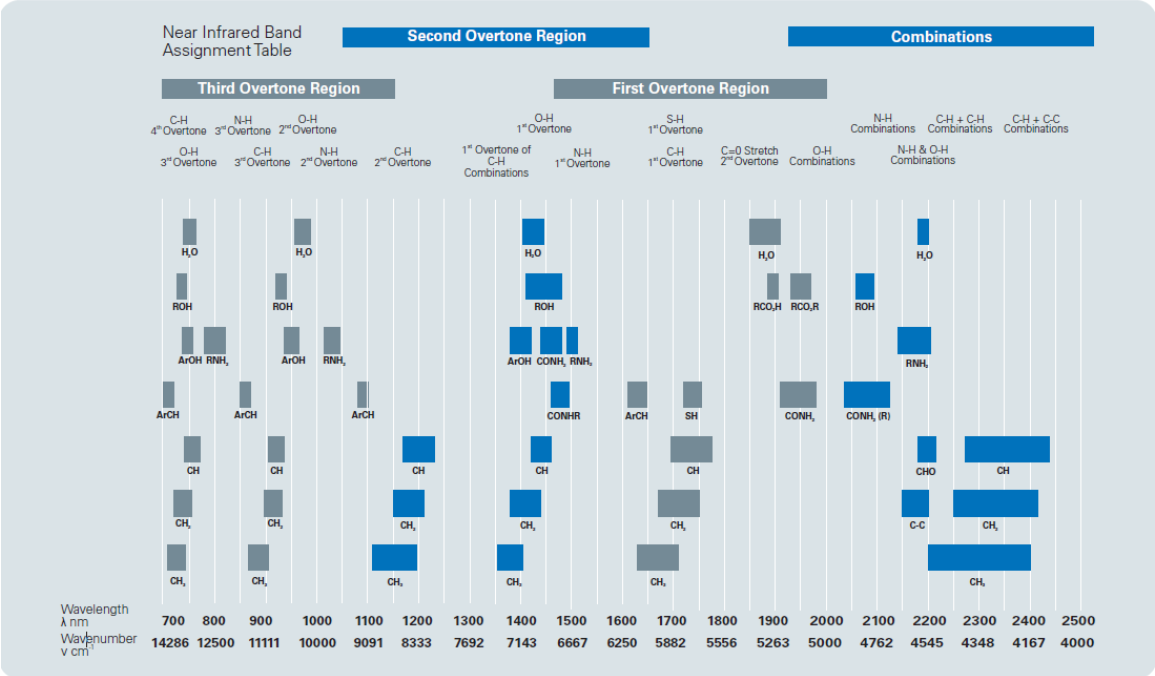


Figure 3. NIR most common absorption bands [8]

3.2.2. NIR spectrometer

The first spectrometers or first-generation spectrometers were dispersive and used prisms as dispersive elements, which later changed over to gratings.

In mid 1960s, second-generation spectrometers appeared using what is called Fourier Transform (FT-IR) and integrated a Michelson interferometer [6] (*Figure 4)*, which ideally transmits the 50% of the light received from the light source and reflects the rest.

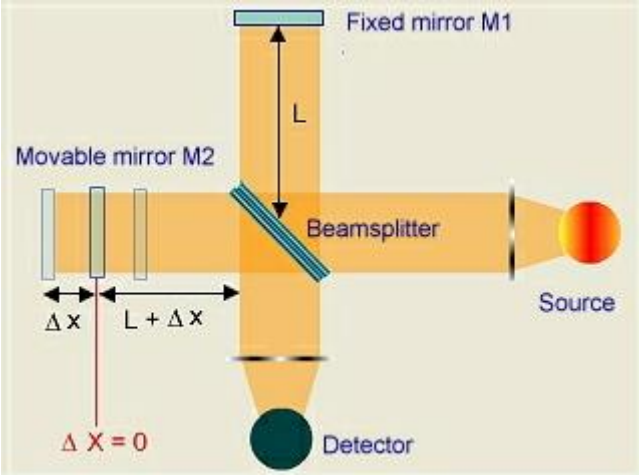


Figure 4. Michelson interferometer [6]

In these spectrometers, the light emitted from a source is directed into an interferometer that modulates the light. Then, the light passes through the sample in the sample compartment and is focused on the detector. The signal the detector measures is called interferogram (*Figure 5)*.

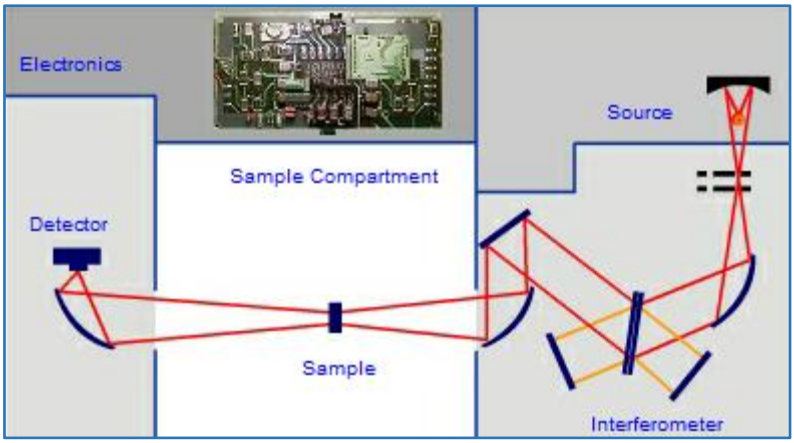


Figure 5. FT-IR spectrometer layout [6]

The NIR spectrometer consists of the four following parts [7], in each of which the characteristics of the spectrometer that will be used for this bachelor's thesis (TANGO-T System (FT-NIR) from Bruker) will be specified:

- *Radiation source:* used to irradiate the sample. The spectrometer used has an air cooled tungsten halogen lamp. It is one of the most used light sources and it can provide a continuous spectrum in the region between 320 – 2500 nm.
- *Wavelength selection system:* depending on the selected wavelength selection system, the instrument can be dispersive or non-dispersive. In this case, the system is non-dispersive and uses FT-IR.
- *Sample compartment:* there are 3 types of NIR spectra registers: transmittance, reflectance and transflectance. In this case, the used one is the transmittance mode, with which the radiation absorption follows Beer-Lambert's law.
- *Detector:* measures the signal. The InGaAs diode used is one of the most common detectors in NIR spectroscopy.

3.3. Multivariate calibration: PLS

In general terms, all quantitative analytical methods [9] have the objective of determining a property *y* of a system from a measured parameter *x* of that same system. This kind of determination usually consists in two steps: first, the calibration, which looks for a correlation (shown in *Equation 6*) between the measured parameter *x* and the property *y*, and then, the analysis or prediction of the property *y*.

$$y = x \cdot b \qquad \textbf{Eq. 6}$$

The parameter *b* is called regression coefficient and is defined by *Equation 7*, in which the individual parameters *X* and *Y* are expressed in matrix form, and *T* indicates the transposed of the matrix.

$$b = (X^T \cdot X)^{-1} \cdot X^T \cdot Y \qquad \textbf{Eq. 7}$$

In this case, as *X* and *Y* are going to represent near-infrared (NIR) absorption spectroscopic data, the spectral intensities will be written point by point into the *X*-matrix in rows, each row corresponding to a different sample, while the corresponding component values of each of those samples will be written in rows into the *Y*-matrix.

There are two types of calibrations that can be done: the univariate and the multivariate calibrations. In the first case, with univariate calibration, a calibration function *b* is obtained from plotting the absorbance values of the peak maximum A_i versus the concentration of the analyte C_i (*Figure 6*).

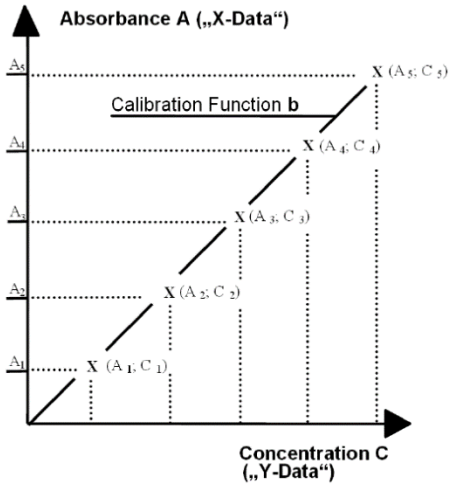


Figure 6. Calibration of absorbance spectra [9]

Then, each unknown sample is measured spectroscopically and a correlation of each of their A_i values with the calibration function *b* gives the corresponding concentration value of the analyte for each of the samples. Generally, a univariate calibration ends up not having enough prediction capability due to the following aspects:

- Cannot recognise outliers or unknown interfering substances.
- The uncertainty must be minimised by multiple sample measurements and subsequent averaging.
- In multi-component systems, the peak maxima are often not separated enough.
- In multi-component analysis, because of the use of the Beer-Lambert law, for many systems it is not valid.

Moreover, univariate calibration methods usually lead to useless results for multi-component systems. This is the reason why multivariate calibration methods, like MLR, PCR or PLS, are used in those cases.

In the case of this bachelor's thesis, a PLS-regression is the method that is going to be used, so it is the only one that will be described.

The PLS [9,10] algorithm is the most used one and it is the major regression technique for multivariate calibration. It has a considerable mathematical complexity because it is based on the fact that information can be obtained without significant loss in a smaller number of variables. For a PLS regression to be used, many samples must be measured to be able to compare the spectral information of the substance with the corresponding concentration values, so that, if changes occur, both data points are correlated and recognised by each other.

Before doing a PLS, generally, a centring of the data matrices is done by subtracting the mean of each column, but, in reality, there is no scientific need to perform this centring. This centring comes from the first applications of PLS in NIR spectroscopy, where due to spectroscopic problems, like ones caused by baselines, the centring of the data matrices is used. However, there are many fields where uncentered PLS is used. The program that will be used for the development of the model in this bachelor's thesis does automatically the centring.

In PLS, both data sets are written down in a data point matrix and the generated eigenvectors are sorted in descending order (*Figure 7*). The first factor has the greatest importance for the calibration model and characterises the main changes of the spectrum. The more factors used, the smaller the changes that are characterised until the higher factors characterise the spectral noise (overfitting), but if too few factors are used, there will be a lack of spectral information (underfitting). This fact makes the selection of the optimum number of factors to be of very high importance for the quality of the model.

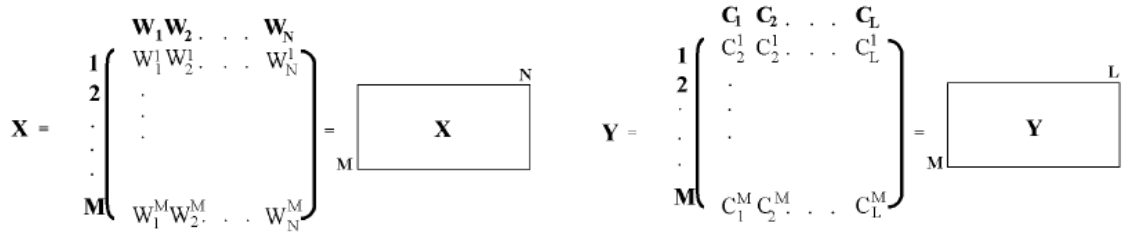


Figure 7. Structure of data point matrices X and Y [9]

In a PLS regression, the data matrix *X* contains the spectral data while the data matrix *Y* contains the concentration data. They are both reduced to a few factors and they are then represented by *Equations 8 and 9*, where $\boldsymbol{t_i}$ are the scores vectors, $\boldsymbol{p_i}$ and $\boldsymbol{q_i}$ are the loadings vectors respectively, ***R*** is the number of factors, ***T*** indicates the transposed of the corresponding vector, and ***F*** and ***G*** are the residual matrices of the corresponding data respectively.

$$\mathrm{X} = t_1 p_1^T + t_2 p_2^T + t_3 p_3^T + \cdots t_R p_R^T + F \qquad \textbf{Eq. 8}$$

$$\mathrm{Y} = t_1 p_1^T + t_2 p_2^T + t_3 p_3^T + \cdots t_R p_R^T + G \qquad \textbf{Eq. 9}$$

Usually, the system is over-determined, which means that the number of absorbance values measured is much larger than the number of components present. This translates in the allowability of the system to be correlated to the whole spectral data structure instead of only to a single spectral data point. It also means that the information obtained is much more than with univariate calibration and it is possible to determine outliers during the analysis.

In addition, the PLS method assumes that, at the given number of factors, the scores vectors for both data sets are identical so that the deviation from the original values is as small as possible. This is because a change in the spectral data should result in a change in the spectrum, making the score vectors identical, but in real samples human and/or instrumental errors will lead to different score vectors.

There are two general types of PLS: PLS 1 and PLS 2. The only difference between them is that, in the second case, a concentration matrix is used instead of concentration vectors like in PLS 1, so in PLS 1 one compound is modelled at a time, while in PLS 2 all known compounds are included in the model simultaneously. There are some cases where PLS 2 is used but its prediction capability is worse than that of PLS 1, so a later individual performance of PLS 1 is done. In the case of this bachelor's thesis, the program that is going to be used exclusively uses PLS 1.

3.4. Validation of the model

Once the multivariate calibration is done [9,11], the spectra of the samples is measured and combined with the calibration function *b* obtained (*Equation 10*), directly giving the concentrations of the analyte from the corresponding spectra.

$$Y_{analysis} = X_{analysis} \cdot b \qquad \textbf{Eq. 10}$$

It is necessary to find the calibration function *b* that provides the better correlation between the spectral and concentration data to obtain results as exact as possible and, in consequence, have the best analysis feasible. That is why the built model must be validated.

The validation consists in a prediction of several samples, from which the concentration of the analyte to analyse is known, using the model to be validated. Then, the predicted values are compared to the real ones to prove the precision of the model using different model parameters and the one that gives the smallest prediction error is the one is the one that characterises the model better. The validation allows to determine the optimum number of factors to be used and to identify the possible outliers.

There are two types of validation: cross validation (CV), which is an internal calibration, and test set validation, that is an external calibration. In the first case, a sample is taken from the calibration data set and a chemometric model is built using the remaining samples (*Figure 8*). Then, the taken sample is predicted with the model and its value is compared to the real value. This is done with all the samples in the calibration data set to assess the precision of the model, as the taken samples are not used to build the model and thus, are independent from it. With this kind of validation, it is possible to calculate a quantitative measure for the average accuracy of prediction of the model, called RMSECV (Root Mean Square Error of Cross Validation). The smaller the RMSECV value, the better the model's quality.

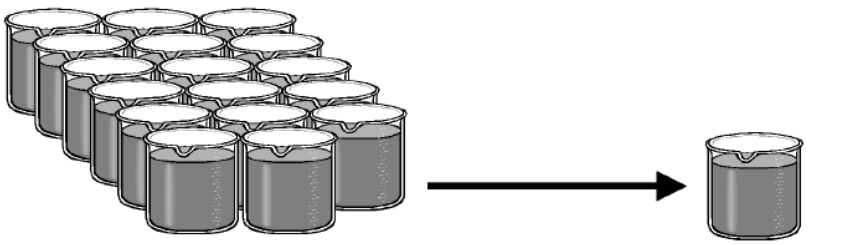


Figure 8. Scheme of a cross validation [9]

For the test set validation, the complete data set is used to build the model and it will remain constant for further validation as there is no sample being removed from the calibration data set. So, to do an estimation of the prediction error, more samples are measured to form another set, called test set (*Figure 9*), and those samples are the ones to be analysed by the method developed with the calibration data set. In this case, the predicted values of the test set are compared to their real values and the RMSEP (Root Mean Square Error of Prediction) is calculated, which is a quantitative measure of the prediction accuracy of the model. As with RMSECV, the smaller the values for the RMSEP, the better the model's quality.

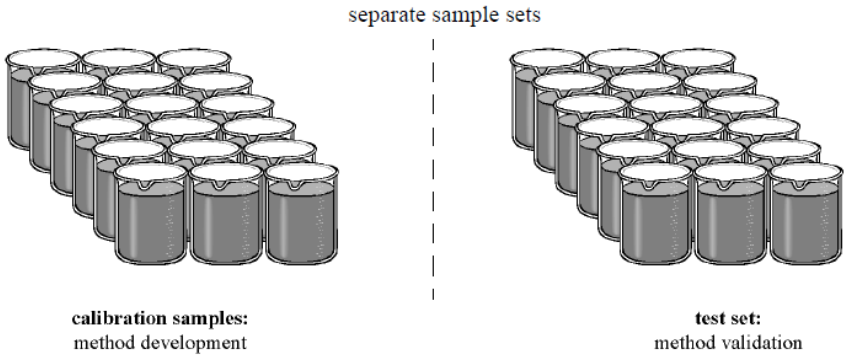


Figure 9. Scheme of a test set validation [9]

Comparing both methods, for the test set validation many more samples are needed than with cross validation and there is no exclusion of samples, but in this last method (CV) only the calibration data set is needed, so it can be used with a limited number of samples. Both methods of validation should lead to comparable results. If not, then more samples are to be analysed to establish a reliable method.

4. Experimental part

4.1. Instrument, measurement parameters and product characteristics

- Software: OPUS [12]
- Instrument: TANGO-T System (FT-NIR)
- Reference substance: ambient air [9]
- Product: trifunctional polyolether (Alcupol F-3231)

Table 1. Product and reagents dangers and handling

Compound	**Dangerousness**	**Handling**
Alcupol F-3231 [13]	–	–
HYDRA-POINT Composite 5 (imidazole, iodine, and sulphur dioxide) [14]	Flammable Acute toxicity Health hazard	Safety glasses Gloves Fume hood
Methanol [15]	Flammable Toxic Health hazard	Safety glasses Gloves Fume hood

4.2. Procedure

4.2.1. Karl Fischer titration

Before doing the NIR spectra analysis to build the model, the concentration of the samples must be known to be able to build the model. For this, each sample to be analysed is titrated with the Karl Fischer reagent (which in this case is made of imidazole, iodine and sulphur dioxide) using an automatic titrator, the program of which will directly display the concentration in % (m/m) of water after the titration is finished. The procedure of this measurement is the following:

1) With a 5 ml syringe, an approximate quantity of 4 to 5 ml of sample is taken.

2) The full syringe is weighted with a balance connected to a small computer with a tactile screen that is connected of the automatic titrator. When the value stabilises, the balance is tared.

3) Then, the option *Start* is selected on the screen and there are 15 seconds to introduce the sample into the cell.

The cell must be half filled with the solvent used, which is methanol in this case, and the drift must be below 0.020 µl/min for the instrument to allow the introduction and subsequent start of the sample measurement. For an accurate measurement, the contents of the cell must be replaced every time before starting the measurement.

4) The cell contains a small cap on the top which is opened to inject the sample into the cell.

5) Immediately after the injection, the cap is closed and the now empty syringe is weighted again.

6) The weight is transferred to the computer or introduced manually, while the titration is, at first, automatically started in second plane.

When the option *Continue* is selected and while the titration is in process, the curve of used Karl Fischer reagent millilitres vs time can be seen, as well as the increasing value of the Karl Fischer reagent millilitres with three decimal figures and the seconds passing.

7) Once the titration has finished, the computer does the corresponding calculation (see *Equation 11* below) and gives the amount of Karl Fischer reagent used and the concentration of water that the sample contains.

$$\%\,(\mathrm{m/m}) = (\mathrm{V} \cdot \mathrm{F})/(10 \cdot \mathrm{m}) \qquad \textbf{Eq. 11}$$

Where:

V: volume of Karl Fischer reagent consumed in the titration in ml

F: factor of Karl Fischer equivalent, specified in the reagent bottle in mg/ml

m: mass of the sample analysed in g

4.2.2. NIR spectrometry measurements

To build the chemometric model, a series of samples with known concentration are measured with the NIR spectrometer and treated with the *Quant2* program that the OPUS software from Bruker Optiks incorporates. As it was explained before, to make the concentration of the samples known, a Karl Fischer titration is done to each of them before the NIR spectrometry measure. The procedure to do a NIR spectrometry measurement is the following:

1) The sample is put in a small, glass, disposable vial, filling 3 quarters of the vial approximately.

2) The disposable vial is closed with a small plastic cap and it is put into a heating device.

The stablished temperature in IQOXE's lab for this measurement is of 50ºC, so the sample needs to achieve that temperature before the NIR spectrometric analysis.

3) While the sample is heating up, a reference measure needs to be done with the NIR spectrometer. This is done simply by taking the screwed, metal cap and doing an ambient air measure, as it is used as reference [9].

4) After 10 to 15 minutes have passed, the sample is removed from the device and introduced immediately into the NIR spectrometer sample slot.

5) A few minutes may be needed for the sample and the NIR spectrometer to adjust their temperatures, and when the sample's temperature is of 50 ± 1ºC, the analysis will begin.

The NIR spectrometer does 31 scans of the sample (by default) and gives the spectrum.

4.2.3. Building of the model

In the *Multivariate Calibration* [9] book provided by the Bruker company, it is stated that: *'subtraction of a straight line, vector normalization or taking the first derivative of a spectrum often leads to optimised PLS models'*. Other information [16,17,18] was found stating that the second derivative of a spectrum also leads to optimised PLS models, as well as the previously mentioned pre-processing methods among others. With all this in mind, for the optimization of the model all these mathematical methods of spectra pre-processing will be considered. The steps used to build the model are the following:

1) The components to build the model are introduced into the program.

2) The spectra are opened in the *Quant2* program and their real concentrations are introduced.

3) A first optimization is required to be done by the program, so the pre-processing methods chosen are:

- Second derivative
- First derivative with vectorial normalization
- Subtraction of a straight line

4) Once the optimization finishes, it gives the lowest RMSECV value and the pre-processing method used to obtain it.

5) The following step is the validation of the optimised model, where internal or external calibration can be chosen.

Both, cross validation and test set validation, can be used for the validation of this model as there are more than enough samples to do a test data set. Both validation methods will be used, and the results will be compared.

For cross validation, one sample is chosen to be excluded from the model at a time, and for test set validation, the data will be divided in half to form the calibration and test sets.

6) As a result of the validation, several plots are obtained: the calibration and the validation plots.

The calibration plots will not really be used, as only the validation ones give information on how to optimise the model.

The *Quant2* program marks in red the points that it considers outliers, which have a very large F-value compared with the other samples and are to be, in principle, taken out to build the model, although they need to be looked at individually to assess if they are really outliers or not.

As it is explained more extensively and detailed in the *Results and discussion* section further on, due to some problems the used spectra lead to, none of them were finally used and another group of samples' spectra was measured and used for the model building.

4.2.3.1. Optimization of the model

A good, optimised model for liquid samples should have a R^2 value larger than 99% when it is validated [9]. There are several parameters that can be modified to optimise the model, for example the data pre-processing method, the recognition and elimination of outliers, etc. In this case, the parameters that will be modified in order to improve the model are the following:

- Concentration range
- Spectral range
- Pre-processing of calibration regions
- Optimization pre-processing

4.2.3.2. Validation of the model

To validate the model both, cross validation and test set validation, will be used. The results obtained from both should be comparable, meaning that they should not differ greatly form one another. The comparison of the results from both validations will assure that the model is good enough for its purpose: prediction.

4.2.4. Testing of the model

This will be the last step and it will be done once the model is proved to be good enough and has been validated. A series of samples with known concentration will be analysed and the

analysis program of the OPUS software, *Quant2 Analysis*, will give the predicted result. It will just be to assure the working condition of the model, as if the validation is good, there should really be no problem with the accuracy of the prediction.

5. Results and discussion

At first, the model was to be built with 496 spectra of different samples collected along the last two years. Out of those, 25 had unknown concentrations, so finally, the model was to be built with the 471 spectra left, which are shown in *Figure 10*.

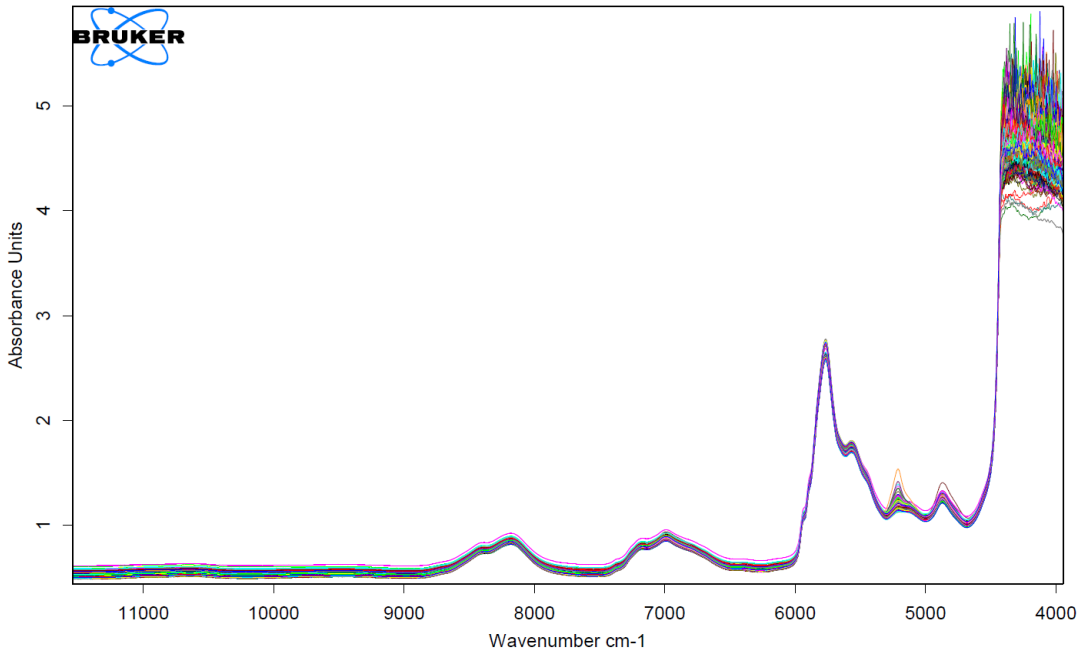


Figure 10. Spectral data used to build the model

There was a problem with all these spectra as the model could not be well enough optimised no matter what pre-processing, concentration range or any other parameter was changed or adjusted. The optimal optimization conditions found for this model are the following:

- Concentration range: 0.027 – 0.300 % (m/m) of water
- Spectral ranges: 10500 – 9300 cm^{-1}; 7100 – 6000 cm^{-1}; 5300 – 4900 cm^{-1}
- Pre-processing of the calibration regions: Second derivative (17 smoothing points by default)
- Optimization pre-processing:
    - Subtraction of a straight line
    - Second derivative
    - First derivative + Vectorial normalization

The optimization results with these conditions can be seen in the following figures, where the validation results achieved are displayed.

As it can be observed in *Figures 11 and 12* below, the rank recommended by the program (in blue) is rank 2. To choose the optimal rank it is needed to look for the minimum in case of *Figure 11*, which is the lowest value for the RMSECV, and to look for the maximum in case of *Figure 12*, which is the highest value for the R^2. In both cases, the minimum and maximum, respectively, would be at rank 4, but there is no significant change of the RMSECV and R^2 after rank 2, and, as it is a lower rank than 4, choosing rank 2 will prevent the model from overfitting.

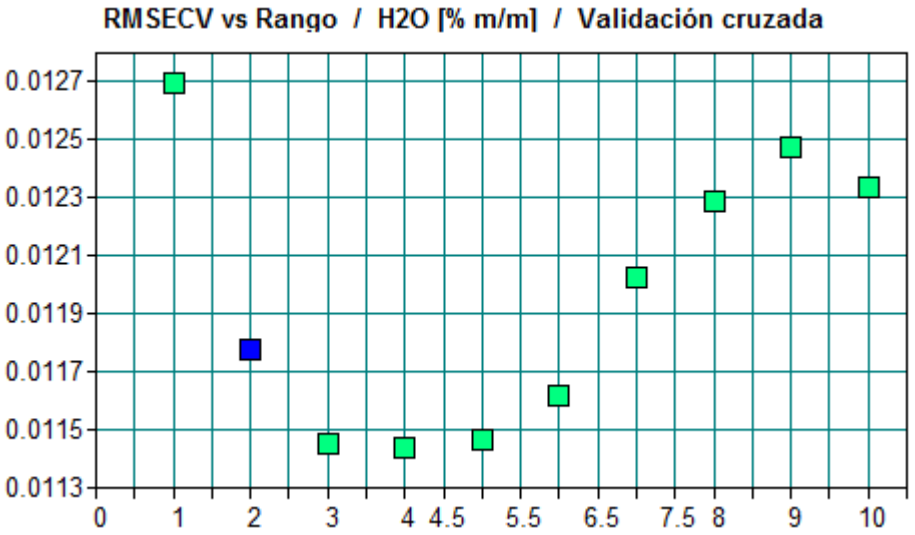


Figure 11. RMSECV vs Rank plot

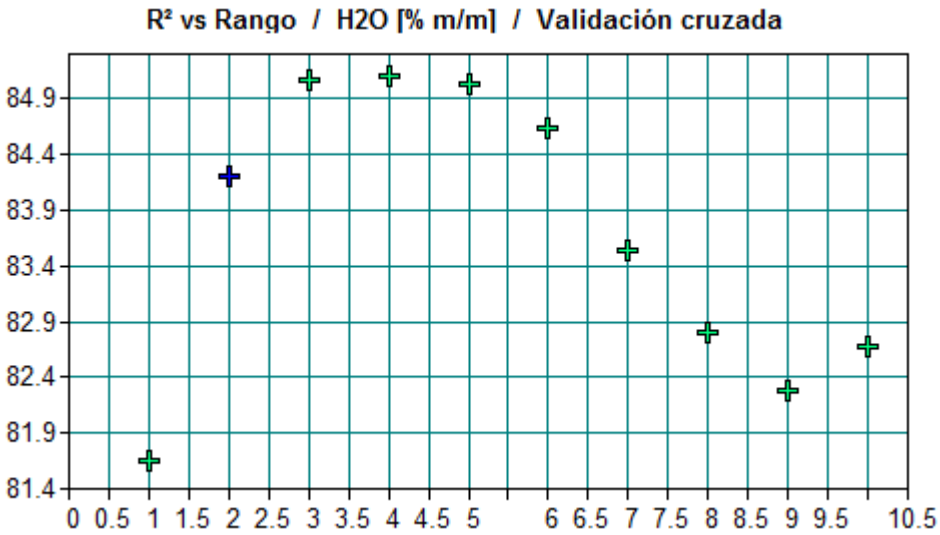


Figure 12. Relative coefficient vs Rank plot

Another factor to take into account to choose the rank is the Residual Prediction Deviation (RPD). The larger this value is, the more accurate the model will be. In this case, the largest value of RPD is for both, rank 3 and 4, but the very small difference between them and rank 2, will make the decision of choosing the last one easier.

As it can be observed in *Figure 13* below, the samples were not homogeneously distributed along the range, so it was not well represented, and the model could not be robust enough to represent this concentration range.

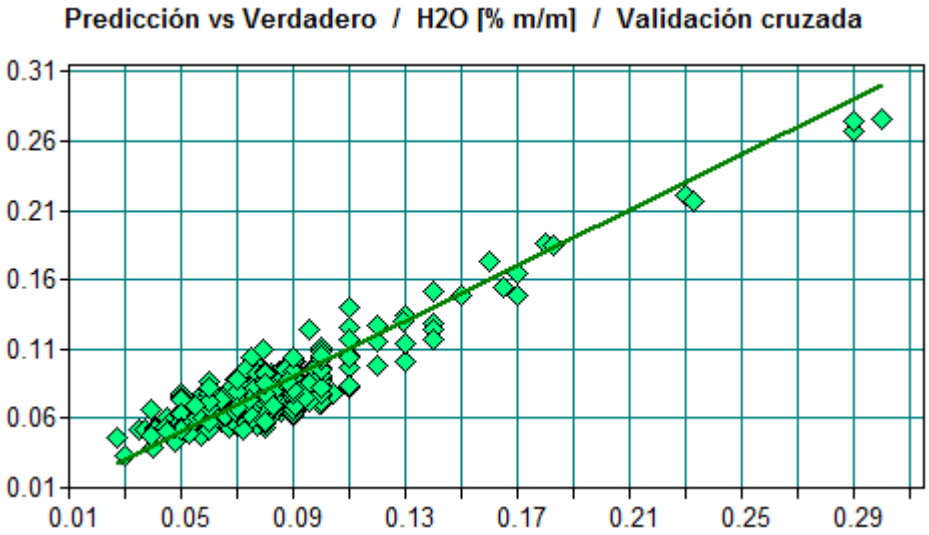


Figure 13. Prediction vs true value plot

When trying to reduce the range and after optimising it, the cloud of samples observed between 0.027 – 0.150% (m/m) of water persisted (*Figure 14*) and the validated calibration curve was even worse than the one from the model before. The optimal rank for this second model was also 2.

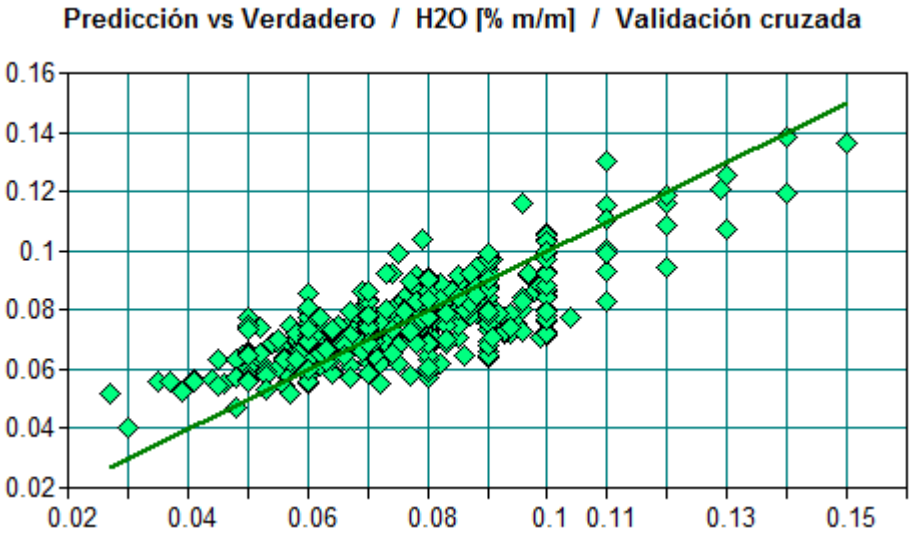


Figure 14. Prediction vs true value plot for restricted range

A summary of the most significant validation results achieved for this optimised model can be seen in *Table 2*.

Table 2. Summary of the validation results of the first model after the optimization

Validation results

R^2	RMSECV	RDP	Opt. Rank
84.20	0.0118	2.52	2

The minimum value of R^2 in the case a model is built for liquid samples must be of 99% or higher for the calibration model to be of sufficient quality, but the maximum value obtained was far from it, as it can be seen in *Table 2*. The RMSECV and RDP values were also quite high and low, respectively, for the model to be accurate. This can mean that the model is not robust enough or that some of the samples' concentrations and/or spectra were not correctly measured. There is also the possibility that, as the product is highly hydroscopic and the concentrations are very small, when doing the Karl Fischer titration and when doing the NIR spectrometry measurement the amount of water has changed slightly because the samples hydrated.

When talking about this problem with my lab mates, the hypothesis we all arrived to was that it may be caused by them not being very strict when doing the measurements, as in the industry they only need an approximate value, and it is good enough while it is below the maximum amount permitted. So, when doing the Karl Fischer titration, they only changed the contents of the cell once it was nearly full and not every time they did a measurement, which can affect said measurement.

To be able to build a good chemometric model for the determination of water contained in this product and test the new aforementioned hypothesis, new samples were measured more carefully and another model was built starting from anew.

5.1. Model's optimization

The building of the new model was started when a few samples were already measured. Then, the other samples were introduced one by one as they were being measured with the NIR spectrometer. As explained before, the samples were already of known concentration because they were previously quantified by Karl Fischer titration when being measured with the NIR spectrometer. This new model finally consisted of 129 spectra, which are showed in *Figure 15*, below.

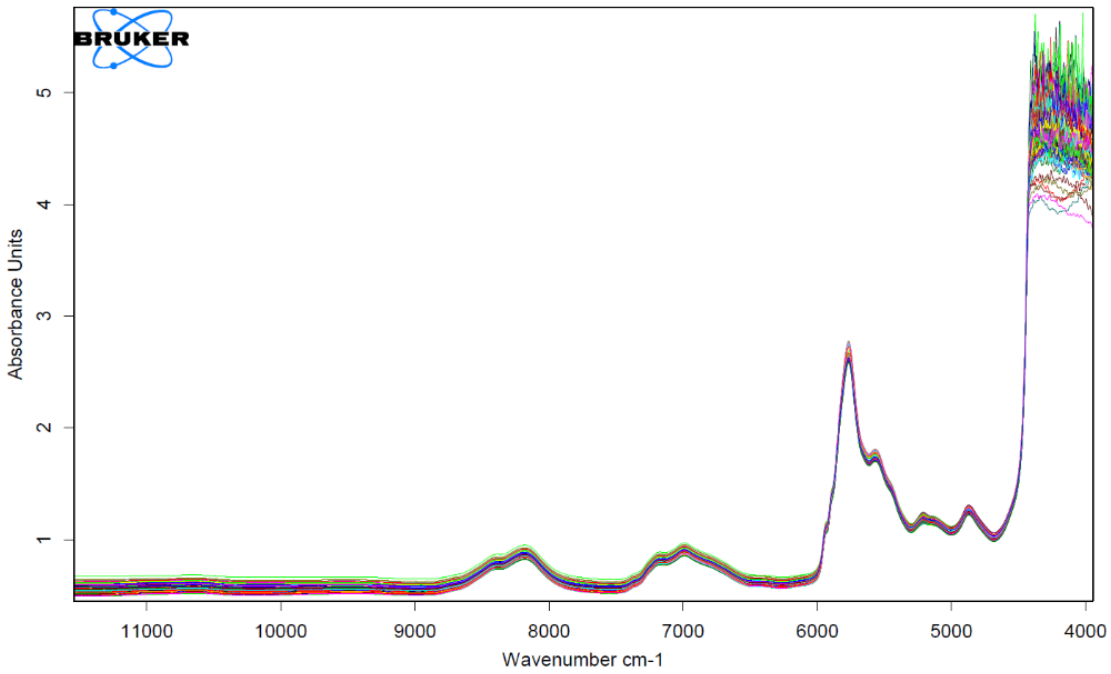


Figure 15. Spectral data used to build the model

The part of the spectra from around 4500 – 4000 cm^{-1} was not considered for any the optimizations due to it being spectral noise generated by strong light loss in the used light fibre.

When trying to find the optimal conditions for the model, the parameters that were changed were the following:

- Concentration ranges tried:
    - 0.027 – 0.100 % (m/m) of water
    - 0.030 – 0.100 % (m/m) of water
    - 0.039 – 0.100 % (m/m) of water
- Spectral ranges tried:
    - 10500 – 9300 cm^{-1}; 7100 – 6000 cm^{-1}; 5300 – 4900 cm^{-1}
    - 10500 – 4500 cm^{-1}
    - 9500 – 4500 cm^{-1}
    - 8900 – 4700 cm^{-1}
- Pre-processing of the calibration regions tried:
    - Subtraction of a straight line
    - Second derivative
    - First derivative
    - Vectorial normalization
    - First derivative + Vectorial normalization
- Optimization pre-processing tried:
    - Subtraction of straight line
    - Second derivative
    - First derivative
    - Vectorial normalization
    - First derivative + Vectorial normalization

These parameters were changed one by one, trying to find the combination that would give the best results. The concentration range was chosen basing on the real concentration values of the samples, as only a couple of them had the maximum concentration and none had a concentration above it. The very small concentration values of 0.027 and 0.030 % (m/m) did not provide good spectra and thus were discarded from the model.

The best results were achieved by the following conditions that were the optimal ones for the model built:

- Concentration range:
    - 0.039 – 0.100 % (m/m) of water
- Spectral range:
    - 8900 – 4700 cm^{-1}
- Pre-processing of the calibration regions:
    - Subtraction of a straight line
- Optimization pre-processing:
    - Subtraction of a straight line

After many tries at optimising, the rank was limited to a maximum of 5, due to the model never giving better results at a rank above this one. This helped to reduce the optimisation and validation times.

5.2. Model's validation

As it is mentioned in the *Experimental part* section, this chemometric model was validated using both, internal and external calibration. For the cross validation (internal calibration), one sample at a time was chosen to be excluded, while in test set validation (external calibration), the spectra were divided randomly in two halves by the program and one half was used as the calibration set and the other as the validation set.

5.2.1. Cross validation

In the case of the cross validation, the optimised model's optimal rank was 3. It may not be quite clear when looking at the maximum and minimum of *Figures 16 and 17*, respectively, as one would probably think that there may be overfitting when choosing this rank instead of rank 2, but the value of RPD for rank 3 (RPD=11.2) is significantly better than the RPD value for rank 2 (RPD=10.2), making rank 3 to be the optimal rank for the validation.

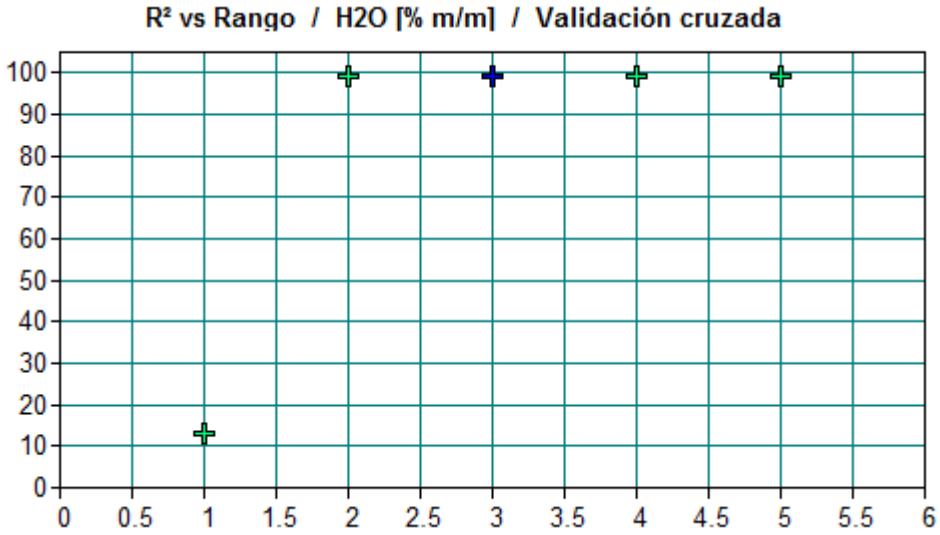


Figure 16. Relative coefficient vs Rank plot for cross validation

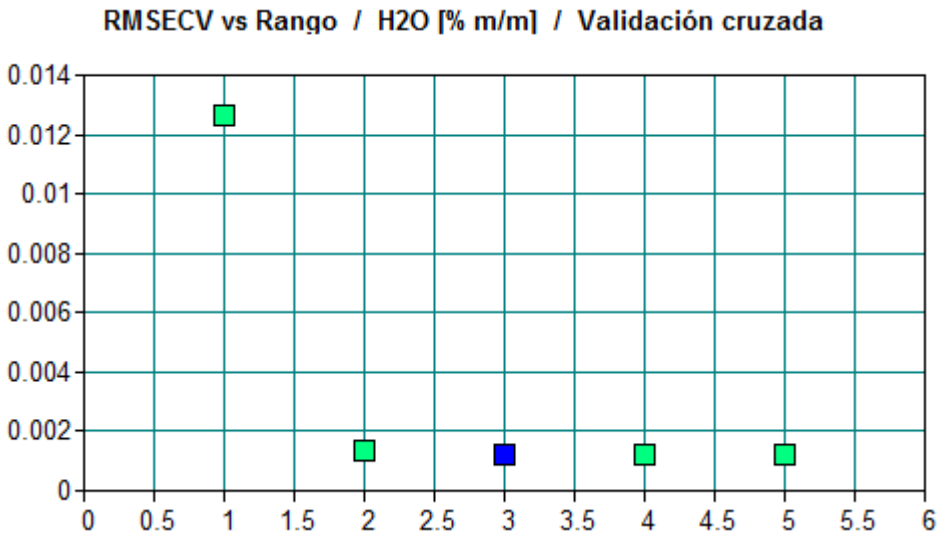


Figure 17. RMSECV vs Rank plot for cross validation

When looking at the Mahalanobis distance in *Figure 18*, it can be seen that one of the samples (Sample 6) has a distance above the limit, which in this case is around 0.16, meaning, this sample could be a possible outlier. When looking closely at the sample, its real concentration value is of 0.085 while its predicted concentration value is of 0.084, and it has an F-value and an F-probability of 0.809 and 0.63, respectively. Taking into account all these values, it can be considered that Sample 6 is not really an outlier, even though its Mahalanobis distance is above the limit.

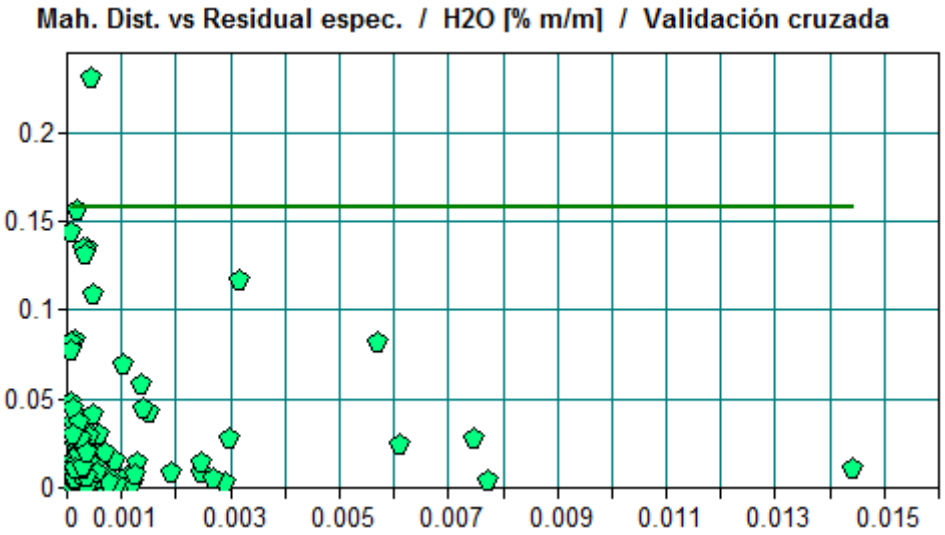


Figure 18. Mahalanobis distance vs Spectral residuals plot for cross validation

Figure 19 is the graphical representation of the validated calibration curve. All the points are homogeneously distributed along the concentration range, having the central part of the range slightly more concentration of points.

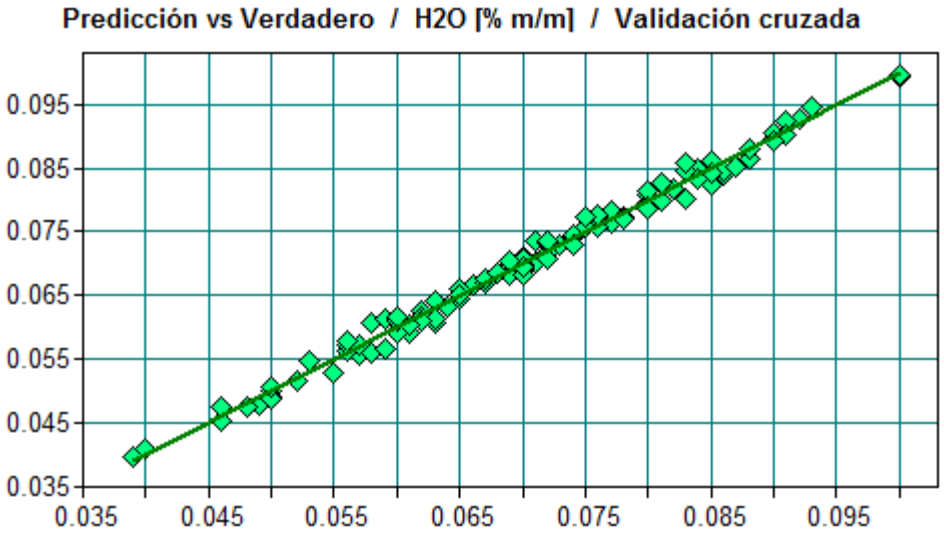


Figure 19. Prediction vs true value plot for cross validation

5.2.2. Test set validation

In the case of the test set validation, the optimised model's optimal rank was also 3. As with cross validation, the rank may not be quite clear either when looking at the maximum and minimum of *Figures 20 and 21*, respectively, because one would think that there might be overfitting when choosing rank 3 instead of rank 2, but the value of RPD for rank 3 (RPD=11.1) is significantly better than the RPD value for rank 2 (RPD=9.89), even more than in the case of cross validation. This makes rank 3 the optimal rank for this validation of the model.

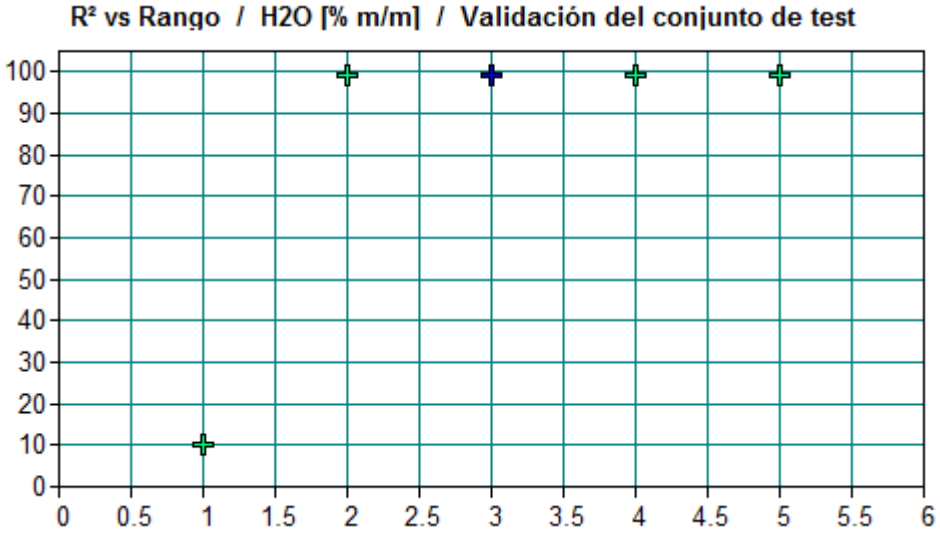


Figure 20. Relative coefficient vs Rank plot for test set validation

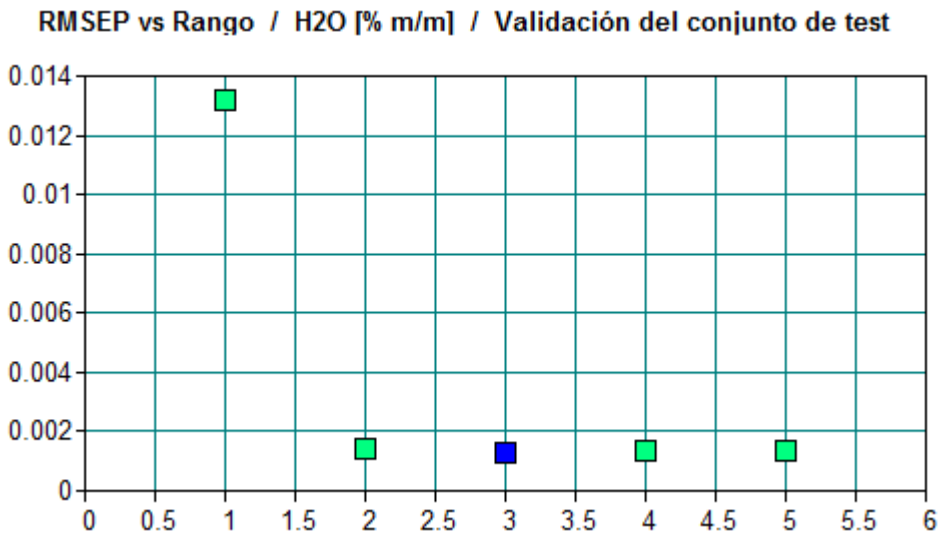


Figure 21. RMSEP vs Rank plot for test set validation

When looking at the Mahalanobis distance in *Figure 22*, it can be seen that one of the samples (Sample 1) has a distance above the limit, which in this case is around 0.30. As it happened before, this means that this sample could be a possible outlier. When examining the sample more closely, its real concentration value is of 0.060 while the predicted value is of 0.059, and it has an F-value and an F-probability of 0.453 and 0.497, respectively. Taking into account all this, it can be considered that Sample 1 is not an outlier, even when its Mahalanobis distance is above the limit.

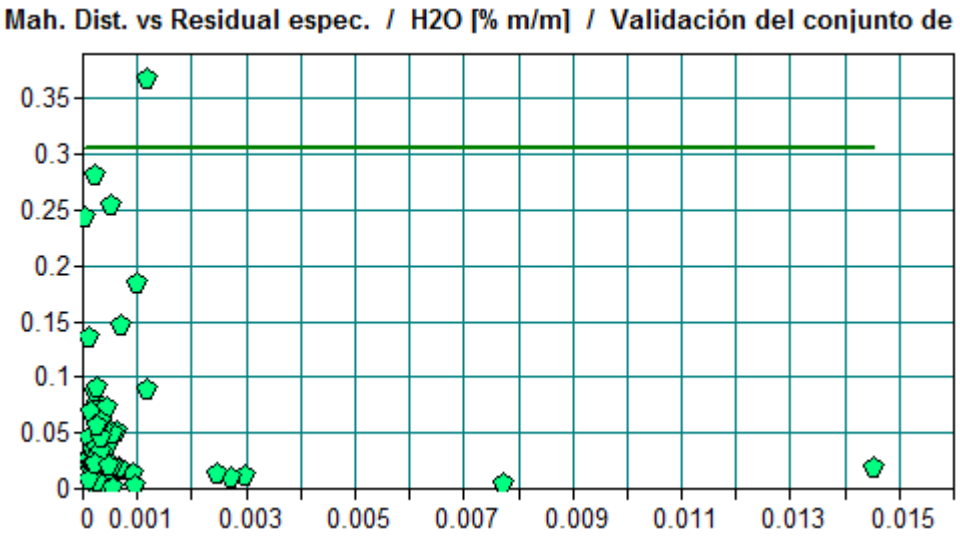


Figure 22. Mahalanobis distance vs Spectral residuals plot for test set validation

Figure 23 is the graphical representation of the validated calibration curve. As in *Figure 19*, all the points are homogeneously distributed along the concentration range, having the central part of the range slightly more concentration of points.

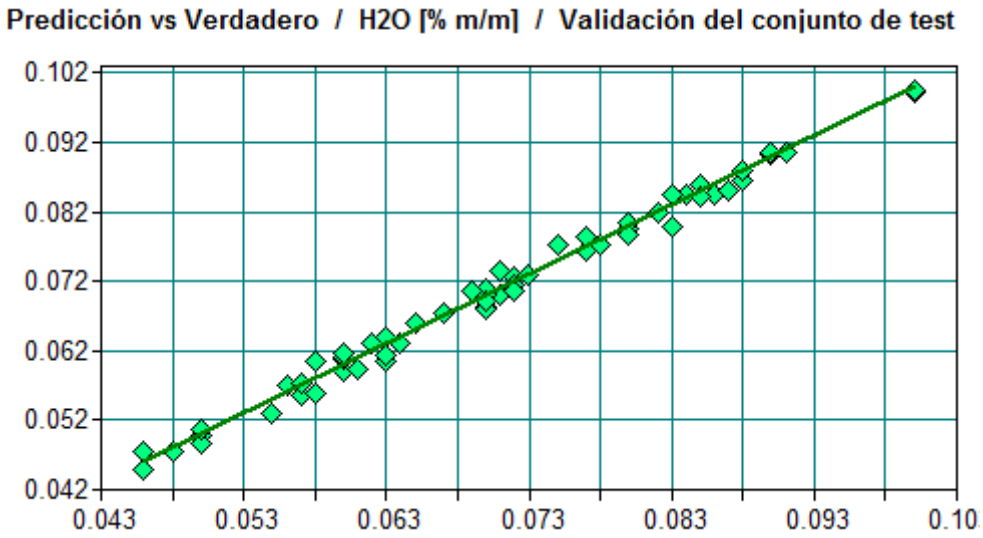


Figure 23. Prediction vs true value plot for test set validation

5.2.3. Comparison of validation methods

The obtained validation results for both types of calibration have been summarised in *Tables 3 and 4* to make the comparison easier to visualise. As it was previously explained in this bachelor's thesis, both validation methods should give comparable results.

Table 3. Summary of the cross validation results of the new model after the optimization

Cross validation results

R^2	RMSECV	RDP	Opt. Rank
99.20	0.00122	11.2	3

Table 4. Summary of the test set validation results of the new model after the optimization

Test set validation results

R^2	RMSEP	RDP	Opt. Rank
99.17	0.00127	11.1	3

As it can be seen in the results obtained in *Tables 3 and 4* above, there is no significant difference between one validation method and the other, meaning the results are comparable, as they should be.

5.3. Testing the model

Once the results of the model optimization were good enough and the comparison of both validation methods was done, an analysis of samples with known concentration was performed to compare the results obtained with the model's prediction to the real concentrations obtained with the Karl Fischer titration.

The TANGO software offers, aside from the option of the building of a model and its optimization and validation (*Quant2*), an option to make an analysis with one of the created models to give prediction values for the selected spectra, called *Quant2 Analysis*.

When doing the validation, two models were created, one for each type of validation, but as their results were really similar and they were built and optimised in the same way (the model was the same one, the only thing that changed was the way it was validated), when doing the *Quant2 Analysis* only one of them was introduced to perform the prediction. In this case, the cross validated model was the one used for predicting.

When comparing the predicted values to the real ones, in general, most of the samples had the same concentration, even if they differed a little from the real concentration values, as it can be seen when looking at the values in *Table 5*.

Table 5. Comparison of predicted and real concentration values

Comparison of concentrations % (m/m)		
Sample	**Predicted**	**Real**
1	0.059	0.060
2	0.062	0.060
3	0.048	0.049
4	0.060	0.058
5	0.052	0.052
6	0.060	0.062
7	0.079	0.080
8	0.063	0.062
9	0.070	0.070
10	0.069	0.069
11	0.093	0.090
12	0.071	0.069
13	0.082	0.082
14	0.090	0.090
15	0.078	0.082
16	0.079	0.080
17	0.075	0.071
18	0.060	0.060
19	0.050	0.050

6. Conclusions

The model that was done with recorded spectra from the last two years was not good. As it can be observed in the results, the points were not homogeneously distributed along the range, which can only mean that the resulting model is not robust and the predictions will not be good. When reducing the range to try and solve this problem, it was clear that it did not help and only made the model worse, making the dispersion of the data points more obvious than before.

As it was explained in the *Results and* discussion section, when commenting this problem with my lab mates, we hypothesised that the possible inaccuracy of the water concentration measurements with Karl Fischer titration may be caused by them not being strict when doing the measurements, as they only needed an approximation of the value and it was good enough for them while the value remained under the maximum. Seeing now the results from the second model built, in which the measurements were done as carefully as possible to reduce the error, I can say that our hypothesis was correct and the problems that surged on the model were caused by the inaccuracy of the Karl Fischer titration measurements. So, this model was completely discarded.

The new model built with the measured samples was much better than the first one. In this second model, the predicted vs real concentration plot was great for both, cross validation and test set validation. The points were homogeneously distributed along the concentration range, none of their predicted concentration values were far from the real ones and there were no outliers in either of the cases.

The results for both, internal and external calibrations, were comparable as there were not significant differences between them, although the cross validation showed a higher relative coefficient and RPD, and a lower error.

The results of the testing of the model also assured that, as the validations showed, the model was good for the prediction of real concentrations. Taking into account that the water concentrations of this product are usually between 0.03 – 0.10 % (m/m), the small differences between the predicted and real values are not significant.

Summarising, the second method build is good for the prediction of unknown samples, as the results of the prediction were very similar if not equal to the results obtained using the official method and both types of validation gave the expected comparable results.

7. Bibliography

(1) ISO - International Organization of Standardization. https://www.iso.org/home.html (accessed Jun 3, 2021).

(2) IQOXE. https://www.iqoxe.com/es/ (accessed Apr 17, 2021).

(3) CL Grupo Industrial. https://grupoindustrialcl.com/ (accessed Apr 17, 2021).

(4) *Alcupol F-3231*. Technical Service and Development: Repsol Technology Centre: Madrid, January 2017.

(5) Bruttel, P.; Schlink, R. *Water Determination by Karl Fischer Titration*; Metrohm Ltd.: Herisau, Switzerland, 2006.

(6) *OPUS: FT-IR Course*, 7.5. Bruker Optik GmbH: Germany, 2014.

(7) Peguero Gutiérrez, A. La Espectroscopia NIR En La Determinación de Propiedades Físicas y Composición Química de Intermedios de Producción y Productos Acabados. Ph. D. Thesis, Universitat Autònoma de Barcelona, May 2010.

(8) *Near Infrared Band Assignment Table*. Bruker Optik GmbH: Germany, 2000.

(9) Conzen, J.-P. *Multivariate Calibration*, 3rd English ed.; Bruker Optik GmbH, 2014.

(10) Brereton, R. G. Partial Least Squares. In *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; John Wiley & Sons, 2003; Vol. 8; pp. 297-313.

(11) Brereton, R. G. Model Validation. In *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; John Wiley & Sons, 2003; Vol. 8; pp. 313-323.

(12) *OPUS*, 7.5; Bruker Optik GmbH: Germany, 2014.

(13) *Alcupol F-3231*; CAS RN: 9082-00-2; REPSOL QUÍMICA, S. A.: Madrid, Spain, May 15, 2015.

(14) *Karl Fischer reagent B pyridine-free*; MSDS No. 109247; Merck KGaA: Darmstadt, Germany, Jul 20, 2018.

(15) *Methanol, >90%*; MSDS No. S25426A; Fisher Science Education: Rochester, NY, Aug 1, 2015.

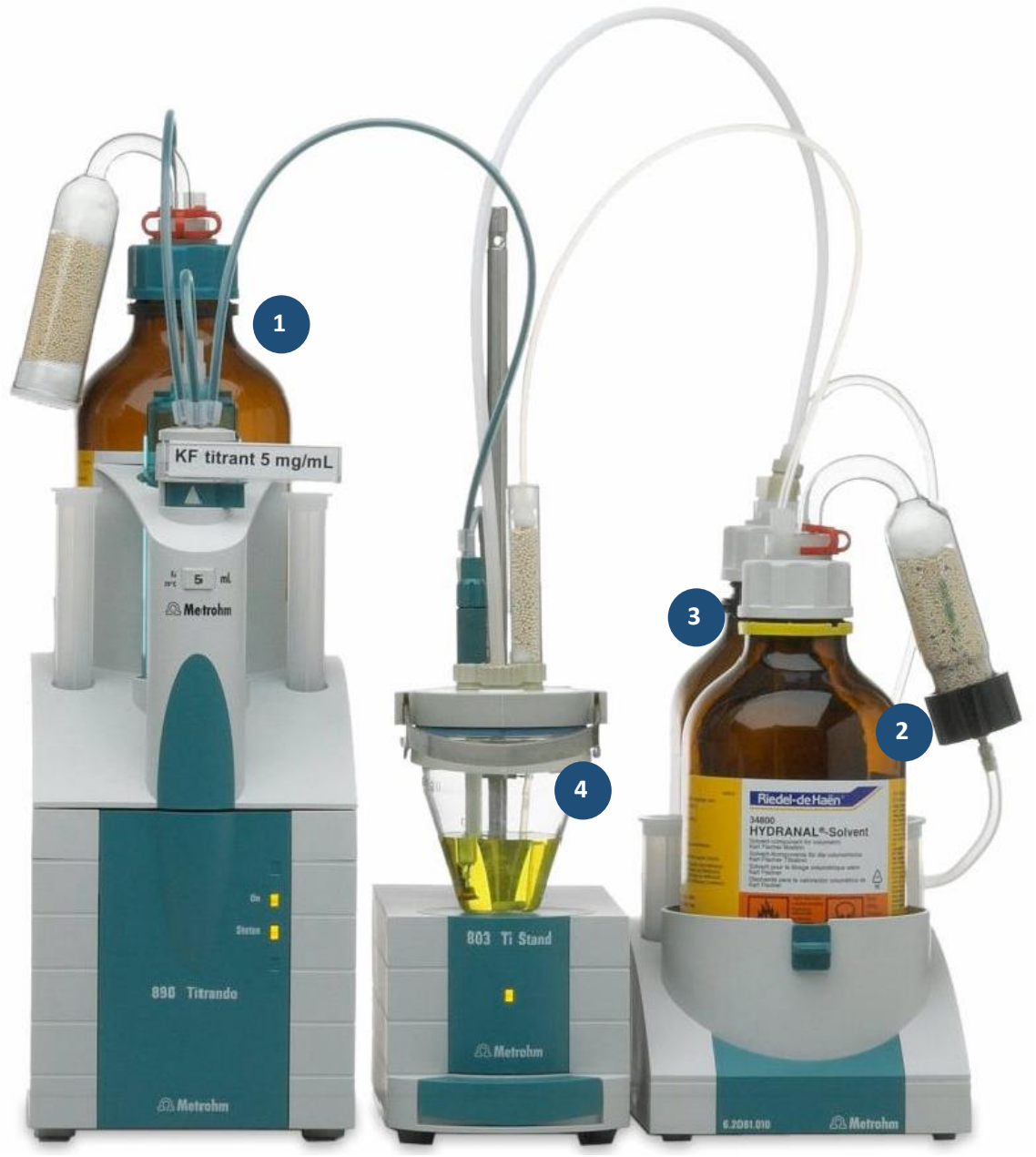
(16) Sánchez-Carnerero Callado, C.; Núñez-Sánchez, N.; Casano, S.; Ferreiro-Vera, C. The Potential of near Infrared Spectroscopy to Estimate the Content of Cannabinoids in Cannabis Sativa L .: A Comparative Study. *Talanta*. **2018,** *190 ,* 147–157.

(17) Suh, E.; Woo, Y.; Kim, H. Determination of Water Content in Skin by Using a FT Near Infrared Spectrometer. *Arch. Pharm. Res*. **2005,** *28* (4)*,* 458–462.

(18) Inarejos-García, A. M.; Gómez-Alonso, S.; Fregapane, G.; Salvador, M. D. Evaluation of Minor Components , Sensory Characteristics and Quality of Virgin Olive Oil by near Infrared (NIR) Spectroscopy. *Food Res. Int.*. **2013,** *50,* 250–258.

8. Annexes

8.1. Karl Fischer titrator



1) Karl Fischer reagent

2) Base bottle

3) Empty bottle for the waste when replacing the cell's contents

4) Titration cell

890 Titrando: Manual. Metrohm's 890 Titrando Manual.

8.2. NIR spectrometer

1 The easy-to-use software interface is operated via a touch screen panel.

2 Sample area for solid and liquid samples. The background is recorded automatically if necessary.

3 Light source and desiccant cartridge can be easily exchanged by the operator.

4 The Internal Validation Unit guarantees that the TANGO is performing at its optimum.

5 The Rocksolid™ interferometer with Cube Corner technology is the heart of the TANGO.

6 Highly reflective gold coated optics ensure optimized light throughput.

TANGO. ANALYSIS TO GO. The next generation FT-NIR spectrometer. Bruker's TANGO Brochure.