Master in Chemical Engineering

Application of PLS models by FTIR for process control of coatings

A Master's dissertation

of

Heloísa Carvalho Reis de Sá

Developed within the course of dissertation

held in

Barbot – Indústria de Tintas, S.A.

BARBOT
Novas emoções

Supervisor at FEUP: **Profª. Margarida Bastos**

Co-Supervisor at FEUP: **Profª. Joana Peres**

Coordinator at Barbot: **Eng. Filipa Oliveira**

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Departamento de Engenharia Química

July of 2019

Acknowledgment

With the completion of this dissertation I want to thank all the people who contributed for the accomplishment of this work

To Professors Margarida Bastos and Joana Peres, academic advisors to this dissertation, I want to express my gratitude for the exemplary guidance they gave me from the beginning to the end of this experience, for their willingness to solve problems, for their constant availability, help, simplicity and understanding demonstrated during these five months.

Engineer Filipa Oliveira, for all the knowledge she shared. And for the accompaniment and help she provided throughout the semester.

To all the people who always have being with me, from the beginning to the end.

Professors Margarida Bastos and Joana Peres, supervisors of this dissertation are integrated members of LEPABE – Laboratory for Process Engineering, Environment, Biotechnology and Energy, finantially supported by: project UID/EQU/00511/2019 - Laboratory for Process Engineering, Environment, Biotechnology and Energy – LEPABE funded by national funds through FCT/MCTES (PIDDAC) and by Project "LEPABE-2-ECO-INNOVATION" – NORTE-01-0145-FEDER-000005, funded by Norte Portugal Regional Operational Programme (NORTE 2020), under PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

Resumo

O trabalho visa realizar análises multivariadas quantitativas em amostras de revestimentos para prever os teores dos quatro principais componentes - polímero, carbonato de cálcio, dióxido de titânio e água - nos produtos acabados com a intenção de utilizar as informações adquiridas para auxiliar no controle do processo.

Para atingir as metas estabelecidas, foram construídas duas bases de dados consoante o tipo de ligante presente nos revestimentos, acrílica e estireno-acrílica. Espectros de revestimentos provenientes da produção diária da empresa e de reformulações feitas em laboratório, foram obtidos por espectroscopia FTIR-ATR e modelos de calibração foram construídos usando o algoritmo dos mínimos quadrados parciais (PLS) combinado com a técnica de validação cruzada k-fold. Para um teste de conjunto de dados independente, foram criados conjuntos de tintas reformuladas - 21 revestimentos de estireno-acrílico e 11 revestimentos de acrílico - com teores específicos, sendo realizada a avaliação de desempenho de todos os modelos.

Por comparação de indicadores estatísticos calculados a partir de ambos os modelos de calibração e conjuntos de dados independentes, oito modelos de previsão foram selecionados como os mais apropriados, um para cada componente principal em revestimentos com os diferentes tipos de ligante.

Por último, foram realizados testes experimentais em conjuntos de revestimentos - 5 tintas estireno-acrílicas e 3 tintas acrílicas - de composição desconhecida para determinar os seus teores de líquidos, compostos orgânicos, carbonato de cálcio e de cinzas. Os melhores modelos foram aplicados, e os teores previstos foram comparados com os resultados obtidos pelas determinações experimentais.

A espectroscopia de FTIR-ATR combinada com a análise PLS de dados provou ser uma ferramenta eficiente na geração de modelos com uma boa capacidade de previsão dos teores dos principais componentes de revestimento. Uma vantagem do método desenvolvido para fins industriais é a rápida quantificação da composição dos produtos fabricados, garantindo a qualidade final do produto. Com base neste método, em casos de erros de produção, através da análise dos produtos não conformes, é também possível identificar e solucionar problemas relacionados com o processo industrial.

Palavras Chave: **Revestimentos, Tintas, Espectroscopia de FTIR-ATR, Regressão dos mínimos quadrados parciais (PLS).**

Abstract

The work aims to perform multivariate quantitative analyses in coatings samples to predict the content of its four main components – polymer, calcium carbonate, titanium dioxide, and water – in the finished products with the intent of using the information acquired to assist in the process control.

To achieve the goals established, two databases were constructed based on the type of binders present in the coatings, namely acrylic, and styrene-acrylic. Spectra from both daily production and manually modified coatings were attained by FTIR-ATR spectroscopy, and calibration models were built using partial least squares (PLS) algorithm combined with k-fold cross-validation technique. For an independent dataset test, sets of reformulated paints – 21 styrene-acrylic's and 11 acrylic's coatings - were created with specific contents and performance evaluation of all models was carried out.

By comparison of statistical indicators computed from both calibration models and independent dataset data, eight prediction models were selected as the most appropriated, one for each main component for coatings with the different kinds of binders.

Finally, experimental tests were performed in sets of coatings – five styrene-acrylics and three acrylics paints - of unknown composition to determine their liquid, organic compounds, calcium carbonate, and ashes' contents. The best models were then successfully applied, and predictions' content values were compared to the results gathered by the experimental determination.

FTIR-ATR spectroscopy combined with PLS data analysis was proved to be an efficient tool in generating models with overall good predictability of the coating's main components contents. The advantages of the method developed for industrial purpose are the performance of rapid quantification of the composition of manufactured goods providing product quality assurance, and in cases of manufacture errors, the essential information provided by analysis of defective good which can assist in process control troubleshooting.

Keywords: **Coatings, Paints, FTIR-ATR spectroscopy, Partial Least Square (PLS).**

Declaration

I hereby declare, on my word of honor, that this work is original and that all non-original contributions were properly referenced with source identification.

Porto, first of July 2019

Index

List of Figures

List of Tables

Notation and Glossary

a_0	Average value of the sample spectrum to be corrected	
a_1	Standard deviation of the sample's spectrum	
A	Number of components in a PC or PLS model	
$\mathbf{b}_k$	Regression coefficient vector of the k^{th} y variable	%, w/w
$\mathbf{B}$	Regression coefficient matrix	
$\boldsymbol{c}$	Concentration	mol L^{-1}
$\boldsymbol{E}$	X-residuals matrix	
$\boldsymbol{F}$	Y-residuals matrix	
$\mathbf{g}_k$	residuals of the kth y variable	%, w/w
$\mathbf{G}$	PLSR Y-residuals matrix	
h	Leverage	
$\boldsymbol{l}$	Pathlength of IR beam	cm^{-1}
n	Number of standards	
$\mathbf{p}_\text{a}$	X loading vector of component a	
$\mathbf{P}$	Loading matrix; $\mathbf{p}_\text{a}$ are columns of $\mathbf{P}$	
R^2	Coefficient of determination	
$\mathbf{s}_\text{a}$	Y loading vector of component a (PLSR Y-weights of component a)	
$\mathbf{S}$	Loading matrix; $\mathbf{s}_\text{a}$ are columns of $\mathbf{S}$	
$\mathbf{t}_\text{a}$	Latent variable a and also X-scores of component a	
$\mathbf{T}$	Score matrix; $\mathbf{t}_\text{a}$ are columns of $\mathbf{T}$	
$\mathbf{u}_\text{a}$	Y-scores of component a	
$\mathbf{U}$	Score matrix; $\mathbf{u}_\text{a}$ are columns of $\mathbf{U}$	
x	Absorbance measured in the different wavelengths of the spectrum	
x_{org}	Original sample spectrum attained by the FTIR analysis	
x_{ref}	Reference spectrum	
x_{cor}	Corrected spectrum	
$\mathbf{x}_\text{j}$	X-variable j	
$\mathbf{X}$	Matrix of predictor variables	
$\mathbf{w}_a$	PLSR X-weights of component a	
W	Mass	g
y	Experimental value of the property	%, w/w
$\hat{y}$	Estimated value of the property	%, w/w
$\bar{y}$	Mean of experimental data	%, w/w
$\boldsymbol{Y}$	Matrix of response variables	
$\hat{\mathbf{Y}}$	Response predictions matrix	

Greek Letters

ε	molar absorptivity	L mol^{-1} cm^{-1}

Indexes

a	index of components ($a = 1, \dots, A$)
i	index of observations ($i = 1, \dots, I$)
k	index of Y-variables ($k = 1, \dots, K$)
j	index of X-variables ($j = 1, \dots, J$)

List of Acronyms

ATR	Attenuated Total Reflection
A/D	Analog-to-Digital
CV	Cross-Validation
EMSC	Extended Multiplicative Scatter Correction
FTIR	Fourier Transform Infrared spectroscopy
GC-MS	Gas Chromatography with Mass Spectrometry
IR	Infrared
IRE	Internal Reflection Element
LV	Latent Variables
MLR	Multilinear Regression
MSC	Multiplicative Scatter Correction
NSE	Nash-Sutcliffe Efficiency
PBIAS	Percent Bias
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least-Square
RMSE	Root Mean Square Error
RSR	Ratio of the Root Mean Square Error to the Standard Deviation of measured data
SEE	Standard Error of Estimate
SEP	Standard Error of Prediction
SNV	Standard Normal Variate
STDEV	Standard Deviation

1 Introduction

1.1 Framing and presentation of the work

Coatings are a mixture of volatile and nonvolatile components. Solvents, binders, extenders, and pigments are responsible for most of the final product's volume. Additives, like anti-foam and neutralizing agents, biocides, thickeners, are accountable for a minimum quantity of its final weight, less than 5 % (w/w) (Bieleman, 2000).

Precise determination of coatings composition, although very challenging, is of extreme importance when it comes to maintaining process control throughout manufacture. Information acquired from the composition analyses of a flawed final product can be of extreme help in process control troubleshooting.

The Fourier transform infrared (FTIR) spectroscopy has become an analytical technique extensively implemented in several different industries (e.g., polymer, pharmaceutical, food). It is used with the purpose of identifying quality consistency of raw materials, verifying finished goods quality, resolving contamination issues and assisting in the process control (Bunaciu, Aboul-Enein, and Fleschin, 2010; Amir et al., 2013).

The ability to perform a short and non-destructive sample analysis that requires no previous sample preparation (Chalmers and Everall, 1999) and provide crucial data about components present in complex mixtures is of invaluable importance for the coating industry.

Held in a business environment, the primary objective of the thesis was the development of chemometric methods for quantitative analysis of the main components present in coatings. The underlying motivation for the study was the necessity of conducting a fast and effective evaluation of the composition of paints in order to secure its validity and provide information about the occurrence of possible manufacturing errors.

This work has as a finality the development of multivariate models capable of predicting paint composition in order to maintain the process control and correct possible manufacture errors, which can infer inferior quality to the final product.

To achieve such goals, the creation of spectra databases was required. Infrared spectra were obtained from liquid samples acquired by Fourier transform infrared (FTIR) spectroscopy equipment incorporated with an attenuated total reflection (ATR) accessory. The data were collected and interpreted by the PerkinElmer Spectrum™ 10 software.

The PerkinElmer Spectrum® Quant software was used for the development of quantitative multivariate analysis methods, with the implementation of partial least-square (PLS) algorithm.

Prediction models for polymer, water, titanium dioxide, and calcium carbonate content were created individually using the assembled databases.

1.2 Presentation of the company

In 1920, Barbot was founded by Diogo Barbot in Porto. The first factory was situated in the Santo Ildefonso. The family business had a traditional image with fundamental characteristics such as rigor, safety, and quality.

In 1958, under Carlos Aires Pereira, a new factory was built in Laborim. Four years later, it was created the brand's most famous paint and current best seller, Barbot Dioplaste.

A violent fire destroyed the entire Santo Ildefonso factory in 1981, indisputably changing the history of the brand. Now led by Carlos Barbot, the company focused on expanding throughout policies of acquisitions and partnerships, which allowed to strengthen the label and extend its range of products. In 1982, occurred the transfer of its premises from Porto to Vila Nova de Gaia.

Throughout the 1990s, Barbot continued to expand, continually launching new products. This expansion led to a fourfold increase in the warehouse area and the opening of a new plant in Canelas in 2002. A higher level of responsiveness and competitiveness was able to be achieved by the acquisition of modern facilities equipped with the most advanced technology, providing Barbot with a field of national scope.

Barbot is currently present from the north to the south of Portugal, with two factories, an extended network of resellers and twenty-four of its own points of sale.

The Barbot group includes eight companies, clearly demonstrating the importance of this continuously growing brand that already operates on three continents: Europe (Belgium, Spain, France, Luxemburg, Switzerland), Africa (Angola, Cabo Verde, Guinea, Mozambique), and South América (Peru).

1.3 Contributions of the Work

The work developed generated two complete spectra databases, which include all daily productions prepared at Barbot, for styrene-acrylic and acrylic coatings. With the use of the databases, multivariate models for identification of the main components – polymer, calcium carbonate, titanium dioxide, and water – for the two types of coatings, were built.

With the implementation of the PLS models faster and effective composition analysis of finished products can be performed, being a valuable tool for process control troubleshooting.

1.4 Organization of the thesis

The thesis consists of 6 chapters. Chapter 1 presents a brief description of the work, stating its goals, as well as the company where it was developed. In chapter 2, the most relevant theoretical information towards the understating of the work, for example, its background, paint systems, and the main methods used, FTIR and multivariate analysis, are addressed. In chapter 3, the materials and the previously explained method's application are described in detailed, for possible work reproducibility. In chapter 4, the results attained are presented as well as discussed, and the conclusions drawn are then indicated in chapter 5. In the last chapter, 6, an assessment of the work performed is made.

2 Context and State of the art

Nowadays, product quality is one of the most critical concerns in industrial activity. Modern production is no longer possible without the search for continuous improvement of the final product's qualitative aspects (Jozsef and Blaga, 2014).

Characterization of paint mixtures has been successfully carried out by employing advanced analytical techniques such as gas chromatography with mass spectrometry (GC-MS) (Colombini *et al.*, 2000) implemented with chemometric methods like principal components analysis (PCA) (Colombini *et al.*, 2010). However, disadvantages related to the use of this method, as for requiring sample preparation and thus time, make it less appropriate for industrial process control. FTIR spectrometry, on the other hand, because of its reliability, little sample preparation, and exceptionally rapid analysis, is widely used for qualitative identification, using PCA (Muehlethaler, Massonnet and Esseiva, 2011; Sciutto *et al.*, 2017) and quantitative determination performing PLS, in some paint components, like pigments (Marengo *et al.*, 2005), and also in oils (Schulz *et al.*, 2002). Nevertheless, all studies previously mentioned are regarding components for artistic purpose, for example paintings rehabilitations, and are not related to industrial activities.

A recent study, conducted with liquid and solid samples, was successfully able to quantify the components present in mixtures of paint consisting of several different binding media, pigments and alkyd resin. FTIR-ATR analysis was carried out followed by the creation of PLS models, and data analysis (Hayes, Vahur and Leito, 2014). The study achieved positive results for coatings containing alkyd resin and pointed out the advantages of applying the method to process and quality control in manufacturing of coatings.

2.1 Composition of coatings

According to the ISO 4618/1; DIN 55945, “Coating” is a general term denoting a material that is applied to a surface. “Paint” indicates a pigmented material, while “varnish” refers to clear lacquer (Stoye and Freitag, 1998). Coatings are defined as any liquid, paste, or powder products designed for application to a substrate, by various methods and equipment, in a thin layer which is converted to an opaque solid film after application (Gürses *et al.*, 2016). Their primary purpose is either to protect surfaces or to create a more aesthetically pleasing environment, however, they can also provide information (traffic signs, information signs, advertising); or have more specific properties.

Coatings systems are made of numerous components, depending on the method of application, the desired properties, the substrate to be coated, and ecological and economic constraints (Stoye, Marwald and Plehn, 2010).

Coatings constituents can be classified as volatile (solvents) or nonvolatile (binders, additives, pigments, and extenders). Solvents, binders, extenders, and pigments account for most of the final product while the proportion of additives are meager, typically, less than 5 % (w/w) (Bieleman, 2000). These components fulfill specifics functions not only in the liquid paint but also in the solid coating film, obtained after hardening. Their leading roles will be briefly described in the subsections below.

2.1.1 Binders

Binders are macromolecular products, which ought to be suitable for physical film formation. Their primary function is to ensure the cohesion of the paint and the connection between the pigments (Alua, 2012). Resin, binder formulation, can be formed by polymers of different molecular sizes. High molecular mass polymers, as cellulose nitrate and polyacrylate and vinyl chloride copolymers, are able to initiate film formation without the need for an additional chemical reaction. However, low molecular mass products which include alkyd resins, phenolic resins, polyisocyanates, and epoxy resins, must be chemically hardened after application to produce high molecular mass cross-linked macromolecules, in order to produce acceptable films.

Binders are of extreme importance once they affect almost all properties of the coating, including adhesion and related properties (resistance to blistering, cracking and peeling), fundamental resistance properties (resistance to scrubbing, chalking and fading), and application properties (flow, leveling and film build, and gloss development).

2.1.2 Pigments and Extenders

Pigments and extenders are responsible for the color and covering power of coatings; in some cases, they also grant improved properties to the coating film. They can be combined in different proportions in order to achieve a specific type of paint.

Pigments are colored, colorless, or fluorescent particulate organic or inorganic finely divided solids which are usually insoluble in, and virtually chemically unaffected by, the vehicle or medium in which they are incorporated (Gürses et al., 2016). They provide one or more of the principal functions, namely color, opacification, and anti-corrosive properties (Stoye and Freitag, 1998). The most important pigment used in the industry is titanium dioxide (TiO_2).

Extenders disperse well in coatings, although they provide much less hiding power when compared to pigments. Their main advantage and the reason why they are used in significant quantities are related to their relatively low cost, which makes the final product more cost-effective. They also impact on many properties, enhancing opacity, controlling surface sheen, and facilitating ease of sanding. Typical extenders are barium sulfate, calcium carbonate, talc, or kaolin.

2.1.3 Solvents

The solvents should be able to disperse the binder, conceding the proper viscosity to the paint and obtaining a clear and homogeneous solution as the final product. Besides, they enable incompatibility between paint components to be overcome, improve pigment wetting and dispersion, and control storage stability of the coating (Stoye, Marwald, and Plehn, 2010).

In coatings, some of the most used organic solvents are aromatic and aliphatic hydrocarbons, esters of acetic acid, glycol ethers, alcohols, and some ketones (Stoye and Freitag, 1998). In water-based systems, the water may act not only as a true solvent for some components but also be a non-solvent for the main film former. This occurs in decorative emulsion paints.

2.1.4 Additives

Paint additives are auxiliary raw-materials added to coatings, in much lower amounts, in order to improve particular technical properties, facilitate manufacture and application, increase stability, and minimize defects of the paints or coating films (Alua, 2012; CSD Engineers, 2016). There are several additives available these being classified according to their mode of action. In coating systems, the most commonly used are thickeners and rheology modifiers, surfactants, biocides, fungicides, defoamers, anti-skinning agents, film-formation promoters, catalysts, neutralizing agents, corrosion inhibitors.

2.2 Fourier-Transform Infrared (FTIR) Spectroscopy: Attenuated Total Reflectance (ATR)

Fourier-Transform Infrared spectroscopy is an analytical technique widely used to provide information about the chemical composition of matters. Also, it possesses leverages such as being a rapid, nondestructive, and time-saving method. There are four main sampling techniques: transmission/absorption, attenuated total reflection (ATR), diffuse reflection (DRIFTS), and specular reflection (Thermo Scientific, 2013b).

The Attenuated Total Reflectance (ATR) technique is used to provide a qualitative and quantitative analysis of solids, gels, pastes, and liquids samples. Its main advantages when compared with the other available methods are requiring minimal sampling preparation, having excellent sensitivity and low detection limits (Thermo Scientific, 2013b, 2013a).

The ATR-FTIR spectrometer contains a polychromatic infrared source (wavelengths range of 10,000 to 100 cm^{-1}), an adapted Michelson interferometer, a sample compartment, an internal reflection element (IRE) (made of Diamond, ZnSe, Ge, or Si), a detector, an amplifier, an analog-to-digital (A/D) converter and a computer.

A beam of light is generated at the source. Afterward, it goes through the Michelson interferometer where it is divided into two by the beam splitter which is designed to transmit half of the light and reflect the other half. One of the recently formed beams reaches a stationary mirror while the other strikes a moving mirror. Both are then reflected and return to the beam splitter where they recombine (LibreTexts, 2019).

Finally, the beam is directed onto the optically dense crystal with a high refractive index at a certain angle. This internal reflectance creates an evanescent wave which extends beyond the surface of the crystal into the sample (penetration length of 0.5 μ-5 μ) held in contact with the crystal (PerkinElmer, 2005). Some of the energy of the evanescent wave is absorbed by the sample, and reflected radiation reaches the detector (Figure 1).

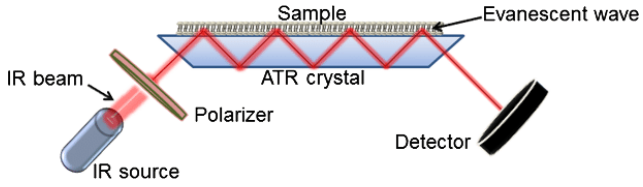

Figure 1. Schematic diagram of the ATR accessory (Adapted from: Ausili, Sanchez and Gomez-Fernandez, 2015).

There the difference in the intensity of the previous two beams is measured as a function of the difference of their paths. The signal is amplified and converted to a digital signal by the amplifier and analog-to-digital converter, respectively. Eventually, the signal is transferred to a computer in which Fourier transform is carried out (LibreTexts, 2019).

The resulting spectrum is more commonly obtained from 4000 cm^{-1} to 400 cm^{-1}, and it represents a molecular fingerprint of the sample. Since every chemical structure produces a unique spectrum fingerprint, this analysis became much appreciated as a tool for chemical identification. In Annex 1 (Table AA1 – AA4) is presented the most relevant information for the FTIR spectra interpretation of coatings. Also, in Appendix 1 (Figure A1.1) can be observed an example of the resulting spectrum obtained from styrene-acrylic coating's analysis.

For quantitative analysis, the Beer–Lambert law (Equation 1) is used. It relates the absorbance to the thickness and the concentration of the sample, since they are directly proportional.

$$A = \varepsilon\, c\, l \tag{1}$$

Where A is the absorbance, and c the concentration of the sample, l the pathlength of the IR beams, and ε is the constant of proportionality, referred to as the molar absorptivity.

In order to analyze the composition of a sample of unknown concentration, it would be necessary to use samples where the concentrations are known, their absorbances then be measured followed by the plot of a calibration graph. The concentration of a specific compound

or even compounds present in the sample, could then be read from the calibration graph, given that their absorbances are known (Stuart, 2004).

2.3 Data pre-processing methods

Pre-processing is of extreme importance for spectroscopy data analysis, since, when raw, the spectra may contain noise, scattering, baseline variations, or other perturbations which can shadow meaningful information (Sjo, 2018).

Methods for signal correction are used to improve the reading of significant data or to minimize perturbations associated with the equipment (Roussel et al., 2014). Some of the preprocessing techniques more commonly applied in spectroscopy are baseline correction, normalization, multiplicative scatter correction (MSC), extended MSC (EMSC), standard normal variate (SNV) and range (Rinnan, Berg and Engelsen, 2009).

Baseline correction

Baseline correction is responsible for compensating amplitudes shifts, by subtracting the undesired spectrum background, formed throughout the different wavelengths which are induced by spectrum signal and lead to wrong results. The traditional way is to subtract the lowest value of each spectrum from all the variables (Roussel et al., 2014; Qian, Wu, and Hao, 2017).

Normalization

Normalization can be done by dividing the spectra by an estimation of its spectrum intensity. It can be based on properties like its area, maximal peak, length, a specific spectrum point, or the sum of the spectrum values (Roussel et al., 2014). After normalization is applied, the data is presented in a patterned matter, for example, having a specific size, facilitating comparisons (Sjo, 2018).

Multiplicative Scatter Correction (MSC)

The MSC removes additive and multiplicative influences from interfering signals, eliminates offsets and baseline effects, and normalizes sets of data (Sjo, 2018). Its concept is to reduce these undesirable effects by fitting each spectrum to a reference spectrum, usually an average calibration database spectrum (Roussel et al., 2014). It comprises two phases. First, it estimates the correction coefficients by performing a linear regression as in Equation 2

$$x_{org} = b_0 + b_1 \, x_{ref} + e \tag{2}$$

where x_{org} is an original sample spectrum attained by the FTIR analysis, x_{ref} is a reference spectrum used to perform this treatment in the entire dataset, and e is the un-modeled. b_0 and

b_1 are the correction coefficients computed for each sample. Second, a correction of the original spectrum is made using these coefficients, as shows Equation 3

$$x_{cor} = \frac{x_{org} - b_0}{b_1} = x_{ref} + \frac{e}{b_1} \tag{3}$$

where x_{cor} is the corrected spectrum (Rinnan, Berg and Engelsen, 2009).

Extended MSC (EMSC)

The extended version of MSC (EMSC) has higher flexibility once it performs more selective correction for distinct types of perturbations which cannot be adjusted with the use of other conventional preprocessing techniques. It is based on a second-order polynomial fitting to the reference spectrum depending on the wavelength, as in Equation 4

$$x_{org} = b_0 + b_1\, x_{ref} + b_2\, \bar{v} + b_3\, \bar{v}^2 + e \tag{4}$$

where $\bar{v}$ is a specific wavenumber. Afterward, the corrected spectrum can be produced in a simplified matter by the use of Equation 5.

$$x_{cor} = \frac{x_{org} - b_0 - b_2\, \bar{v} + b_3\, \bar{v}^2}{b_1} \tag{5}$$

This expansion also allows the use of a priori knowledge from the spectrum of interest or spectrum interferents (Rinnan, Berg and Engelsen, 2009; Sjo, 2018).

Standard Normal Variate (SNV)

SNV correction has similar concepts to normalization and MSC, as it also is a path-length variation correction method. Its goal is to limit the spectrum intensity variation problem by processing each spectrum individually, but unlike MSC this step does not require a common reference signal, as can be observed in equation 6

$$x_{cor} = \frac{x_{org} - a_0}{a_1} \tag{6}$$

where, a_0 and a_1 are the average value of the sample spectrum to be corrected and the standard deviation of the sample's spectrum, respectively.

Range

The range treatment is used when a specific region of the spectrum is considered to have more relevance or the opposite, when a region presents interference or does not have significant meaning to analyses, and therefore can be removed from consideration. The Range pre-processing step enables to select an interval of the data to be included or excluded from the method calculations, and the data region to be used for all standard and sample spectrum (PerkinElmer, 2014).

2.4 Quantitative predictive models

2.4.1 Multivariable analyses

Multivariate analysis is a collection of statistical methods that copes with a large number of variables. These variables may present correlations between one another that can be taken into consideration when such techniques are applied (Olkin and Sampson, 2001; Beukelman and Brunner, 2016).

In order to transform sets of data into useful statistical information, a relatively new discipline, named chemometric, is implemented. It applies mathematical, statistical, and other logic-based methods to analyze chemical data, in particular, in analytical chemistry (Sánchez Rojas and Cano Pavón, 2005). Its goal is to develop models which correlate the information in the set of known measurements to the desired property (Camo Analytics, 2019). The most common methods used in multivariate analysis are principal component analysis (PCA), multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) regression.

Principle component analysis (PCA)

The method's main intention is to characterize a system of a large number of interrelated variables with the minimum number of variables possible which comprises the most significant variation present on it thus, reducing the dimensionality of the data set (Jolliffe, 2002). The original variables are transformed to new variables, named principal components (PCs) that are uncorrelated and which are obtained by descending order of explained variance present in the variables. Thus, the first PC retains the most significant variance present, the second PC the second largest variance, and this descending order of the variance magnitude prevails until the last PC. Thus, this method transforms the original coordinate system axes in a new orthogonal coordinate system where the axes are defined by the directions of the PCs. Having in mind the dimensionality reduction, only a few PCs are considered, and the data set matrix X is transformed by Equation 7

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{T} + \mathbf{E} \tag{7}$$

where $\mathbf{T}$ is the scores matrix, i.e., the projection of the data points coordinates onto the PCs axes, and $\mathbf{P}$ is the loadings matrix, i.e., the coefficients that define the directions of each PC in the space of the original variables. The error matrix $\mathbf{E}$ comprises the PCs that retain insignificant variance in the data, so it can be neglected.

Score plots (score as a function of principal components) allow a better conception of the sample variability while maintaining the distances between samples and the plot scale (Roussel

et al., 2014) and can indicate how are the relationship among distinct samples, while principal component plots show how different variables are related to each other (Peter *et al.*, 2018).

The PCA method is usually used to help visualize samples represented by several variables, detect outliers, select best-representing variables, and compress data by removing noise (Roussel *et al.*, 2014). This method was the basis for other multivariate methods, such as previously mentioned.

2.4.2 Partial Least Square (PLS) Regression

PLS regression is of particular interest as it can analyze data with strongly correlated, noisy, and numerous X-variables as in the case of spectra data. It is also able to simultaneously model several response variables, Y (Wold *et al.*, 2001). Whereas PCA maximizes only the variance in X, PLS tries to maximize covariance between X and Y (Grimnes and Martinsen, 2015). The PLS model (Figure 2) decomposes X and Y data in scores and loadings through Equations 8 and 9

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^T + \mathbf{E} \tag{8}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{S}^T + \mathbf{F} \tag{9}$$

where matrix $\mathbf{U}$ and $\mathbf{S}$ are the scores and loadings of Y, respectively, and $\mathbf{F}$, the error matrix.

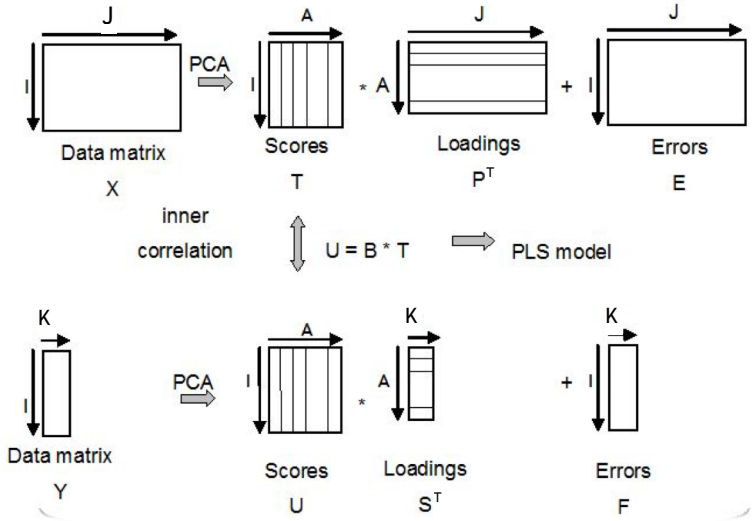

Figure 2. Decomposition of X and Y into matrices of scores and loadings. The I observations described by K dependent variables are stored in the matrix Y, the values of J predictors collected on these I observations are collected the matrix X. E and F are the errors matrices (Adapted from Bohm, Smidt and Tintner, 2013).

The PLS algorithm searches for a set of components, named latent vectors, that performs a simultaneous decomposition of X and Y, i.e., $\mathbf{T}$ and $\mathbf{U}$ are determined together, with the constraint that these components must explain as much as possible of the covariance between X and Y (Abdi, 2003). The PLS model calculates a few new variables (latent variables) denoted

by $\mathbf{t}_a$ $(a = 1, \cdots, A)$ which can be expressed as a linear combination of the $\mathbf{x}_j$ variables (Equation 11) with the coefficients w_{ja} (Equation 10), where A is smaller than J (Kunal Roy, 2015).

$$t_{ia} = \sum_j w_{ja}\, x_{ij} \tag{10}$$

The X scores ($\mathbf{t}_a$) are orthogonal then it can be written that

$$x_{ij} = \sum_a t_{ia}\, p_{aj} + e_{ij} \tag{11}$$

where p_{aj} represents the loadings and e_{ij} the X-residuals. Similarly, the Y variables can be expressed as Y scores ($\mathbf{u}_a$) such that

$$y_{ik} = \sum_a u_{ia}\, s_{ak} + f_{ik} \tag{12}$$

where s_{ak} represents the loadings so that the residuals, f_{ik}, are "small".

The X-scores are good predictors of Y so that it can be written that

$$y_{ik} = \sum_a s_{ka} \sum_j w_{ja}\, x_{ij} + g_{ik} = \sum_j b_{kj}\, x_{ij} + g_{ik} \tag{13}$$

where the Y-residuals (g_{ik}), express the difference between the observed and modeled responses. The b_{kj} are the PLS regression coefficients obtained from the loadings, s_{ka}, and weights, w_{ja}, and can be written as

$$b_{kj} = \sum_a s_{ka}\, w_{ja} \tag{14}$$

The prediction of a PLS method is summarized in the regression coefficient matrix, $\mathbf{B}$, as shown in Equation 15

$$\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{B} \tag{15}$$

in which $\hat{\mathbf{Y}}$ are the matrix of the response predictions.

When the w_{ja} weights present a substantial value, the X-variables are considered to provide relevant information to the model, while if the weight values are similar, the X-variables are considered to provide the same information. Therefore, this coefficient allows a crucial understanding of how important the X-variables are (Kunal Roy, 2015). Significant b_{kj} regression coefficients also allow identifying valuable X-variables. Large Y-residuals suggest that the model has bad prediction abilities, while X-residuals are used to identify outliers in the X-space.

In PLS modeling, it is considered that only a few variables (latent variables, LV) influence the system. At first, the exact number of LV's are unknown, with the performance of several iterations, the PLS regression, is capable of estimating the ideal LV number (Wold *et al.*, 2001).

2.4.3 Cross-Validation

Cross-validation (CV) is a practical and consistent way to analyze the predictive capabilities of a model. This can be of extreme importance since, when dealing with a high number of correlated X-variables, the resulting model can be over-fitting, presenting very little or no estimation power (Wold *et al.*, 2001). There are two types of cross-validation strategies: leave one out and k-Fold.

Leave one out

In the leave one out, each standard is taken out of the data set, one at a time. Calibration is performed, and the standard eliminated is then used to validate the model (Roussel *et al.*, 2014).

K-Fold

This approach divides the samples into k groups (Figure 3), developing parallel models, excluding one of the groups at a time. This deleted group is later used to validate the model.

The sequence separation consists of sequentially allocating the samples into the different groups. For example, if six samples were divided into two groups, A and B, the sequence division would result in the odd samples going to the group A and the even samples to group B.

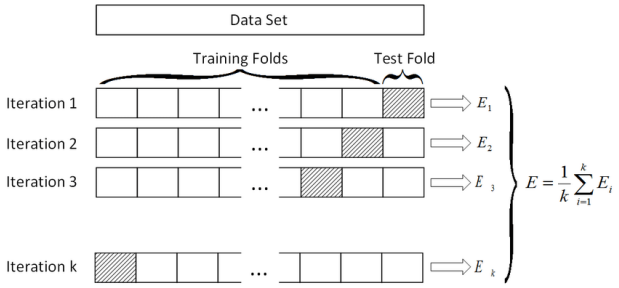

Figure 3. k-Fold cross-validation scheme (adapted from Oliveira *et al.*, 2017).

There are three different ways of dividing the standards into groups, by sequence, blocks, or random partition.

The block approach separates samples into groups. Using the same example as before, the result would be, the first three samples going to group A while the three lasts would go to group B.

The last tactic of splitting samples commonly used, as the name suggests, arbitrarily distributes the standards into the k groups (PerkinElmer, 2014).

After developing models, the residual variance between the predicted and actual Y-values are calculated for all deleted data in each model, as is the sum of squares of these differences. The later parameter measures how well-fitted a model is to a set of new standards, not used for its development. With these calculated values, the standard error of prediction (SEP) estimate can be computed, as shown in Equation 16.

$$\mathrm{SEP_j}^{estimate} = \sqrt{\frac{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1 - h_{ii})^2}}{n - 1}} \tag{16}$$

where j is the number of LV considered in the model, y_i and $\hat{y}_i$ are the experimental and estimated value of the property, respectively, and n the number of standards. The h_{ii} is the leverage of the i th sample in the property subspace (PerkinElmer, 2014).

The $\mathrm{SEP_j}^{estimate}$ value establishes how many latent variables will be included in the model. The $\mathrm{SEP_j}^{estimate}$ value associated with the lowest number of latent variables, considered not being significantly different from the $\mathrm{SEP_j}^{estimate}$ of the maximum latent variable number will be the chosen value. This can be explained by the principle of the PLS algorithm works, which is, as stated before, to explain a complex system as best as possible by using a minimum of latent variables, considered to be the optimum number of latent variables to the model. An F-test is performed in order to determine if there is a significant change between the statistical indicator values ($\mathrm{SEP_j}^{estimate}$) (PerkinElmer, 2014).

3 Materials and Methods

3.1 Samples origins

In order to achieve well-adjusted models, it was necessary to build databases with coatings spectra's information, which contained a proper extent of the desired components' contents. In other words, these contents needed to vary as much as possible, so they would cover a significant range of values and therefore be able to produce reasonable predictions results.

The use of samples acquired only from daily production was thought not to be enough to obtain the models stated before. With that in mind, sampling modifications were conducted, adjusting the desired components' contents. In Table 1, it is shown the number of samples gathered of each type of coating, what kind of resin they possessed, and their origin.

Table 1. Summary of obtained samples

		Daily productions			Laboratory Modifications	
Coating's type		Paint	Varnish	Primer	Paint	**Total**
Resin	Acrylic	35	1	8	15	59
	Styrene-acrylic	89	1	3	35	128

Daily Production

The samples were derived from daily production of coatings – paints, varnishes or primers – which were formulated with either acrylic or styrene-acrylic resins.

Laboratory modifications

Alterations were proceeded by using some of the samples mentioned above to formulate newly remodeled paints. This was done by the addition of previously calculated quantities of water, calcium carbonate, and titanium dioxide dispersions, and resin. These contents' values were chosen based on their expected scope on coatings, and so gaps were filled. This strategy allowed a better representation of the standards' resulting spectra according to changes in the sought components' concentration leading to the production of more complete and consequently reliable models.

3.2 Spectra FTIR analysis

Two separate databases, one for each type of resin, were created with the spectra collected from analysis done in the Spectrum Two FT-IR Spectrometer by PerkinElmer equipped with a diamond crystal ATR accessory, optical potassium bromide beam splitter, and a lithium tantalite ($LiTa0_3$) detector. The liquid samples, previously properly blended, were placed in

small amounts on the ATR accessory, assuring complete coverage of the diamond crystal. In every sample, measurements were performed in triplicate where eight scans were accumulated at a resolution of 2 cm^{-1}, and the wavelength range of registration was 400-4000 cm^{-1}. The background was executed between every analysis. The spectrum data were collected by the PerkinElmer Spectrum™ 10 software.

3.3 Development of PLS prediction models

Applying the acquired databases, several models were constructed to predict either polymer, water, calcium carbonate, or titanium dioxide content, using the PerkinElmer Spectrum® Quant software. In all models, a PLS algorithm was used with random k-Fold cross-validation. Minor versions of models were created with different pre-processing treatments. The first one contained only a baseline correction using a second derivate with nine points smoothing window for noise reduction. While the second an MSC normalization was added and in the third a pre-processing named range, where only a specific range of the spectrum is used for the method's calculations, was added. The number of iterations groups (k), in which the samples were randomly divided, was also changed according to the number of standards used to create the models. Data pre-processing consisted of a baseline correction using a second derivate with nine points smoothing window for noise reduction, an MSC normalization and a specification on the range of spectra regions used for the method's calculations.

General prediction models

For both kinds of resin, and for each needed content, general models were developed. Updated versions were established according to the introduction of the reformulated coatings' data or the deletion of samples with zero percentage of the component to be predicted.

Specific contents' range models

Concern arose when the polymer content's model for styrene-acrylic resins was being tested. It was observed that in coatings with very low concentrations of this constituent, the resulting predictions had significant errors. Two ideas were considered for dealing with this obstacle. The first was to fill gaps in the lower contents' range with modified paints to assure this region was better represented. The second was to construct models based on a more limited amplitude of polymer content. With that, models containing only spectra of polymer contents within the interval of 0-6% and 0-10% were composed.

3.4 Evaluation of models' performance

In the way of evaluating predicting abilities of the calibration models, also known as reduced models, obtained by the Quant PLS analysis, a set of statistical parameters were adopted, them

being the Nash-Sutcliffe Efficiency (NSE), the percent bias (PBIAS), and the ratio of the root mean square error to the standard deviation of measured data (RSR).

Nash-Sutcliffe Efficiency (NSE)

The NSE is a normalized statistic that determines the relative magnitude of the residual variance ("noise") compared to the measured data variance ("information"). Its optimal value is 1. When between 0.0 and 1.0, the model is generally viewed as having acceptable levels of performance. Whereas, values lower than zero, indicates that the mean observed value is a better predictor than the simulated value, which indicates unacceptable performance (Moriasi *et al.*, 2007). In Equation 17, it is possible to view how the parameter is calculated.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{17}$$

where y_i is the observed value, $\overline{y_i}$ is the mean of observed data, and $\hat{y}_i$ is the estimated value for the constituent being evaluated.

Percent bias (PBIAS)

It measures the average tendency of the simulated contents to be larger or smaller than their observed counterparts. Its optimal value is 0. Positive values indicate an underestimated model; contrarily, it indicates an overestimated model (Gupta, Sorooshian and Yapo, 1999). The parameter is computed using Equation 18

$$\mathrm{PBIAS} = \left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i) \times 100}{\sum_{i=1}^{n} y_i}\right] \tag{18}$$

Ratio of the root mean square error to the standard deviation of measured data (RSR)

It is calculated as the ratio of the RMSE and standard deviation of measured data. RSR optimal value is 0, which indicates zero RMSE or residual variation and therefore, perfect model simulation (Moriasi *et al.*, 2007). It is calculated as shown in Equation 19

$$\mathrm{RSR} = \frac{\mathrm{RMSE}}{\mathrm{STDEV}} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}} \tag{19}$$

The parameters above mentioned were classified, similarly to Moriasi *et al.*, 2007, either as unsatisfactory, satisfactory, good, or very good, in correspondence with an established array of values, as represented in Table 2. Based on the resulting ratings, it was possible to conclude how well was the model's performance.

Table 2. General performance ratings for recommended statistics for model's evaluation (Moriasi *et al.*, 2007)

Performance Rating	NSE	PBIAS	RSR
Very good	$0.9 < NSE \leq 1.0$	$\lvert PBIAS \rvert < 5$	$0.0 \leq RSR \leq 0.5$
Good	$0.8 < NSE \leq 0.9$	$\lvert PBIAS \rvert < 10$	$0.5 < RSR \leq 0.6$
Satisfactory	$0.6 < NSE \leq 0.8$	$\lvert PBIAS \rvert < 15$	$0.6 < RSR \leq 0.7$
Unsatisfactory	$NSE \leq 0.6$	$\lvert PBIAS \rvert > 20$	$RSR > 0.7$

Calibration model evaluation

The data used was gathered at the end of each model's calibration, acquired with a complete statistical review performed by the software and presented in a report. At the calibration report, the experimental and estimated values for the reduced models were exported, and further statistical calculations were performed as described previously.

Independent dataset test

For validation of the models, sets of coatings were formulated, 21 using styrene-acrylics resin and 10 using acrylics resin. The constituent's contents were established to reside inside or close to their general models' range. Analyses were performed ten times for each of these paints in the FTIR-ATR. Once again, eight scans were accumulated at a resolution of 2 cm^{-1}, and the wavelength range of 400-4000 cm^{-1}. Between every analysis, background checks were done. After, a mean of the spectra was acquired. All models were tested, and statistical data, as mentioned before, was determined for each one of them.

3.5 Application of the best models to analyze the composition of coatings

Two sets of paints composed of the two different types of resin were analyzed, five being styrene-acrylics, and three acrylics coatings. Predictions of the four main components were performed by applying the best-established models for the different type of resin. Experimental results were attained from tests performed in the laboratory and which will be further explained in subsection 3.5.2. The produced contents in both analyses were then compared.

3.5.1 Model's prediction

For each kind of resin, the best content's models achieved by both the train and the test groups, regarding performance ratings, were used to perform predictions.

3.5.2 Experimental determination of constituents' content

In an analytical balance, porcelain melting pots were weighted. In each of those, was added 1 gram of coating sample well homogenized. The determinations were performed in triplicates,

and the final results were reached by their average calculation when presenting 15 % concordance.

Solid content (SC)

To determine the solid content on the sample, the melting pots were subjected to the muffle at 105 ºC for 3 hours. After, the porcelain melting pots were left resting for in the desiccator and when the room temperature was reached, their mass was determined by Equation 20

$$\text{SC}\ (\%, w/w) = \frac{W_{105^{o}C} - W_{melting\ pot}}{W_{sample}} \times 100 \quad (20)$$

where $W_{105^{o}C}$ is the mass of the porcelain melting pot after drying at 105 ºC, $W_{melting\ pot}$ is the empty porcelain melting pot weight, and W_{sample} is the mass of the sample added.

Organic compounds content (OCC)

The procedure was done by following an internal method used at Barbot, which performed the calcination of coatings samples in a muffle. The procedure allows the determination of the organic compound degraded, mainly composed of the polymer, organic solvents, and additives. To determine the organic compounds' content, Equation 21, the porcelain melting pots stated before were brought to the muffle, now at 450 ºC, for 6 hours. After the porcelain melting pots rested in the desiccator till their temperature reached room temperature, and then they were weighted. The OCC was calculated with Equation 21

$$\text{OCC}\ (\%, w/w) = \frac{W_{105^{o}C} - W_{450^{a}C}}{W_{sample}} \times 100 \quad (21)$$

where $W_{450^{a}C}$ is the mass weighted after the muffle test at 450 ºC.

Carbonates content (CC)

The melting pots and sample used for the OCC determination, Equation 22, were placed in the muffle at a temperature of 950 ºC, for 6 hours. In order to avoid thermic shock and possible sample loss, the melting pots initially cooled in the muffle and then proceeded to the desiccator where they reached room temperature, afterward, their weighted were determined. The CC was calculated with the Equation 22

$$\text{CC}\ (\%, w/w) = \frac{W_{450^{o}C} - W_{950^{a}C}}{W_{sample}} \times 2.27 \times 100 \quad (22)$$

where $W_{950^{a}C}$ is the mass weighted after the muffle test at 450 ºC and 2.27 corresponds to a stoichiometric factor which correlates the molar mass of the calcium carbonate and the carbon dioxide.

4 Results and discussion

4.1 Coatings databases

Two different sets of spectra databases, acquired by FTIR-ATR analysis, were constructed, one for each type of resin. They were separately built since the polymers' bands regarding the carbonyl bond are formed in very similar ranges of the spectrum, 1730 cm^{-1} for styrene-acrylic and 1727 cm^{-1} for acrylic (Annex 1). Since paint systems are composed of several raw materials, small variations can occur in the range of expected bands in its resulting spectra. Therefore, these slights shifts could prompt to a misinterpretation, which would result in a wrongful prediction of binder's content.

Initially, the sets consisted of spectra from coatings' samples derived from daily production containing a number of 273 and 132 standards for styrene-acrylic and acrylic coatings, respectively. Later, as previously mentioned, spectra of modified samples were introduced, resulting in two new databases of 383 and 177 standards for styrene-acrylics and acrylics coatings, correspondingly (Table 3). The strategy was taken with the intent of completing the information contained in the databases so models with better predicting abilities could be generated.

Table 3. Summary of databases range of component's contents regarding each resin kind

	Content range (%)			
Database	**Polymer**	**$CaCO_3$**	**Water**	**TiO_2**
Styrene-Acrylic resin	0.8 – 42.1	0.0 – 62.6	26.3 – 86.4	0.0 – 15.3
Acrylic resin	8.8 – 40.8	0.0 – 67.2	7.1 – 75.5	0.0 – 22.3

4.2 Development and performance evaluation of PLS models

Several predictions models were created with the spectra databases. The firsts versions were all obtained from spectra derived from daily production. After preliminary testing, it was verified the need for a better representation of some of the regions, with that, remodeled paints were made, which were then used to build optimized versions of the component's models. All first minor versions were built having baseline correction as a pre-processing treatment. In the second and third minor versions pre-processing methods were always added, namely MSC normalization and range, respectively.

For each constituent, different strategies were followed in pursuit of the best-resulting model. The calibrations performed by the PerkinElmer Spectrum® Quant software computed and presented a few model's statistical indicators, the main ones being the number of latent

variables, the coefficient of determination (also known as R^2), the standard error of estimate (SEE) and the standard error of prediction (SEP) Additionally, after every model's calibration, a review report was generated. Information about the calibration models (a group of data which accomplished a model with the best predicting ability also known as reduced models), was given, including calculated values of contents for each sample held on it. Based on these values and the actual known contents, evaluations of performance could be completed, as is referend in section 3.4.

Performance ratings were also reached from the independent dataset. This evaluation consisted in the creation of sets of paints (Table 4), for both types of coatings, which possessed specific computed compositions; in other words, were formed of unknown samples to the models and therefore, did not have any matching spectra on it. The resulting computed ratings reached from this particular assessment are of great importance once they allow a more realistic review of the model's predictability capacities.

Table 4. Summary of the ranges for each components' contents of the paints created for the independent dataset test

	Content range (%)			
Dataset	**Polymer**	**$CaCO_3$**	**Water**	**TiO_2**
Styrene-Acrylic resin	2.15 – 27.79	9.00 – 60.03	37.25 – 82.20	1.95 – 15.01
Acrylic resin	12.33 –36.99	5.46 – 30.38	38.86 – 69.94	2.22 – 20.40

With performance evaluations' statistical indicators values and their corresponded ratings, a single model was chosen for each component for the different type of resins, having a total of eight PLS predicting models with the best-obtained capacities.

4.2.1 Models for prediction of components' content in styrene-acrylics coatings

Polymer's content models

Numerous models (Table 5) for predicting polymer's content were constructed, where different approaches were taken. The need for creating a high number of models for this property had two underlying reasons. Firstly, the polymer's content is considered a crucial element of a coating's composition; hence, it demands a better accuracy in predictions results. Secondly, and more mattering, this kind of resin is used in lowers quantities in paints when compared to the acrylics type, with that, the outcomes had more significant relative errors associated with it. The different strategies were taken with the intent of minimizing those errors and achieving higher precisions in the estimated values.

Table 5. Summary and calibration statistical indicators of polymer's models for styrene-acrylics coatings

Model's name	Smp	Samples description	K-Fold	Pre-processing			LV	R^2	SEE	SEP
				BC	MSC	Range				
StyPol	273	BP	3	✓	×	×	6	0.993	0.562	0.570
StyPolOpt1.1.1	333	BP + 20 RP	12	✓	×	×	8	0.995	0.454	0.462
StyPolOpt1.1.2			22	✓	×	×	7	0.993	0.506	0.518
StyPol(0-10%)1.1.1	293	BP + RP pol<10%	12	✓	×	×	6	0.968	0.339	0.347
StyPol(0-10%)1.1.2			15	✓	×	×	7	0.975	0.302	0.311
StyPol(0-6%)1.1.1	269	BP + RP pol<6%	12	✓	×	×	6	0.973	0.235	0.239
StyPol(0-6%)1.1.2			20	✓	×	×	6	0.975	0.227	0.230
StyPolOpt2.1	377	BP + 35 RP	22	✓	×	×	8	0.994	0.463	0.471
StyPolOpt2.2				✓	✓	×	8	0.995	0.418	0.425
StyPolOpt2.3				✓	✓	✓	6	0.972	0.321	0.334
StyPol(0-10%)2.1	339	BP, RP pol<10%	15	✓	×	×	8	0.972	0.323	0.333
StyPol(0-10%)2.2				✓	✓	×	7	0.977	0.222	0.225
StyPol(0-6%)2.1	308	BP + RP pol<6%	20	✓	×	×	7	0.976	0.226	0.229
StyPol(0-6%)2.2			18	✓	✓	×	8	0.994	0.451	0.460
StyPolOpt3.1.1	383	New BP added+ RP	22	✓	×	×	8	0.994	0.463	0.472
StyPolOpt3.1.2			18	✓	×	×	8	0.994	0.463	0.472
StyPolOpt3.1.3			45	✓	×	×	8	0.994	0.463	0.472
StyPolOpt3.1.4			3	✓	×	×	8	0.994	0.463	0.472
StyPolOpt3.2			22	✓	✓	×	8	0.995	0.427	0.435
StyPolOpt3.3			22	✓	✓	✓	7	0.993	0.490	0.499

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Reformulated Paint

The first obtained model (StyPol), as mentioned before, had data acquired only from daily productions samples. The model's statistical indicators and calibration model's evaluation ratings (Table 6) were acceptable; however, when independent dataset performance evaluation was conducted results pointed out an overestimation in the predicted values, since PBIAS was equal to 12.59.

Considering the performance ratings, better spectrum representation of the polymer's content was sought as an attempt to achieve better models. Bearing in mind that approximately 86 % of the samples used to build the first model (StyPol) had a polymer content equal to or lower than 10 %, and also that this range is usual for the styrene-acrylic coatings, two approaches were considered. Twenty paints were formulated to fill gaps in the amplitude between 0-10 % of polymer's content (StyPolOpt1.1.1, StPolOpt1.1.2) and not only an optimized version of the previous model was generated, where the new samples were included, but also two other models were created having restricted ranges of polymer's contents between 0-10 % (StyPol(0-10%)1.1.1, StyPol(0-10%)1.1.2) and 0-6 % (StyPol(0-6%)1.1.1, StyPol(0-6%)1.1.2).

Table 6. Performance rating computed from values of polymer's content for styrene-acrylics coatings

	Calibration model ratings						Independent datasets ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
StyPol	0.993	VG	0.00	VG	0.08	VG	0.906	VG	12.59	S	0.31	VG
StyPolOpt1.1.1	0.995	VG	0.00	VG	0.07	VG	0.915	VG	10.19	S	0.29	VG
StyPolOpt1.1.2	0.993	VG	0.00	VG	0.08	VG	0.903	VG	12.67	S	0.31	VG
StyPol(0-10%)1.1.1	0.977	VG	0.00	VG	0.15	VG	0.671	S	-16.13	U	0.57	G
StyPol(0-10%)1.1.2	0.975	VG	0.00	VG	0.16	VG	0.694	S	-17.65	U	0.55	G
StyPol(0-6%)1.1.1	0.973	VG	0.00	VG	0.16	VG	0.110	U	-46.92	U	0.94	U
StyPol(0-6%)1.1.2	0.975	VG	0.00	VG	0.16	VG	0.133	U	-44.80	U	0.93	U
StyPolOpt2.1	0.994	VG	0.00	VG	0.08	VG	0.937	VG	4.56	VG	0.25	VG
StyPolOpt2.2	0.995	VG	0.00	VG	0.07	VG	0.941	VG	7.31	G	0.24	VG
StyPolOpt2.3	0.994	VG	0.00	VG	0.08	VG	0.943	VG	6.48	G	0.24	VG
StyPol(0-10%)2.1	0.972	VG	0.00	VG	0.17	VG	0.762	S	-17.14	U	0.49	VG
StyPol(0-10%)2.2	0.972	VG	0.00	VG	0.17	VG	0.861	G	-10.64	S	0.37	VG
StyPol(0-6%)2.1	0.977	VG	0.00	VG	0.15	VG	0.184	U	-44.97	U	0.90	U
StyPol(0-6%)2.2	0.976	VG	0.00	VG	0.16	VG	0.278	U	-42.79	U	0.85	U
StyPolOpt3.1.1-3.1.4	0.984	VG	-2.50	VG	0.13	VG	0.937	VG	4.60	VG	0.25	VG
StyPolOpt3.2	0.995	VG	0.00	VG	0.07	VG	0.943	VG	6.40	G	0.24	VG
StyPolOpt3.3	0.993	VG	0.00	VG	0.08	VG	0.941	VG	7.95	G	0.24	VG

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Also, each pair of models were constructed using the same samples and pre-processing treatments, changing only the k-fold number. This action was taken with the intent of analyzing the influence of this variable in the models' results.

Evaluations were proceeded to judge the models' performances. Regarding the general models (StyPolOpt1.1.1, StyPolOpt1.1.2), better results were attained for the model StyPolOpt1.1.1, the version with the smallest k-Fold number, equal to 12. The evaluation's parameters were slightly better, for both the calibration model and independent dataset, when compared to the other two general versions (StyPol, StyPolOpt1.1.2), and the PBIAS showed a reduction (from 12.69/12.67 to 10.19) which means the model was predicting values with better accuracy.

For the restricted range models, even though all calibration model's ratings were considered very good, the results attained from the independent dataset test had an opposing conclusion. The results confirmed the lack of extrapolation capacities of the models, since when using datasets with a larger range than the range of the actual model. The statistical indicators were significantly distant from their ideal values. Both models were proved to be underestimated (range of attained PBIAS values of -16.13 to -44.97) with significant residual values, especially the 0-6% polymer's content model (StyPol(0-6%)1.1.1, StyPol(0-6%)1.1.2). As for the k-fold influence, since inconsistent results were achieved, no reliable conclusion was able to be drawn. Further investigation was carried out in a later version.

The addition of reformulated coatings to the model was proved to be effective in increasing the general model's prediction abilities. Accordingly, 35 new reformulations were produced, analyzed in the FTIR-ATR, and introduced to generate a second optimized version of the general model.

In the second general optimized version (StyPol Opt2.1) two new minor versions (StyPol Opt2.2, StyPol Opt 2.3) were produced, which defer from the first one in pre-processing treatments, as mentioned in section 3.2. The models' statistical indicators and calibration models' performance evaluation presented very similar and satisfactory results. Nevertheless, the ratings obtained from independent validation's performance evaluation inferred the model with only baseline correction as pre-processing with having the best predicting abilities. Since its bias held the lower absolute value (4.56), even though it was considered to be slightly overestimated.

As for the limited range's models (StyPol(0-10%)2.1, StyPol(0-10%)2.2, StyPol(0-6%)2.1, StyPol(0-6%)2.2), only a second minor version was carried out. And, performance evaluation proved this later minor version to be significantly better than the first, especially in the 0-10% polymer's content model (StyPol(0-10%)2.2).

A third and last optimized version was built (StyPolOpt3.1.1-3.1.4, StyPolOpt3.2, StyPolOpt3.3). It presented now the spectra of all styrene-acrylic coatings produced at Barbot. In this final version, the k-fold influence was once again studied. The same strategy as for the first optimized version attempt was taken, but four similar models (StyPolOpt3.1.1-3.1.4) were created with different k-fold numbers, 3, 18, 22, and 45. The models' statistical indicators and consequently, the performance ratings derived from both evaluations were identical. With that, it is possible to presume that different k-Fold numbers for random cross-validation can achieve the same results, hence changing between them does not bear any significance to the resulting model. The best way of choosing the most appropriated one is only based on the statistical indicators obtained after the calibration of each model. In other words, different k numbers should be tested until the best calibration review is conferred.

Styrene polymer best model – StyPolOpt2.1

The best obtained model for polymer's content prediction in styrene-acrylics coatings was the model named as "StyPolOpt2.1".

The model was tested as having different latent variable numbers, for each of them the SEP estimate number was computed (Figure 4) and also the cumulative X and Y-variance, which shows how much of both variables can be explained by the model. The optimum number of LV's was determined as 8, having a SEP estimate of 0.47 and as expected translated in a much better explanation of the variables (having a cumulative X-variance and Y-variance of 50.1 % and 99.4

%, respectively, Table 7) when compared to the same model containing only one latent variable. This optimum number was established by the performance of an F-test.

Table 7. Cumulative variance values for StyPolOpt2.1 model for the different number of LV

LV	Cum X-Var (%)	Cum Y-Var (%)
1	25.2	84.6
2	34.2	93.6
3	40.2	96.4
4	43.1	97.9
5	45.0	98.7
6	46.9	99.1
7	48.4	99.3
8	50.1	99.4

LV = Latente Variables, Cum = Cumulative, Var = Variance

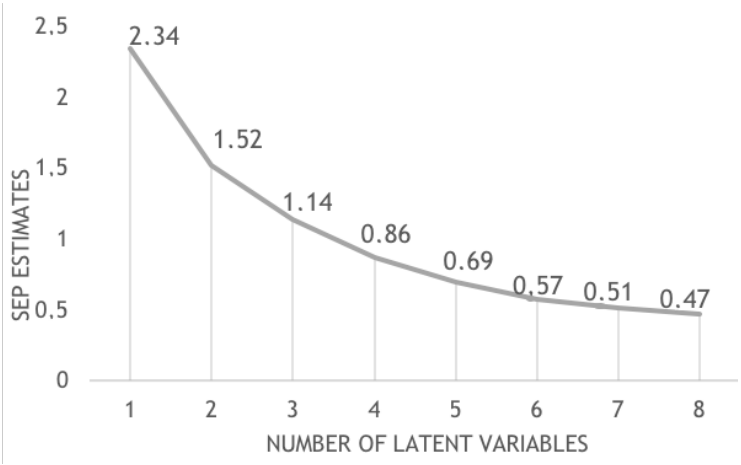

Figure 4. SEP estimates as a function of Latent variables for "StyPolOtm2.1" model.

A first visible confirmation is that most samples are indeed comprised in the lowers content ranges, up to 9 % (Figure 5-a). It is possible to verify that the model can be considered as having a good bias since computed values for PBIAS were 0.0 and 4.56 for the calibration model (Figure 5-a) and independent dataset test, respectively (Table 6). Even though slight overestimation can be observed in the validation test (Figure 5-b), both performance evaluation ratings were very good.

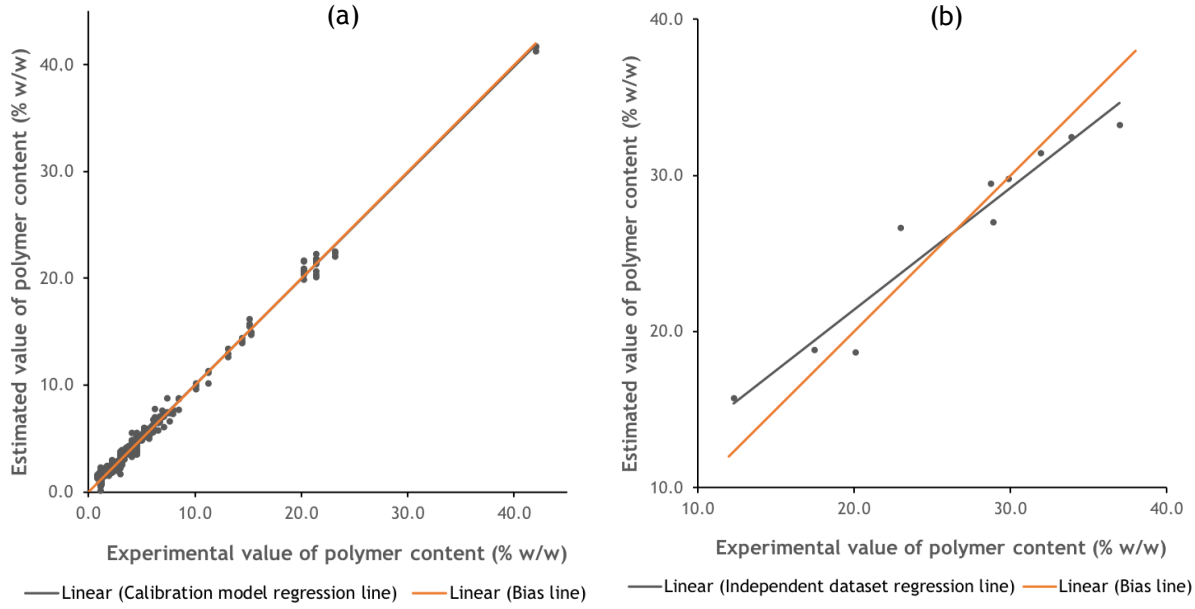

Figure 5. Estimated values as a function of experimental values of polymer content attained from "StyPolOpt2.1" calibration model (a) independent dataset test (b).

Also, both regressions lines presented concordant results, having NSE indicators of 0.994 and 0.937 and RSR indicators of 0.08 and 0.25, for calibration model and independent dataset, correspondingly. Their ratings were also classified as very good, which pointed to a model very well adjusted to the data, presenting all points with no or little significant residual variance.

It is possible to confirm that results from both performances evaluations are concordant, and prediction abilities of the attained model for polymer's content prediction in styrene-acrylics resin were good and, most importantly, the model was able to demonstrate its applicability.

Calcium carbonate's content models

Extenders present a particular situation. In coatings like varnishes which are transparent, there is no amount of extenders of any kind, including calcium carbonates, present in their composition. With that in mind, it was considered that when present in coatings, more precisely in paints or primers, the lowest content typically used in styrene-acrylics is close to 8 % and in acrylics is 5 %. Moreover, by having spectra of coatings with no content of this component in the database used to build the models (Table 8), the gap between the 0 % and the lowest content of the other type of coatings that present this constituent, could affect the regression line and consequently the accuracy of predictions. In search of a better understanding of the influence these coatings' spectra presence in the models, versions without data derived from them were built ($StyCaCO_3$1.3, $StyCaCO_3$1.4, $StyCaCO_3$Opt1.3, $StyCaCO_3$Opt1.4).

Table 8. Summary and calibration statistical indicators of calcium carbonate's models for styrene-acrylics coatings

				Pre-Processing						
Model's name	**Smp**	**Samples description**	**k-Fold**	**BC**	**MSC**	**Range**	**LV**	**R^2**	**SEE**	**SEP**
$StyCaCO_3$1.1	278	BP	25	✓	×	×	7	0.978	2.46	2.51
$StyCaCO_3$1.2				✓	✓	×	7	0.978	2.48	2.52
$StyCaCO_3$1.3	269	BP no varnishes		✓	×	×	6	0.976	2.36	2.40
$StyCaCO_3$1.4				✓	✓	×	5	0.966	2.77	2.81
$StyCaCO_3$Opt1.1	383	BP + 35 RP	35	✓	×	×	8	0.977	2.30	2.33
$StyCaCO_3$Opt1.2				✓	✓	×	7	0.973	2.27	2.30
$StyCaCO_3$Opt1.3	374	BP + 35 RP no varnishes		✓	×	×	7	0.973	2.27	2.30
$StyCaCO_3$Opt1.4			12	✓	✓	×	7	0.973	2.27	2.29

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

Results gathered from the calibration models' reviews, and performance ratings were all refereed as very good, and little difference between statistical indicators values can be observed. However, when analyzing the results attained from independent validation (Table 9),

the first conclusion to be drawn was that models with the varnishes spectra – $StyCaCO_3$1.1, $StyCaCO_3$1.2, $StyCaCO_3$Opt1.1, and $StyCaCO_3$Opt1.2 – presented the worst performance, showing significant underestimated content's values having computed PBIAS of -27.82, -25.85, -19.87, and -15.40, correspondingly.

Table 9. Performance rating computed from values of Calcium Carbonate's content for styrene-acrylics coatings

	Calibration model ratings						Independent dataset ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
$StyCaCO_3$1.1	0.978	VG	0.00	VG	0.15	VG	0.537	U	-27.82	U	0.68	S
$StyCaCO_3$1.2	0.978	VG	0.00	VG	0.15	VG	0.596	U	-25.85	U	0.64	S
$StyCaCO_3$1.3	0.976	VG	0.00	VG	0.15	VG	0.798	S	9.23	G	0.45	VG
$StyCaCO_3$1.4	0.966	VG	0.00	VG	0.18	VG	0.645	S	14.95	S	0.60	G
$StyCaCO_3$Opt1.1	0.973	VG	0.00	VG	0.16	VG	0.706	S	-19.87	U	0.54	G
$StyCaCO_3$Opt1.2	0.975	VG	0.00	VG	0.16	VG	0.674	S	-15.40	U	0.57	S
$StyCaCO_3$Opt1.3	0.973	VG	0.00	VG	0.16	VG	0.630	S	13.98	S	0.61	S
$StyCaCO_3$Opt1.4	0.973	VG	0.00	VG	0.16	VG	0.630	S	13.98	S	0.61	S

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Calcium carbonate best Model – $StyCaCO_3$1.3

The model with best performance ratings for predicting calcium carbonate's content in coatings containing styrene-acrylics resin was the model named as "StyCaCO31.3". The optimum number of latent variables for this model was 5 with a SEP estimate equals to 2.81 and a cumulative X and Y-variance of 47.4 % and 97.6 %, as suggests Table A1.1 and Figure A1.1.

Initially, predictions reached by this model suggested no presence of bias, and calculation of PBIAS from values gathered from calibration review report confirms this statement. However, after analyses of independent dataset results, and calculation of PBIAS (9.23), an overestimation bias was observed, which indicated predictions values to present higher contents of $CaCO_3$ than the actual sample content's value. The NSE indicators of the calibration model and independent dataset regression lines (Figure 6), were rated as very good and good, having values of 0.976 and 0.798, and both RSR were considered very good, 0.15 and 0.45, correspondingly. The indicators computed describe a fairly-adjusted model with no significant residual variance between estimated and experimental contents' values.

Water's content models

The approach selected for the water content's models were the simplest and identical for both kinds of resin. It can be explained by the fact that the high amounts of water used in coatings generate small relative prediction errors. Thus, a high accuracy of the model is not an essential factor. Also, in the case of uncertainty, verification of water's content can be determined by

performing the muffle test as described in section 3.5.2 or a more rapid one: *Halogen Moisture Analyzer.* While there are available tests able to indicate a very accurate content of water presented in the coating, they have the disadvantage of taking several minutes or even hours to generate the results required. By using the work in development, only a few minutes would be necessary to attain such information. Four models were constructed to determine the content of the constituent in question (Table 10).

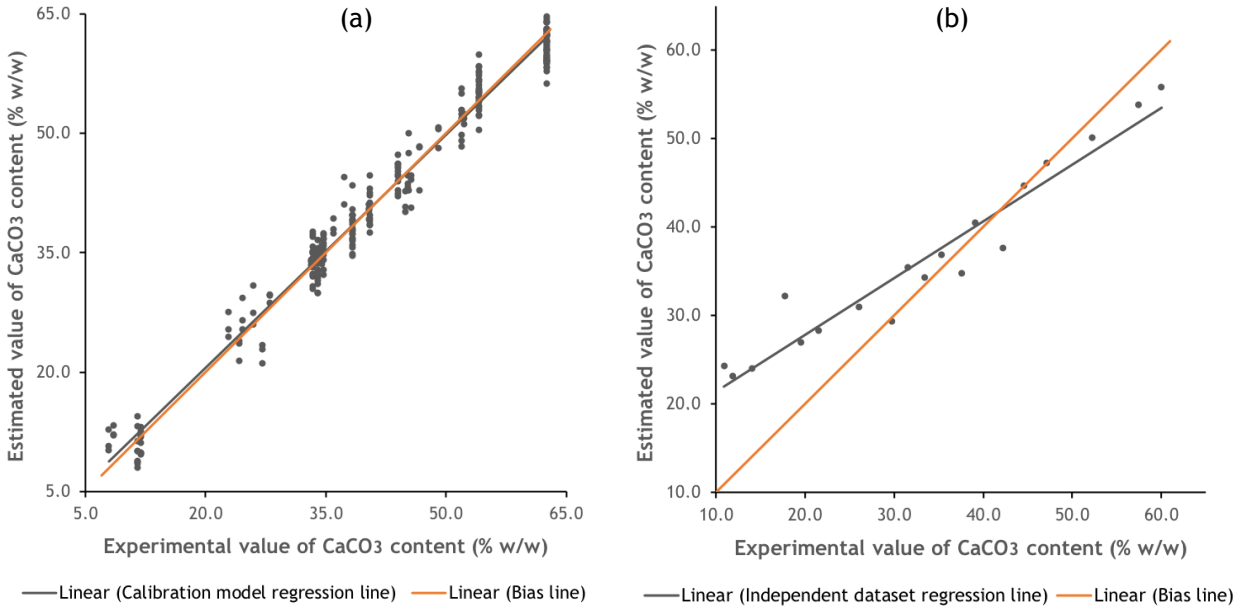

Figure 6. Estimated values as a function of experimental values of $CaCO_3$ content attained from "StyCaCO21.3" calibration model (a) and independent dataset test (b).

Table 10. Summary and calibration statistical indicator of calcium carbonate's models for styrene-acrylics coatings

Model's name	Smp	Samples description	k-Fold	Pre-Processing			LV	R^2	SEE	SEP
				BC	MSC	Range				
StyWtr1.1	278	BP	8	✓	×	×	6	0.935	2.12	2.14
StyWtr1.2			20	✓	✓	×	7	0.945	1.95	1.98
StyWtrOpt1.1	383	BP + 35 RP	12	✓	×	×	7	0.938	1.95	2.00
StyWtrOpt1.2			8	✓	✓	×	7	0.936	1.97	2.00

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

All calibration models presented excellent evaluation concerning the reduced model's performance ratings (Table 11). Moreover, very few distinctions can be made regarding the results achieved through independent validation. The main difference is that the optimized models –StyWtrOpt1.1, StyWtrOpt1.2 – are producing slightly less overestimated values than the first versions' models – StyWtr1.1, StyWtr1.2 – which can be observed by the calculated

PBIAS results which are lower for the optimized versions (8.78 and 9.39, respectively). This means that, even though in all models, the estimated content is higher than the experimental content value, the later versions present slightly more accurate results.

Table 11. Performance rating computed from values of water's content for styrene-acrylics coatings

Model's name	Calibration model ratings						Independent dataset ratings					
	NSE	PI	PBIAS	PI	RSR	PI	NSE	PI	PBIAS	PI	RSR	PI
StyWtr1.1	0.939	VG	0.04	VG	0.25	VG	0.53	U	9.57	G	0.69	S
StyWtr1.2	0.947	VG	0.07	VG	0.23	VG	0.53	U	9.62	G	0.69	S
StyWtrOpt1.1	0.938	VG	0.00	VG	0.25	VG	0.54	U	8.78	G	0.68	S
StyWtrOpt1.2	0.936	VG	0.00	VG	0.25	VG	0.51	U	9.39	G	0.70	S

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Water best Model – StyWtrOpt1.1

The best model for water's content prediction in styrene-acrylics paints was the StyWtrOpt1.1. The selected latent variable number for the model was 7, having a cumulative X and Y-variances and a SEP estimate associated to it of 48.2 %, 93.8 %, and 2.00, correspondingly, Table A1.2 and Figure A1.2.

Analysis of the calibration model regression line (Figure 7-a) confirms ratings of performance evaluation determined as very good. The PBIAS, NSE, and RSR had computed values of 0.0, 0.938, and 0.25, respectively. The first observable characteristic of the validation test (Figure 7-b) is the model's overestimation bias toward predicted values, and calculated PBIAS of 8.78 corroborates with this assumption. Performance ratings were found unsatisfactory and satisfactory for the NSE and RSR parameters, which presented values of 0.54 and 0.68, correspondingly.

Important information, which can be noticeable by analyzing both graphics, is that there is a range, around 30-50 %, where most samples where comprised. It was expected since this region is so well characterized, high accuracy of prediction in contents of paints held within this interval. However, with results from the independent dataset test, in this range of content, all predictions presented overestimated values. This could mean that the model, even though well represented in those regions, drawn wrong assumptions because other ranges were not as well defined or spectra of coatings with the same content of water presented contradictory spectra. A possible solution to the problem could be producing more coatings with different water contents and introducing their spectra in the database as a way of allowing better identification of the changes caused by different contents quantities in spectra. By taking this approach, the model could then make more coherent distinctions and, as a consequence, predict more reliable results.

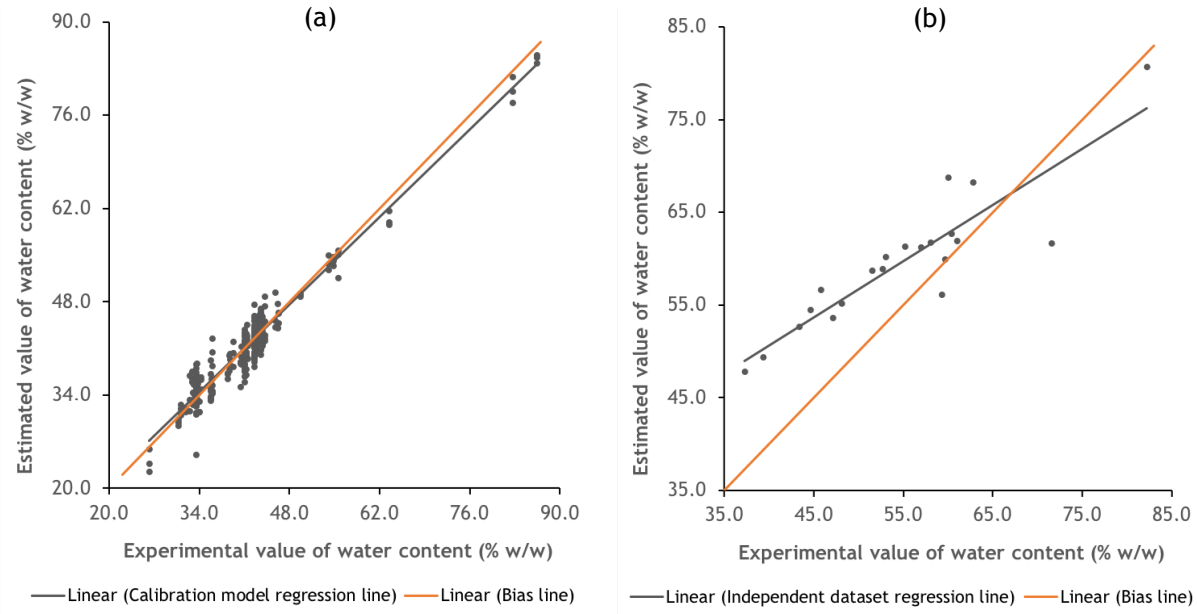

Figure 7. Estimated values as a function of experimental values of water content attained from "StyWtrOpt1.1" calibration model (a) and independent dataset test (b).

Titanium dioxide's content models

Strategies for titanium dioxide (TiO_2) models' development (Table 12) did not differ based on the resin type. Since varnishes also do not present any content of TiO_2 in their composition, in the second version optimized versions of models ($StyTiO_2Opt2.1$-2.3), no spectrum corresponding to this type of coating was introduced. Also, is worth mentioning, that spectra database was very limited in range, since as specified before, this constituent has a high price associated with it, and therefore it is not frequently used in large amounts.

Different from previous constituent's contents models, a third minor version ($StyTiO_2Opt1.3$, $StyTiO_2Opt2.3$) was created. It consisted of applying a specific range as a pre-processing treatment. In the case of TiO_2, the bands responsible for its identification in a spectrum are located close to the region of 500 - 400 cm^{-1} (Bobrova *et al.*, 1968). Hence this was defined as the range to be taken into consideration in the pre-processing of standards.

Calibration models' performance ratings revealed acceptable results for all models, although the minor third versions presented slight poorest values. Ratings from Independent validation analyses demonstrated unsatisfactory predictions produced by most models. Table 13 presents performance ratings computed for this constituent's developed models.

Titanium dioxide best Model – Sty TiO_2Opt2.1

The most adjusted model for estimation of titanium dioxide's content attained for coatings composed of styrene-acrylics resin was the $StyTiO_2Opt2.1$. It was created with spectra only from coatings containing this component.

Table 12. Summary and statistical indicators calibration of titanium dioxide's models for styrene-acrylics coatings

Model's name	Smp	Samples description	k-Fold	Pre-Processing			LV	R^2	SEE	SEP
				BC	MSC	Range				
StyTiO2Opt1.1	383	BP + 35 RP	18	✓	×	×	7	0.971	0.73	0.74
StyTiO2Opt1.2				✓	✓	×	7	0.970	0.75	0.76
StyTiO2Opt1.3				✓	✓	✓	4	0.808	1.87	1.91
StyTiO2Opt2.1	335	BP + 35 RP no varnishes	10	✓	×	×	6	0.965	0.76	0.76
StyTiO2Opt2.2			18	✓	✓	×	6	0.961	0.80	0.81
StyTiO2Opt2.3			10	✓	✓	✓	4	0.771	1.9	1.96

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

Table 13. Performance rating computed from values of titanium dioxide's content for styrene-acrylics coatings

Model's name	Calibration model ratings						Independent dataset ratings					
	NSE	PI	PBIAS	PI	RSR	PI	NSE	PI	PBIAS	PI	RSR	PI
StyTiO2Opt1.1	0.971	VG	0.00	VG	0.17	VG	0.278	U	-34.14	U	0.85	U
StyTiO2Opt1.2	0.813	G	0.00	VG	0.43	VG	0.506	U	-24.14	U	0.70	U
StyTiO2Opt1.3	0.808	G	0.00	VG	0.44	VG	0.272	U	-29.08	U	0.85	U
StyTiO2Opt2.1	0.965	VG	0.00	VG	0.19	VG	0.750	S	-15.00	S	0.50	VG
StyTiO2Opt2.2	0.970	VG	-0.56	VG	0.17	VG	0.316	U	-31.83	U	0.83	U
StyTiO2Opt2.3	0.771	S	0.00	VG	0.48	VG	-0.690	U	-6.35	G	1.30	U

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

The optimum number for the latent variables was defined as 6 having an associated SEP estimate of 0.76, and cumulative X and Y-variances were 43.3 % and 96.5 %, correspondingly, as it is indicated in Table A1.3 and Figure A1.3.

The regression line resulted from the calibration model (Figure 9-a) had all statistical indicators classified as very good regarding the performance evaluation. Its calculated values for NSE of 0.964 and RSR of 0.19, indicated a very well-adjusted model with minimal residual variance between the experimental and estimated values. The bias of the model was considered very good as well, being the computed value of PBIAS equal to zero.

The regression line from the independent dataset test (Figure 9-b) results had a computed value of NSE of 0.75 and RSR of 0.50, being judged as satisfactory and very good, respectively. However, when analyzing Figure 8-a, it is noticeable that a consistent underestimation of the predicted values, confirmed by a PBIAS calculated value of -15.0.

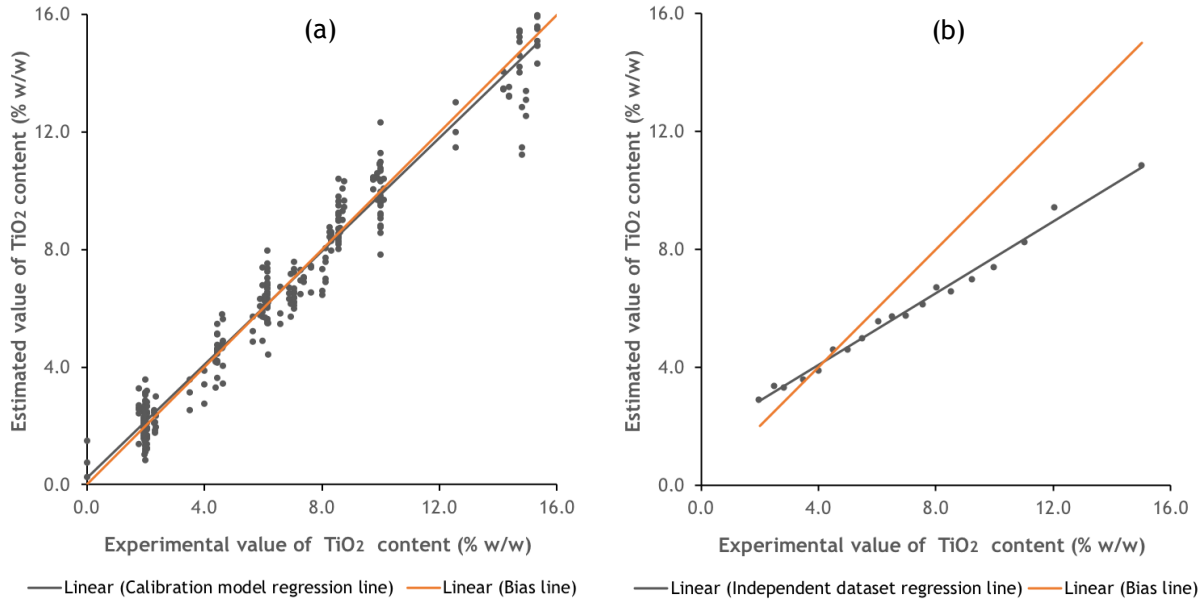

Figure 8. Estimated values as a function of experimental values of TiO_2 content attained from "StyTiO2Opt2.1" calibration model (a) and independent dataset test (b).

4.2.2 Models for prediction of components' content in acrylics coatings

Polymer's content models

Since the quantities of polymer in acrylic coatings were observed to be higher than in styrene-acrylics coatings, and the limited range approach did not generate satisfactory results, only four models were built for predicting polymers content in acrylics coatings (Table 14).

Table 14. Summary and statistical indicators calibration of polymer's models for acrylics coatings

				Pre-Processing						
Model's name	**Smp**	**Samples description**	**k-Fold**	**BC**	**MSC**	**Range**	**LV**	**R^2**	**SEE**	**SEP**
AcrylPol1.1	132	BP	8	✓	×	×	4	0.963	1.62	1.65
AcrylPol1.2				✓	✓	×	3	0.967	1.53	1.55
AcrylPolOpt1.1	177	BP + 15 RP	8	✓	×	×	3	0.933	2.12	2.20
AcrylPolOpt1.2				✓	✓	×	3	0.958	1.73	1.75

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

Calibration's parameters and reduced model's ratings were more satisfying for the models made only from Barbot's productions (AcrylPol1.1, AcylPol1.2) while independent validation's ratings pointed out better performances for the minor versions only with baseline correction pre-processing (AcrylPol1.1, AcrylPolOpt1.1). In Table 15, the obtained performance ratings can be observed.

Table 15. Performance rating computed from values of polymer's content for acrylics coatings

	Calibration model ratings						Independent dataset ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
AcrylPol1.1	0.963	VG	0.00	VG	0.19	VG	0.913	VG	0.01	VG	0.29	VG
AcrylPol1.2	0.967	VG	0.00	VG	0.18	VG	0.922	VG	2.54	VG	0.28	VG
AcrylPolOpt1.1	0.933	VG	0.00	VG	0.26	VG	0.929	VG	2.52	VG	0.27	VG
AcrylPolOpt1.2	0.958	VG	0.00	VG	0.20	VG	0.921	VG	3.50	VG	0.28	VG

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Acrylic polymer best Model – AcrylPol1.1

The model with best predictable abilities for polymer's content in acrylics coatings was the "AcrylPol1.1". As for latent variables, the optimum number was defined as 4, having a cumulative X and Y-variances of 41.4 % and 96.3 %, respectively, and a SEP estimate of 1.65. In this specific case, the number of latent variables was considerably smaller than the one with minimum SEP estimate associated (Table A1.4 and Figure A1.4). This can happen if the F-test determines that there is no significant difference in the model's results when it is calibrated with those numbers of latent variables (4 and 8). Once one of the intents of the PLS algorithm is to represent a system by using the smallest number of latent variables as possible, it chooses the optimum number of latent variables as being 4.

The model is well adjusted, having values of NSE and RSR of 0.963 and 0.19, correspondingly, and also has a good estimation, presenting a calculated PBIAS of zero. An observation to be made when analyzing Figure 9-a is that the concentration of the majority of samples used to develop the model contained polymer's content in the range of 14-25 %.

For the independent validation (Figure 9-b), although the regression line does not present a flawless adjustment, once its NSE and RSR are 0.913 and 0.29, the predictions do not present a substantial residual variation and model is still considered very good. It produces an overall good estimation of content values, having a computed value for PBIAS of just 0.01.

Calcium carbonates' content models

The creation of models for predicting this component content in acrylic coatings followed the same line of thinking, regarding as which samples were included and pre-processing treatments were used, of when creating the styrene-acrylics models for this constituent and a summary is found in Table 16.

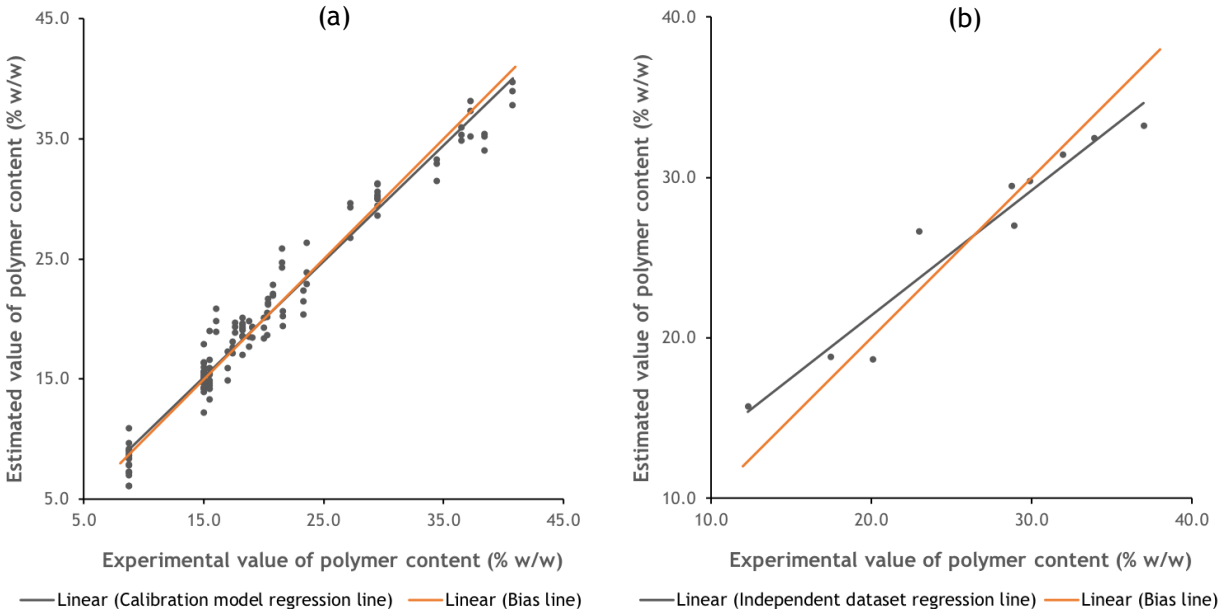

Figure 9. Estimated values as a function of experimental values of polymer content attained from "AcrylPol1.1" calibration model (a) and independent dataset test (b).

Table 16. Summary and statistical indicators calibration of calcium carbonate's models for acrylics coatings

Model's name	Smp	Samples description	k-Fold	Pre-Processing			LV	R^2	SEE	SEP
				BC	MSC	Range				
AcrylCaCO$_3$1.1	132	BP	5	✓	×	×	3	0.942	2.94	2.98
AcrylCaCO$_3$1.2				✓	✓	×	4	0.964	2.32	2.38
AcrylCaCO$_3$1.3	99	BP no varnishes	3	✓	×	×	5	0.983	1.19	1.22
AcrylCaCO$_3$1.4				✓	✓	×	3	0.950	2.02	2.06
AcrylCaCO$_3$Opt1.1	177	BP + 35 RP	8	✓	×	×	2	0.883	3.87	3.91
AcrylCaCO$_3$Opt1.2				✓	✓	×	2	0.855	4.33	4.37
AcrylCaCO$_3$Opt1.3	144	BP + 35 RP no varnishes	5	✓	×	×	2	0.823	3.65	3.68
AcrylCaCO$_3$Opt1.4				✓	✓	×	2	0.819	3.69	3.70

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

Performance evaluation (Table 17) of calibration models showed better results for models created from daily (AcrylCaCO$_3$1.1-1.4) production but opposing conclusions could be drawn when examining ratings gathered from validation test (AcrylCaCO$_3$Opt1.1-1.4). Most models obtained negative NSE, which is a concern, once, as stated before, it means that the mean observed value is a better predictor than the estimated value. Since this is considered unacceptable for a good prediction model, the models which possessed this result were not analyzed any further. As for the remaining models (AcrylCaCO$_3$Opt1.1, AcrylCaCO$_3$Opt1.3), two consisted of models containing varnishes spectra and all only had baseline correction as a pre-processing treatment.

Table 17. Performance rating computed from values of calcium carbonate's content for acrylics coatings

	Calibration model ratings						Independent dataset ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
Acryl$CaCO_3$1.1	0.942	VG	0.00	VG	0.24	VG	0.023	U	-14.14	S	0.99	U
Acryl$CaCO_3$1.2	0.964	VG	0.00	VG	0.19	VG	-0.095	U	-15.74	U	1.05	U
Acryl$CaCO_3$1.3	0.987	VG	0.00	VG	0.11	VG	-0.816	U	141.73	U	1.35	U
Acryl$CaCO_3$1.4	0.963	VG	0.00	VG	0.19	VG	-0.721	U	25.48	U	1.31	U
Acryl$CaCO_3$Opt1.1	0.883	G	0.00	VG	0.34	VG	0.108	U	-10.72	S	0.94	U
Acryl$CaCO_3$Opt1.2	0.855	G	0.00	VG	0.38	VG	-0.013	U	-6.90	G	1.01	U
Acryl$CaCO_3$Opt1.3	0.853	G	0.00	VG	0.38	VG	0.033	U	3.96	VG	0.98	U
Acryl$CaCO_3$Opt1.4	0.850	G	0.00	VG	0.39	VG	-0.255	U	2.90	VG	1.12	U

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Calcium carbonate best Model – Acryl$CaCO_3$Opt1.3

In acrylics coatings, the model "Acryl$CaCO_3$Opt1.3" was found to be the best-representing model for this constituent. It is composed with spectra only from coatings where this constituent is present. The number of latent variables considered for this model as optimum was 2. A SEP estimate, cumulative X, and Y-variance of 3.68, 24.7 %, and 82.3 %, is associated with it, as can be observed in Table A1.5 and Figure A1.5.

The calibration model can be considered as possessing no bias, which not only can be confirmed by analyzing Figure 10-a but also by the calculated PBIAS, equal to 0.0 with an NSE and RSR of 0.823 and 0.38, respectively, it is notable the presence of residual variance.

The results achieved from the independent dataset (Figure 10-b) test for the model were not so satisfactory; however, it presents interesting data. The results gathered are approximately half overestimated, and the other half is underestimated. When the PBIAS is calculated (Equation 2), a sum of variances is made. Since the resulting value of 3.96 is relatively close to zero, because there is a compensation, the final score of the parameter is viewed as very good when, in reality by evaluating the data, the model does not seem to present accurate results.

Hence, the importance of the graphic analysis for this statistical indicator. It is possible to determine that the angles and the crossing position between those two regression lines (independent dataset and bias) are also important for bias analysis. The bigger the angle between them, the worst the model's bias is. Moreover, depending on where the lines cross each other, will determine the type of bias the model presents. If the crossing happens in the lower contents' region, the model will present underestimated values, on the contrary, if they crossed in the higher content ranges than the model would perform overestimation.

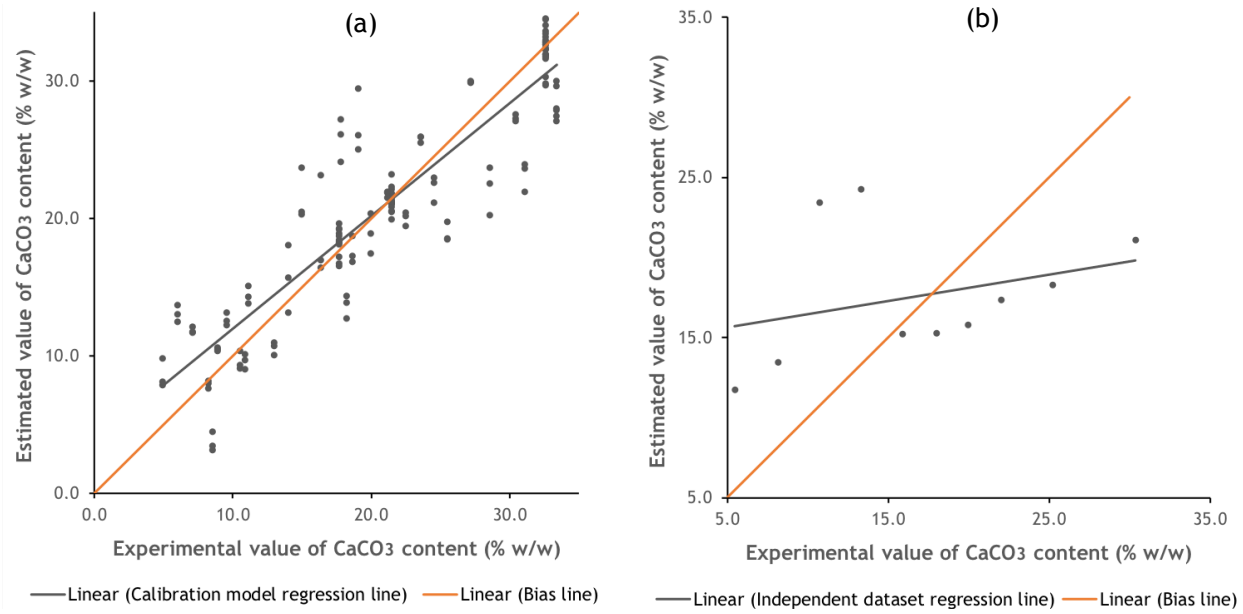

Figure 10. Estimated values as a function of experimental values of $CaCO_2$ content attained from "AcrylCaCO₂Opt1.3" calibration model (a) and independent dataset (b).

Water's content models

PLS models constructed for the component's content prediction in acrylic coatings are represented in Table 18.

Table 18. Summary and statistical indicators calibration of water's models for acrylics coatings

Model's name	Smp	Samples description	k-Fold	Pre-Processing			LV	R^2	SEE	SEP
				BC	MSC	Range				
AcrylWtr1.1	132	BP	8	✓	×	×	4	0.923	2.66	2.71
AcrylWtr1.2				✓	✓	×	4	0.909	2.89	2.98
AcrylWtrOpt1.1	177	BP + 15 RP	8	✓	×	×	4	0.866	3.91	3.94
AcrylWtrOpt1.2				✓	✓	×	4	0.860	4.01	4.07

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

In this type of resin, conflicting performance evaluation (Table 19) outcomes were observed for the calibration model, and independent dataset test collected data. In the first, the best results were attained from the model with only daily productions data and minor versions conducting baseline correction exclusively (AcrylWtr1.1). However, in the validation test significantly better performance ratings were attained by the optimized version and minor versions with not just baseline correction but also MSC normalization pre-processing (AcrylWtrOpt1.2).

Table 19. Performance rating computed from values of water's content for acrylics coatings

	Calibration model ratings						Independent dataset ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
AcrylWtr1.1	0.923	VG	0.00	VG	0.28	VG	0.496	U	-9.40	G	0.71	U
AcrylWtr1.2	0.909	VG	0.00	VG	0.30	VG	0.543	U	-9.75	G	0.68	S
AcrylWtrOpt1.1	0.866	G	0.00	VG	0.37	VG	0.674	S	-4.35	VG	0.57	G
AcrylWtrOpt1.2	0.860	G	0.00	VG	0.37	VG	0.740	S	-3.95	VG	0.51	G

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

Acryl water best Model – AcrylWtrOpt1.2

The preferred estimation model for water content in acrylic coatings was the AcrylWtrOpt1.2. For the model in question, the latent variable number defined as optimum was 4, having cumulative X and Y-variances of 35.8 % and 86.0 %, correspondingly, generated SEP estimate of 4.07 (Table A1.6 and Figure A1.6).

The calibration model regression line presents significant residual variance, but still, the model's statistical indicators are considered as good and very good, presenting values of 0.860 and 0.37 for the NSE and RSR, respectively. The adjustment of the independent dataset regression line was considered satisfactory, even though significant residual variation can be observed. The values calculated for NSE and RSR parameters were 0.74 and 0.51, correspondingly. The computed PBIAS indicates no significant bias, and a similar deduction could be made by graph analyses (Figure 11-a). However, independent dataset regression line results, noticeably reveal the underestimation bias (Figure 11-b), and the PBIAS value of -3.95 shows agreeance.

Values presenting more inaccurate estimations present content values in ranges where there is more than one sample with the same experimental content presenting different estimated value attributed to it. This can indicate a possible misinterpretation of the spectrum by the model which would translate into an incorrect prediction. Reducing this error can be done by instead of introducing triplicates of samples' spectrum in the databases used in the construction of models, the use of a medium spectrum should be preferred. Also, since variation in between spectra can be quite significant, more reads should be performed, in order to achieve more accurate results.

Titanium dioxide's content PLS models

Summary of models generated for the prediction of the pigment's content in question is represented in Table 20.

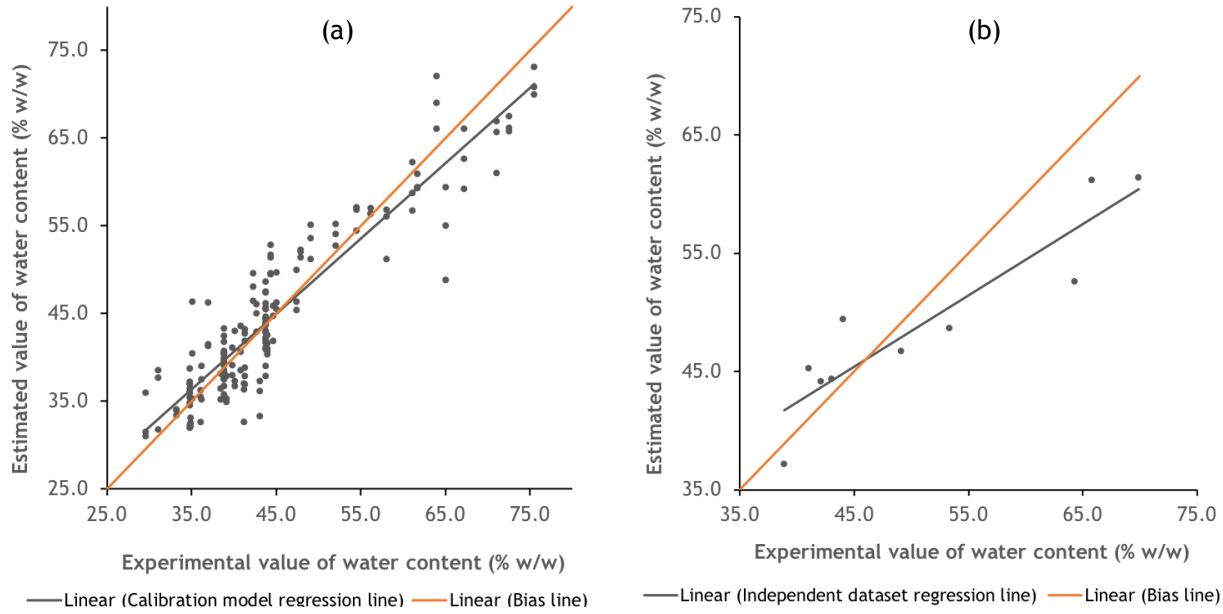

Figure 11. Estimated values as a function of experimental values of water content attained from "AcrylWtrOpt1.2" calibration model (a) and independent dataset test (b).

Table 20. Summary and statistical indicators calibration of water's models for acrylics coatings

				Pre-Processing						
Model's name	**Smp**	**Samples description**	**k-Fold**	**BC**	**MSC**	**Range**	**LV**	**R^2**	**SEE**	**SEP**
AcrylTiO₂Opt1.1				✓	×	×	7	0.989	0.75	0.76
AcrylTiO₂Opt1.2	174	BP + 15 RP	14	✓	✓	×	7	0.986	0.83	0.85
AcrylTiO₂Opt1.3				✓	✓	✓	4	0.947	1.63	1.68
AcrylTiO₂Opt2.1				✓	×	×	6	0.985	0.73	0.75
AcrylTiO₂Opt2.2	132	BP + 15 RP no varnishes	12	✓	✓	×	6	0.981	0.82	0.85
AcrylTiO₂Opt 2.3				✓	✓	✓	3	0.906	1.80	1.85

Smp = number of samples, BC = Baseline Correction, LV = Latent Variables, MSC = Multiplicative Scatter Correction, SEP = Standard Error of Prediction, SEE = Standard Error of Estimate, BP = Barbot's Production, RP = Remodeled Paint

Performance scores for this kind of resin were significantly better. All models' parameters were judged as very good, in the calibration model's evaluation. The versions with the varnish's spectra (AcrylTiO₂ Opt1.1-1.3) and the minor version with baseline correction and MSC normalization pre-processing of samples (AcrylTiO₂Opt1.2, AcrylTiO₂Opt2.2), was able to make more accurate predictions. In Table 21, parameters and ratings calculated for each model can be observed.

Titanium dioxide best Model - AcrylTiO₂Opt1.2

The best predicting model for TiO_2 content in acrylics coatings was AcrylTiO₂ Opt1.2. The optimum number of latent variables was 7, and the attained SEP estimate was 0.85. Cumulative X and Y-variance was 54.1 % and 98.6 % (Table A1.7 and Figure A1.7). The calibration regression line was considered well-adjusted with no meaningful residual variance between values,

presenting the values for NSE and RSR of 0.813 and 0.17, correspondingly. Calculated PBIAS of 0, shows no model's bias as the line is almost overlapped with the bias line. Moreover, the regression line shows a good adjustment with no significant residual variance between values (Figure 12-a).

Table 21. Performance rating computed from values of titanium dioxide's content for acrylics coatings

	Calibration model's ratings						Independent dataset ratings					
Model's name	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**	**NSE**	**PI**	**PBIAS**	**PI**	**RSR**	**PI**
AcrylTiO_2Opt1.1	0.989	VG	0.00	VG	0.10	VG	0.937	VG	-9.33	G	0.25	VG
AcrylTiO_2Opt1.2	0.986	VG	0.00	VG	0.12	VG	0.948	VG	0.74	VG	0.23	VG
AcrylTiO_2Opt1.3	0.947	VG	0.00	VG	0.23	VG	0.882	G	-6.86	G	0.34	VG
AcrylTiO_2Opt2.1	0.985	VG	0.00	VG	0.12	VG	0.894	G	4.74	VG	0.33	VG
AcrylTiO_2Opt2.2	0.981	VG	0.00	VG	0.14	VG	0.915	VG	9.06	G	0.29	VG
AcrylTiO_2Opt 2.3	0.906	VG	0.00	VG	0.31	VG	0.895	V	-5.04	G	0.32	VG

PI = Performance Index, VG =Very good, G = Good, S = Satisfactory, U = Unsatisfactory

The independent dataset results presented similar statistical indicators to the calibration model. The regression line was determined as having very good adjustment, and residual variance between experimental and estimated values were not considered as being substantial. Computed values of o.948 and 0.23 for the parameters NSE and RSR, respectively, corroborates with the conclusions gathered from the plot analysis (Figure 12-b). Also, the model's bias could be considered as insignificant once, since the two lines are very close to each other throughout the entire range, and the regression line seems to compensate the over with the underestimated values. With calculation of the PBIAS the previous assumption was able to be confirmed having the parameter a value of 0.74. The model attained can be presumed as a generator of good estimation in the hole range tested.

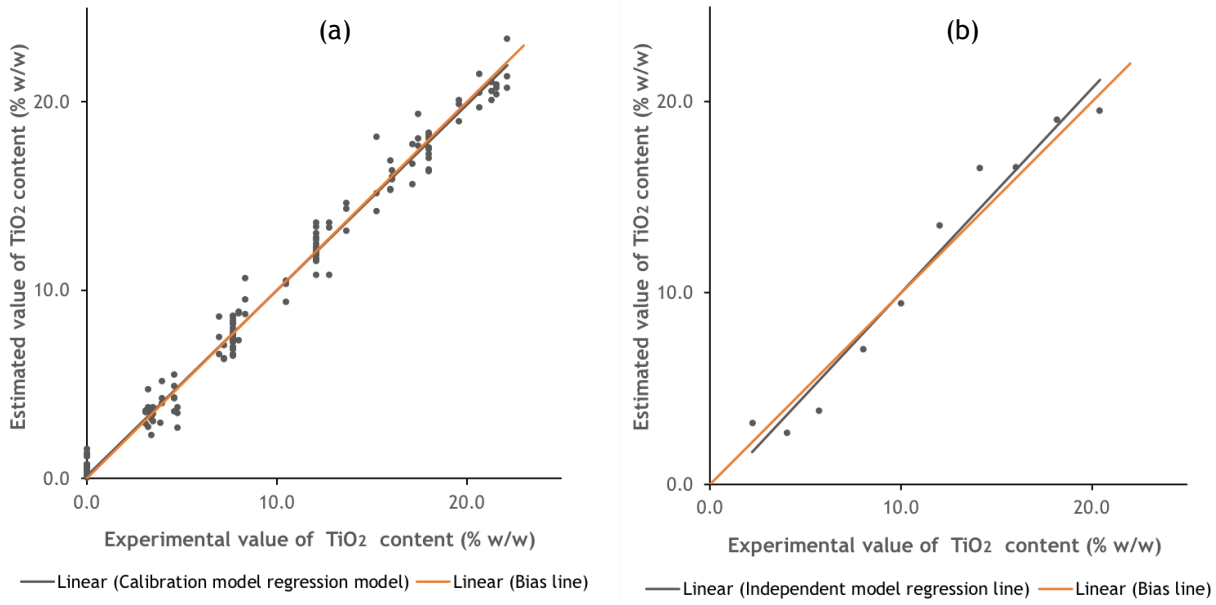

Figure 12. Estimated values as a function of experimental values of TiO_2 content attained from "AcrylTiO_2Opt1.2" calibration model (a) and independent dataset test (b).

4.3 Application of the best models to analyze the composition of coatings

Two sets of coatings (five and three containing styrene-acrylics and acrylics resin, respectively) were analyzed by FTIR-ATR and predictions models were applied. Since there was no information about the composition of the tested coatings, a deeper studied of the paint system were performed by comparison of the predicted contents with the contents gathered from the experimental analysis performed as described in Section 3.5. The results attained are represented in Table 22.

Table 22. Results obtained from the experimental determinations and PLS models for coatings

	Content (%, w/w)							
	Experimental results				Prediction results			
Coatings	Org. Comp.	$CaCO_3$	Water	Ashes	Polymer	$CaCO_3$	Water	TiO_2
Acryl #1	22.52 ± 0.76	16.13 ± 0.40	43.28 ± 0.70	33.96 ± 0.28	15.35	17.81	38.72	18.10
Acryl #2	16.34 ± 0.18	11.73 ± 1.13	47.54 ± 0.03	31.13 ± 0.65	16.48	15.95	47.91	15.74
Acryl #3	20.83 ± 0.27	13.23 ± 0.73	47.68 ± 0.20	26.52 ± 0.44	15.74	11.77	46.62	15.83
Sty #1	6.43 ± 0.07	38.03 ± 1.28	37.03 ± 0.35	39.91 ± 0.64	3.35	48.32	31.78	9.77
Sty #2	9.26 ± 0.05	29.13 ± 1.54	36.44 ± 0.26	41.51 ± 0.82	6.91	26.54	23.68	9.88
Sty #3	7.66 ± 0.02	25.27 ± 0.99	40.85 ± 0.06	40.43 ± 0.47	2.24	32.73	26.24	5.45
Sty #4	10.22 ± 0.05	25.23 ± 1.21	34.90 ± 0.05	44.04 ± 0.42	7.50	18.10	27.16	4.47
Sty #5	9.03 ± 0.01	5.55 ± 0.73	52.03 ± 0.18	40.26 ± 0.40	10.94	9.41	52.04	11.07

Water content

The constituent in question is the simplest to verify and acquire definitive results. Results obtained for the acrylic's paints are very satisfying, having the most considerable value of relative errors of only 10 %, w/w (Acryl #1). As for the styrene-acrylic's coatings, conflicting results were attained, only one result showed high similarity (Sty #5), while all others were significantly underestimated.

Polymer content

The majority organic part of a coating's composition consists of binders, organic solvents, and additives. Since the analyzed paints were all aqueous dispersions, organic solvents are used only as additives (e.g., as a coalescent agent). Therefore, the organic part of the type of coatings in the study can be considered as formed of just binders and additives.

Since the typical content of additives present in coating varies up till 5 %, w/w, it can be expected the maximum organic compound content to be the sum of polymer's content and the maximum content of additives frequently used. In some coatings (Acryl #1, Acryl #3, Sty #3), there is a slightly more significant organic compound value than the maximum expected. This

can mean that either more additives were used in these coatings or that their polymer's content predictions were slightly overestimated.

Calcium carbonate content

The results attained for the prediction of the content of calcium carbonates presented a significant discrepancy between one another. However, lower relative errors were associated with the contents' values of the acrylic coatings. For this component, the best prediction model for styrene-acrylics coatings was judged as possessing overestimation bias, which can be confirmed in three of the analysis (Sty #1, Sty #3, Sty #5) performed in this type of paint. An observation can be made when comparing the paints Sty #3 and Sty #4. The results gathered from the experimental test showed very similar content of, 25.27 and 25.23, respectively. However, predicted results were overestimated for the first, 32.73, and underestimated for the later,18.10. By evaluating the unexpected results, was assumed that one of the coatings presented a component which interfered with the spectrum interpretation related to the band of the CaCO3 characteristic bonds. Since the model's bias was considered overestimated, the assumption would be that the paint with the unknown interferent was the Sty #4.

TiO_2 content

By using the experimental test, only a mere comparison was able to be made. The ashes are formed of the inorganic components of the coatings which are mainly consisted of the inorganic parts of the extenders and inorganic pigments compounds, one of them being the TiO_2.

The information derived from the value of ashes content combined with the results from calcium carbonate determination and the TiO_2 content prediction allows a better understanding of the presence of other types of extenders which might have been used in the coatings formulation.

$$Ashes\ (\%, w/w) = \ CaO\ (\%, w/w) + TiO_2\ (\%, w/w) + Other\ extenders\ and\ pigments\ (\%, w/w)$$

where the calcium oxide (CaO) present in coatings is the product of thermal decomposition of the calcium carbonate of the coating’s formulations. With that, by using the attained values for ashes and $CaCO_3$ content determined by the experimental test and the predicted content values for TiO_2 given by the PLS models, an approximate value of the content of other extenders and inorganic pigments present in the coatings can be estimated, Table 23.

The results point to a meaningful presence of other types of extenders and pigments in the analyzed composition of styrene-acrylic's coatings. It is also possible to observe a significant difference in the quantity of these other constituents between the two types of coatings studied. It is clear that a more substantial part of the styrene-acrylic coatings' composition

consists of extenders when compared to the part related to pigments, however, the opposite was noticed when analyzing the results gathered from acrylic coatings' constitution.

Table 23. Estimated computed values for different extenders in the tested coatings

Coatings	Content (%, w/w)			
	Ashes	CaO *	TiO_2 **	Other extenders and inorganic pigments
Acryl #1	33.96	9.04	18.10	6.82
Acryl #2	31.13	6.57	15.74	8.82
Acryl #3	26.52	7.41	15.83	3.28
Sty #1	39.91	21.31	9.77	8.83
Sty #2	41.51	16.32	9.88	15.31
Sty #3	40.43	14.16	5.45	20.82
Sty #4	44.04	14.14	4.47	25.43
Sty #5	40.26	3.11	11.07	26.08

* Value of CaO calculated from decomposition of $CaCO_3$, present in the ashes ** Value estimated by the PLS model

4.4 Exemplification of PLS model's applicability in process control

The key finalities of the work accomplished were to assist in maintaining process control throughout the manufactures and to assist in the first stage of troubleshooting. This later is a systematic approach to locate and solve a one-time problem. Its first step consists of determining what changed in the production process, which may have caused the problem.

By performing an FTIR-ATR analysis of a defective good followed by the use of PLS models' information regarding its composition is promptly provided. This strategy can, in a matter of minutes, suggest the source of the manufactured failure, since the error in a constituent content indicates, either a problem is related to the raw material or the occurrence of an error in a process step associated to that specific component.

An example of that was encountered when performing the independent dataset test of the TiO_2 model in styrene-acrylic's coatings previously described in section 4.2.1.

The set of paint mixtures, used to accomplish the performance evaluation, were created using a dispersion of the TiO_2 powder. The prediction values attained from the model were considerably underestimated, as can be observed in Figure 9-b, and suggest the occurrence of a systematic error. With that, a few considerations for why the error occurred were made.

The dispersion used was already created, and the content of TiO_2 was assumed to be correct when calculations for the paints' contents were made. Therefore, if the production of the raw material were defective, it would directly affect the later produced coatings. Another possibility would be the incorrect homogenization of the dispersion before been added to the

coatings. Once TiO_2 tends to sediment, the formulated paints would possess a smaller content of this constituent in their composition, assuming the dispersion was taken from an upper level of the recipient where lowers contents of the component were present.

Similar incidents are frequent in a larger scale production when failure of equipment or even human error occurs. By applying the developed tool to analyze the final product, the abnormalities were easily located. In the case, if the samples were derived from commercialized paint manufacture, the conclusion drawn would be related to the TiO_2, simplifying the search for the origin of the failure. Since it concerns a systematic error and not a one-time error, further analyses should proceed, and different batches investigated in order to attain a history of the productions.

Also, by performing the independent dataset test, was confirmed the ability of the models to detect one-time errors. After conducting the validation test in acrylic's coatings of the $CaCO_3$ model and when analyzing the plot data (Figure 11-b), two outliers were detected.

The information presented can be interpreted, as being an indicator of oversights when preparing these two samples, once the predicted value is almost double as the original supposed content. Since $CaCO_3$ is also used in dispersions and its sedimentation occurs fast, periodic homogenization is necessary. Once the amounts of $CaCO_3$ estimated are much higher, a possible explanation could be the raw material dispersion recipient was not full, and when used before, the mixture was not well homogenized, and a part of it was removed, which provoked a change of the concentration of $CaCO_3$ in the remaining dispersion.

These situations were excellent examples of how the work in development can be used in industrials facility to assist in not only the detection of one-time error but also systematic manufacture errors, initiate process control troubleshooting and after normalization of the production process, perform periodic analyses in order to verify the quality of the finished and therefore the solution of the problem.

5 Conclusion

The work was able to demonstrate the applicability of FTIR-ATR spectroscopy in conjunction with multivariate analyses, more particularly the PLS method, to quantify the majority composition of acrylics' and styrene-acrylics' coatings, namely the polymer, calcium carbonate, titanium dioxide, and water's content.

In order to create the PLS models, two databases, one for each type of resin, were constructed, containing all Barbot's produced coatings. Besides, with the intent of improving the predictability capacity of the models, reformulated coating's spectra were introduced to those databases. A total of 59 and 128 samples of acrylic's and styrene-acrylic's coatings were analyzed, generating databases containing a total of 177 and 384 IR spectra, respectively.

The eight best models selected presented overall very good prediction results in both the independent dataset test and in the final analysis of the composition of coatings (Section 4.4).

Most of the selected models, 75 %, were optimized versions and performed only baseline correction as a pre-processing sample treatment, which indicates them as the best approach. Regarding the calcium carbonate's models, the use of a database without the spectra of varnishes (no presence of $CaCO_3$) was proved to generate models with better results.

Further studies should consider the construction of models using databases created with solid samples' spectra. This strategy would allow better prediction results for the solid constituents present in paints since the water bands would not appear in the spectrum, and better interpretation of characteristics bands of the solids components would be possible. Another recommendation would be creating models with the medium spectrum obtained from the triplicates. This tactic would minimize residual variance and therefore, errors related to the predictions' results. Furthermore, taking into consideration the two previous suggestions, the development of PLS models for the quantification of other commonly used extenders, namely talc, barium sulfate, silicates, and kaolin, would be of interest, since in most coatings they also represent a significant amount of their total weight.

The main applications of the developed work could be to detect the origin of a manufacturing error, by correlating composition's defects of flawed products with a particular step of the production process and also perform routine analysis to maintain process control throughout manufactures. Since composition's analysis can be made in a matter of minutes and the results attained from the PLS models were proven reliable, the work developed can be a valuable tool to be implemented in an industrial unit.

6 Assessment of the work done

6.1 Objectives Achieved

Complete sets of databases were created for the two main types of coatings – acrylics and styrene-acrylics – produced at Barbot.

PLS models were developed, with the use of the databases, to predict the contents of the four main components present in coatings: polymer, calcium carbonate ($CaCO_3$), titanium dioxide (TiO_2) and water. For each resin type, after the validation test, the models with best predicting abilities were selected, one for each component.

6.2 Other Work Carried Out

Performance of quality control on samples of all daily productions was conducted. It consisted of verifying if the most critical properties of the coatings, mainly viscosity, density, opacity, glow, and pH, were contained within the required range established internally.

FTIR analyses of raw materials samples were carried out either to confirm their compositions or, with the use of a tool present in the Spectrum Software called compare. With the later, similarity analyses were conducted in order to find similar ones present at Barbot's raw material database.

Spectrum analysis of paints related to customers complaints were also completed, using not only the Spectrum software tools but also the PLS models in development in the Quant software, to detect possible composition errors.

Samples derived from residues which led to equipment failure were also analyzed in the FTIR-ATR, so composition was discovered.

6.3 Limitations and Future Work

Paint media is an extremely complex mixture. Therefore, spectrum data has difficulties when it comes to identifying some components present on it, for example, additives, owing to the existence of overlapping and very low-intensity bands (since they are present in small quantities).

As for limitations regarding the PLS models, the presence of overlapping and low-intensity band can interfere with the accuracy of the main components' prediction. Another related complication is the presence of unexpected compounds in coatings which could lead to erroneous results, once their respective bands are not present in the database with which the model was created.

In order to use the developed prediction PLS models for the acquisition of information about coatings of unknown compositions, it necessary that the user possesses enough knowledge of the paint system and spectrum analyses to be able to draw right conclusions.

As for future work, development of PLS models using spectra acquired for solid samples could be a way of obtaining more accurate predictions' results. This can be explained by the high intensity of bands related to water molecules' bonds, which usually causes overlapping of bands leading to possible misinterpretation of spectra information and latter incorrect predictions. It is also advisable to create databases using medium spectrum only, once it was observed during the development of models an increase of prediction error associated with the use of triplicates spectra.

Further studies related to other frequently used extenders, for example, talc, and their main spectrum bands could provide enough information for the development of PLS models for predicting their content’s. Also, models should be built for prediction of components in vinylics’ coatings, once it is the third main type of coating produced at Barbot.

6.4 Final Assessment

The applicability of the PLS models as a process control tool can be expected to present outstanding outcomes, once it can generate fast and considerably accurate prediction results concerning the final product's composition.

The performed work can also be used with other finalities, for example, the models can be used to determine the defective composition of paints related to customer complaints, simplifying the search for the error while producing quick definite results.

7 References

Abdi H (2003) Partial least squares (PLS) regression. In: Lews-Beck M, Bryman A, Futing T (eds) Encyclopedia of Social Sciences Research Methods. Thousand Oaks. Sage. pp 792–795.

Alua, P. M. (2012) 'Opacity optimization of waterborne paints', Masters Thesis, Universidade de Lisboa. Lisboa. Portugal.

Amir, R. M., Anjum, F. M., Khan, M. I., Pasha, I., Nadeem, M. (2013) 'Application of Fourier Transform Infrared (FTIR) Spectroscopy for the identification of wheat varieties', Journal of Food Science and Technology, 50(5), pp. 1018–1023.

Ausili, A., Sanchez, M. and Gomez-Fernandez, J. C. (2015) 'Attenuated total reflectance infrared spectroscopy: A powerful method for the simultaneous study of structure and spatial orientation of lipids and membrane proteins', *Biomedical Spectroscopy and Imaging*, 4(2), pp. 159–170.

Beukelman, T. and Brunner, H. I. (2016) 'Trial Design, Measurement, and Analysis of Clinical Investigations', in *Textbook of Pediatric Rheumatology*. 7th Edition. Elsevier, pp. 54-77.e2.

Bieleman, J. (2000) *Additives for Coatings*. Wiley. New York.

Bobrova, A. M., *Zhigun, .I G., Bragina, M. I., Fotiev, A. A.* (1968) 'Infrared absorption spectra of various titanium compounds', *Journal of Applied Spectroscopy*, 8(1), pp. 59–63.

Bohm, K., Smidt, E. and Tintner, J. (2017) 'Application of Multivariate Data Analyses in Waste Management', in Beata Akselsen (ed.) *Multivariate Analysis in Management, Engineering and the Sciences*. Scitus Academics LLC. Valley Cottage.

Bunaciu, A. A., Aboul-Enein, H. Y. and Fleschin, S. (2010) 'Application of Fourier Transform Infrared Spectrophotometry in Pharmaceutical Drugs Analysis', Applied Spectroscopy Reviews, 45(3), pp. 206–219.

Camo Analytics (2019) *Chemometrics*. Available at: https://www.camo.com/chemometrics/ (Accessed: 10 June 2019).

Chalmers, J. M. and Everall, N. J. (1999) 'Polymer Analysis and Characterization by FTIR, FTIRMicroscopy, Raman Spectroscopy and Chemometrics', International Journal of Polymer Analysis and Characterization, 5(3), pp. 223–245.

Colombini, M. P., Modugno, F., Giannarelli, S., Fuoco, R., Matteini, M. (2000) 'GC-MS characterization of paint varnishes', *Microchemical Journal*, 67(1–3), pp. 385–396.

Colombini, M. P., Andreotti A., Bonaduce I., Modugno F., Ribechini E. (2010) 'Analytical strategies for characterizing organic paint media using gas chromatography/mass spectrometry', *Accounts of Chemical Research*, 43(6), pp. 715-727.

CSD Engineers, I. (2016) *Paint and Varnishes Industry*. UNIDO, Environmental Management Branch. Vienna.

Faia, C. (2018) *Determinação de componentes de tintas através de FTIR*. Dissertação de

Mestrado. Universidade do Porto. Porto. Portugal.

Grimnes, S. and Martinsen, Ø. G. (2015) 'Data and Models', in Academic Press (ed.) *Bioimpedance and Bioelectricity Basics*. 3rd edition. Elsevier. San Diego. pp. 329–404.

Gupta, H. V., Sorooshian, S. and Yapo, P. O. (1999) 'Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration', *Journal of Hydrologic Engineering*, 4(2), pp. 135–143.

Gürses, A., Açıkyıldız, M., Güneş, K., Gürses, M. S. (2016) 'Their structure and properties', in *Dyes and pigments*. 1st edition. Springer International Publishing, pp. 13–29.

Hayes, P. A., Vahur, S., Leito, I. (2014) 'ATR-FTIR spectroscopy and quantitative multivariate analysis of paints and coating materials', *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. Elsevier, 133, pp. 207–213

Jolliffe, I. T. (2002) *Principal Component Analysis*. 2nd edition. Springer-Verlag. New York.

Jozsef, B. and Blaga, P. (2014) 'Production Quality Control in the Process of Coating in an Electrostatic Field', *Procedia Technology*, 12, pp. 476–482.

Kunal Roy, S. K. and R. N. Das (2015) *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. 1st edition. Elsevier. San Diego.

LibreTexts (2019) 'How an FTIR spectrometer operates', pp. 1–9. Available at: https://chem.libretexts.org/Textbook_Maps/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy/How_an_FTIR_Spectrometer_Operates.

Marengo, E., Liparota, M. C., Robotti, E., Bobba, M. (2005) 'Multivariate calibration applied to the field of cultural heritage: Analysis of the pigments on the surface of a painting', *Analytica Chimica Acta*, 553(1–2), pp. 111–122.

Moriasi, D. N., Arnold J. G., Van Liew M. W., Bingner R. L., Harmel R. D., Veith T. L., (2007) 'Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations', *Transactions of the ASABE*, 50(3), pp. 885–900.

Oliveira, R., Araújo, R., Barros, F., Segundo, A., Zampolo, R.,Fonseca, W., Dmitriev, V. Brasil, F. (2017) 'A system based on artificial neural networks for automatic classification of hydro-generator stator windings partial discharges', *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, 16(3), pp. 628–645.

Olkin, I. and Sampson, A. R. (2001) 'Multivariate Analysis: Overview', in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 10240–10247.

PerkinElmer (2005) 'Technical note: FT-IR Spectroscopy Attenuated Total Reflectance (ATR)', PerkinElmer Life and Analytical Sciences, p. 1. http://www.utsc.utoronto.ca/~traceslab/ATR_FTIR.pdf (Accessed: 30 march 2019).

PerkinElmer (2014) SPECTRUM QUANT - User's Guide. PerkinElmer Ltd. United Kigdom.

Peter, S. C., Dhanjal, J. K., Malik, V., Radhakrishnan, N., Jayakanthan, M., Sundar, D. (2018)

Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications, *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier.

Qian, F., Wu, Y. and Hao, P. (2017) 'A fully automated algorithm of baseline correction based on wavelet feature points and segment interpolation', Optics and Laser Technology, 96, pp. 202–207.

Rinnan, Å., Berg, F. van den and Engelsen, S. B. (2009) 'Review of the most common pre-processing techniques for near-infrared spectra', *TrAC - Trends in Analytical Chemistry*, 28(10), pp. 1201–1222.

Roussel, S., Preys, S., Chauchard, F., Lallemand, J. (2014) 'Multivariate Data Analysis (Chemometrics)', in *Process Analytical Technology for the Food Industry*. Springer Science. New York. pp. 15–18.

Sánchez Rojas, F. and Cano Pavón, J. M. (2005) 'Spectrophotometry - Biochemical Applications', *Encyclopedia of Analytical Science*, (April), pp. 366–372.

Schulz, H., Schrader, B., Quilitzsch, R., Steuer, B. (2002) 'Quantitative analysis of various citrus oils by ATR/FT-IR and NIR-FT Raman spectroscopy', *Applied Spectroscopy*, 56(1), pp. 117–124.

Sciutto, G., *Oliveri, P., Prati, S., Catelli, E., Bonacini, I., Mazzeo, R.* (2017) 'A Multivariate Methodological Workflow for the Analysis of FTIR Chemical Mapping Applied on Historic Paint Stratigraphies', *International Journal of Analytical Chemistry 2017, Article ID 4938245, 12 pages, https://doi.org/10.1155/2017/4938145* .

Sjo, H. I. (2018) *Using Multivariate Data Analysis and ATR- FTIR Spectroscopy for Modeling Components Present During CO_2 Capture with Amines*. Masters Thesis. University of Bergen. Bergen. Norway.

Stoye, D. and Freitag, W. (1998) *Paints, coatings and solvents*. 2nd edition. Wiley-VCH Verlag GmbH & Co. KGaA. Weinheim.

Stoye, D., Marwald, B. and Plehn, W. (2010) 'Introduction', in *Paints and Coatings*. Wiley-VCH Verlag GmbH & Co. KGaA. Weinheim. pp. 1–28.

Stuart, B. (2004) 'Spectral analysis', in *Infrared Spectroscopy: Fundamentals and Applications*. 1st edition. John Wiley & Sons Inc. New York. pp. 45–70.

Thermo Scientific (2013a) 'Introduction to Fourier Transform Infrared Spectroscopy'. Thermo Scientifc. http://www.ftpx.com/ftpintro.aspx. (Accessed: 25 July 2019).

Thermo Scientific (2013b) 'Introduction To FT-IR Sample Handling', *Thermo Fisher Scientific Inc.*, pp. 1–8. https://tools.thermofisher.com/content/sfs/brochures/BR50557_E_IntroSampHand0613M_H_1.pdf. (Accessed: 25 July 2019).

Wold, S., Sjöström M., Eriksson L., (2001) 'PLS-regression: A basic tool of chemometrics.', *Chemometrics and Intelligent Laboratory Systems*, 58, pp. 2001–109.

Annex 1 IV Spectra Interpretation

Resins e respective monomers

*Table AA1.1. Characteristic peaks of main resin and their respective monomers in Infrared Spectroscopy (adapted from:*Faia, 2018)

		Wavenumber (cm^{-1})							
		=C-H (3200-3000)	C-H (3000-2800)	C=O (1800-1700)	CH_2 (1470-1450)	CH_3 (1415-1360)	C-O (1280-1200)	C-O-C (1200-1000)	=C-H (710-690)
Resins	Acrylic		2956 (w)	1727 (s)	1450 (m)	1385 (w)	1236 (m)	1143 (s)	
	Styren-Acrylic	3069 (w)	2958 (w)	1730 (s)	1453 (m)	1385 (w)	1270 (m)	1162 (s)	698 (vs)
	Vinyl-VeoVa		2959 (w)	1735 (s)	1434 (w)	1372 (m)	1235 (vs)	1022 (s)	
Monomers	Butyl acrylate		2962 (w)	1723 (s)	1464 (w)	1408 (m)	1273 (m)	1186 (s)	
	Styren	3083 (w)			1450 (w)	1413 (w)	1202 (w)	1083 (w)	695 (s)
	Vinyl acrylate	3095 (w)		1740 (s)		1371 (m)	1225 (s)	1133 (s)	710 (w)
	VeoVa	3040 (w)	2962 (w)	1735 (s)	1463 (w)	1384 (w)	1206 (w)	1015 (vs)	698 (w)

w- weak intensity peak; m -medium intensity peak; s-strong intensity peak; vs - very strong intensity

Solvents and diluents

Table AA1.2- Characteristic peaks of solvents in Infrared Spectroscopy (adapted from: Faia, 2018)

	Wavenumber (cm^{-1})							
	O-H (3400-3000)	C-H (3000-2800)	C=O (1800-1700)	CH_2 (1470-1450)	CH_3 (1380-1360)	C-O (1270-1200)	C-O (1200-1000)	
Esters		2961 (m)	1739 (vs)	1463 (w)	1366 (m)	1228 (vs)	1065 (m)	1031 (m)
Ketones		2980 (w)	1713 (s)	1458 (w)	1364 (m)	1240 (w)	1171 (m)	1086 (w)
Ethers		2974 (m)				1263 (w)	1200 (m)	1102 (s)
Alcohols	3325 (w)	2970 (m)/ 2884 (s)		1467 (w)	1378 (m)	1230 (w)	1160 (m)	1128 (s)

Table AA1.3- Characteristic peaks of diluents in Infrared Spectroscopy (adapted from: Faia, 2018)

	Wavenumber (cm^{-1})							
	C-H insat (3100-3000)	C-H sat (3000-2800)			C=C (1610-1460)			CH_3 (1380-1360)
H- aromatics	3027 (w)	2920 (w)			1604 (w)	1495 (m)	1460 (w)	1380 (w)
H- aliphatics	2957 (m)	2923 (m)	2856 (s)		1461 (m)			1378 (w)

w- weak intensity peak; m -medium intensity peak; s-strong intensity peak; vs - very strong intensity

Extenders

Annex AA1.4- Characteristic peaks of solvents in Infrared Spectroscopy (adapted from: Faia, 2018)

	Wavenumber (cm^{-1})									
	3240-3220	1640-1620	1422-1390	1200-1100	1100-1000	1000-900	900-800	800-700	700-600	600-400
Calcium Carbonate			1412 (s) C-O				875 (m) O-C-O	712 (w) O-C-O		
Talc						981 (s) Si-O			666 (m) Si-O/Si-MgO	411 (s) SiO/MgO
Kaolin					1074 (s) Si-O-Si	811 (w) Si-O-Si				450 (vs) Si-O-Si
Silica					1058 (s) Si-O-Si			778 (m) Si-O-Si		451 (vs) Si-O-Sl
Barium Sulfate				1189 (w) S-O	1062 (s) S-O	983 (m) S-O			637 (m) O-S-O 605 (s) O-S-O	
Titanium Dioxide	3236 (m) O-H	1635 (w) H-O-H								482 (s) O-Ti-O

w- weak intensity peak; m -medium intensity peak; s-strong intensity peak; vs - very strong intensity

Appendix 1 **Example of coating's spectrum**

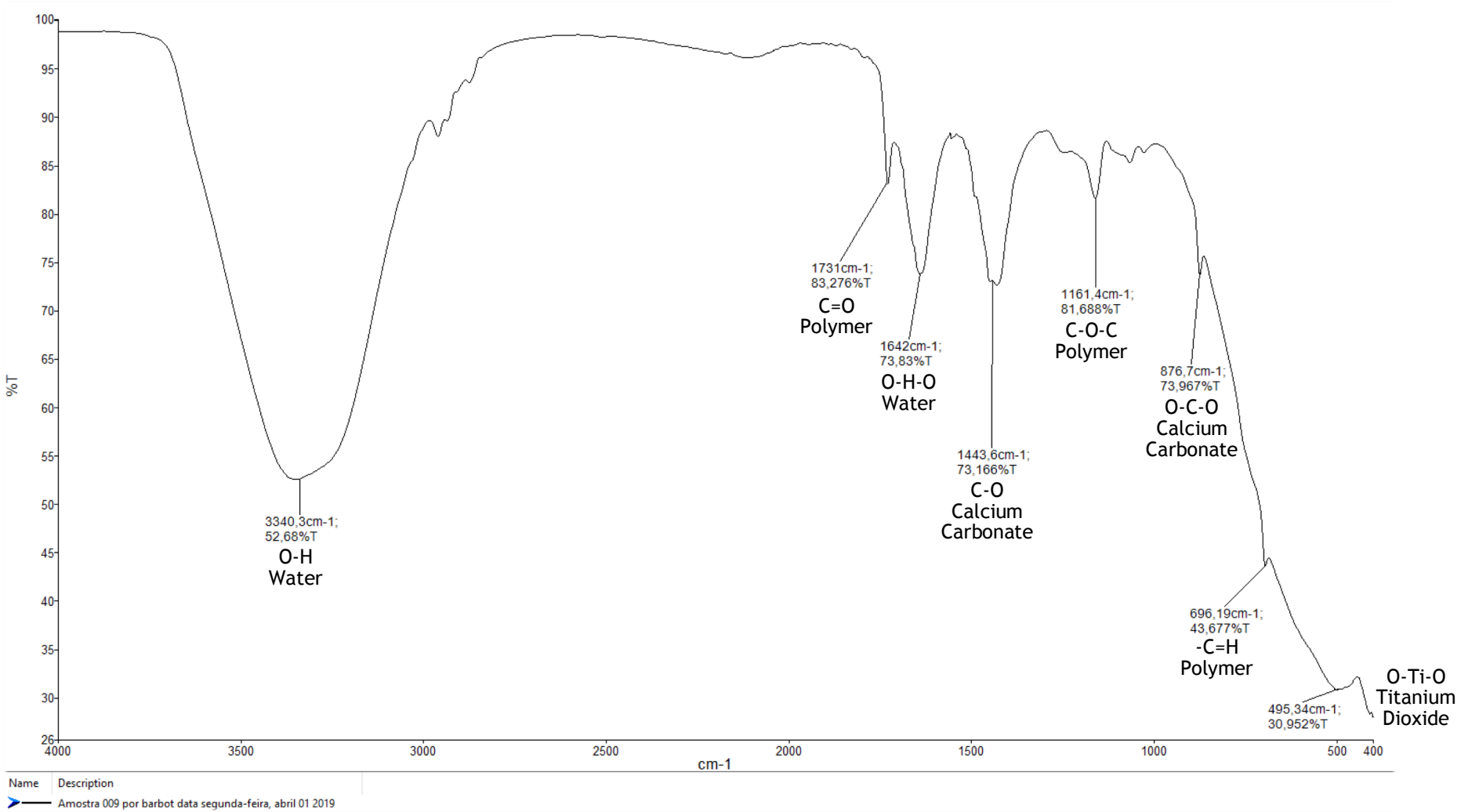

Figure A1.1. Example of resulting spectra obtained from styrene-acrylic coating's analysis.

Appendix 2 Data for selection of optimum number of latent variables for the different PLS models

Table A2.1. Cumulative variance values for StyCaCO$_3$1.4 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	29.7	86.8
2	34.1	90.4
3	42.2	93.9
4	44.7	95.8
5	47.1	96.6
6	49.1	97.6
7	50.4	98.0
8	52.5	98.3

LV = Latente Variables, Cum = Cumulative, Var = Variance

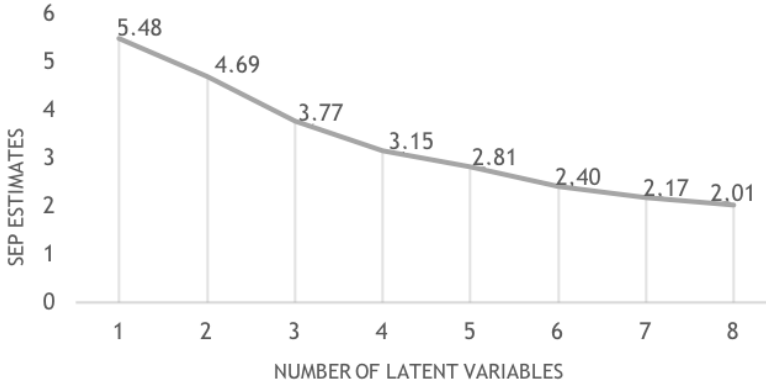

Figure A2.1. SEP estimates as a function of Latent variables for StyCaCO$_3$1.4 model.

Table A2.2. Cumulative variance values for StyWtrOpt1.1 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	13.0	34.8
2	30.2	64.9
3	38.6	78.7
4	42.2	85.8
5	44.4	90.1
6	46.0	92.2
7	48.2	93.8
8	49.9	94.7

LV = Latente Variables, Cum = Cumulative, Var = Variance

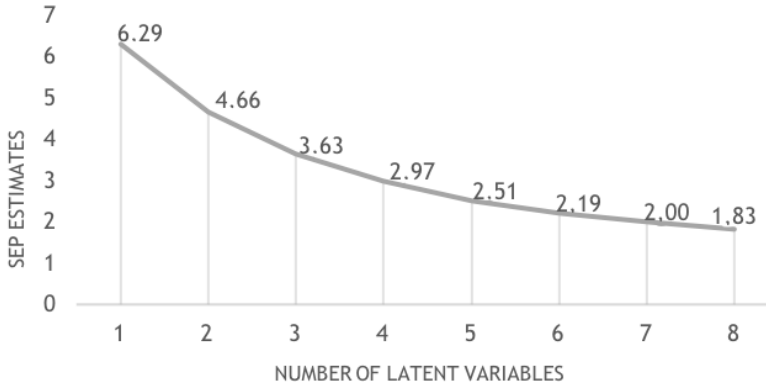

Figure A2.2. SEP estimates as a function of Latent variables for StyWtrOpt1.1 model.

Table A2.3. Cumulative variance values for StyTiO$_2$Opt2.1 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	23.7	58.7
2	30.2	74.9
3	37.9	87.0
4	40.5	93.3
5	42.0	95.6
6	43.3	96.5
7	46.5	96.9
8	48.7	97.4

LV = Latente Variables, Cum = Cumulative, Var = Variance

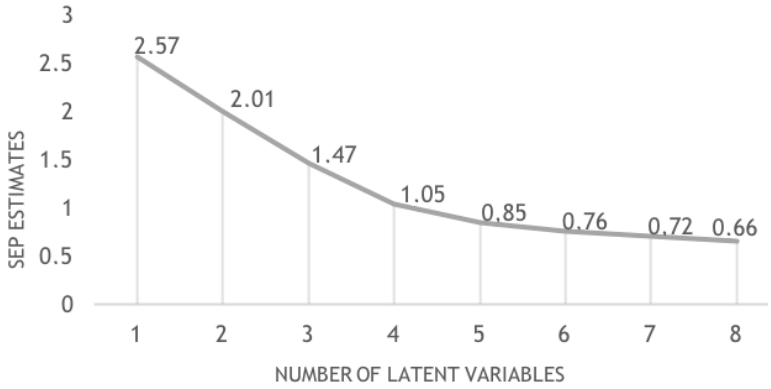

Figure A2.3. SEP estimates as a function of Latent variables for StyTiO$_2$Opt2.1 model.

Table A2.4. Cumulative variance values for AcrylPol1.1 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	17.8	85.1
2	26.6	92.4
3	36.8	94.4
4	41.4	96.3
5	45.6	97.5
6	48.0	98.1
7	51.8	98.6
8	55.2	98.9

LV = Latente Variables, Cum = Cumulative, Var = Variance

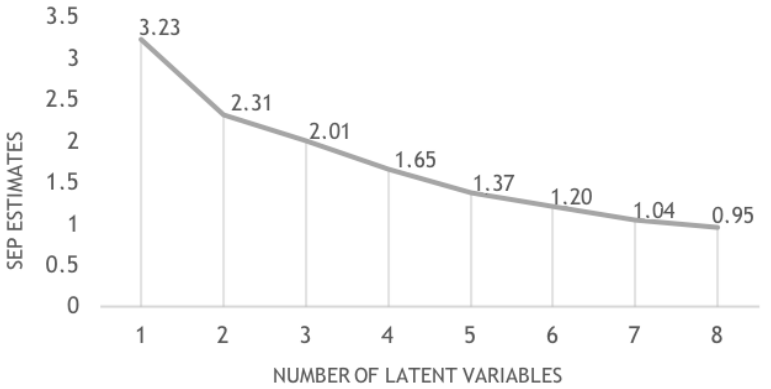

Figure A2.4. SEP estimates as a function of Latent variables for AcrylPol1.1 model.

Table A2.5. Cumulative variance values for AcrylCaCO$_3$Opt1.3 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	13.2	65.5
2	24.7	82.3
3	27.2	85.9
4	33.7	90.3
5	41.6	91.1
6	47.6	93.5
7	50.9	95.2
8	53.5	96.4

LV = Latente Variables, Cum = Cumulative, Var = Variance

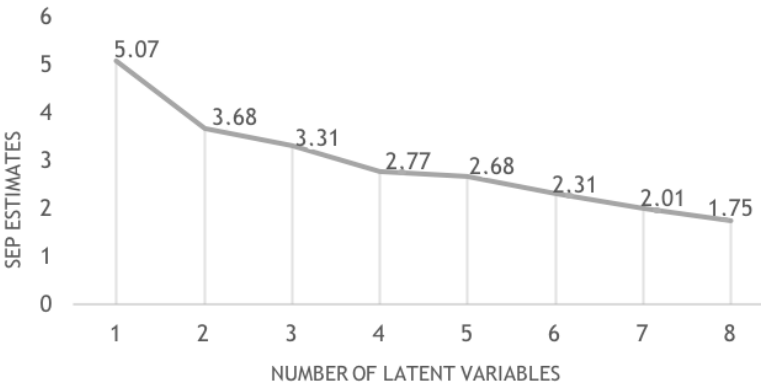

Figure A2.5. SEP estimates as a function of Latent variables for AcrylCaCO$_3$Opt1.3 model.

Table A2.6. Cumulative variance values for AcrylWtrOpt1.2 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	11.8	63.6
2	20.3	75.0
3	28.6	79.9
4	35.8	86.0
5	43.2	87.8
6	50.5	90.1
7	53.5	92.8
8	55.2	94.4

LV = Latente Variables, Cum = Cumulative, Var = Variance

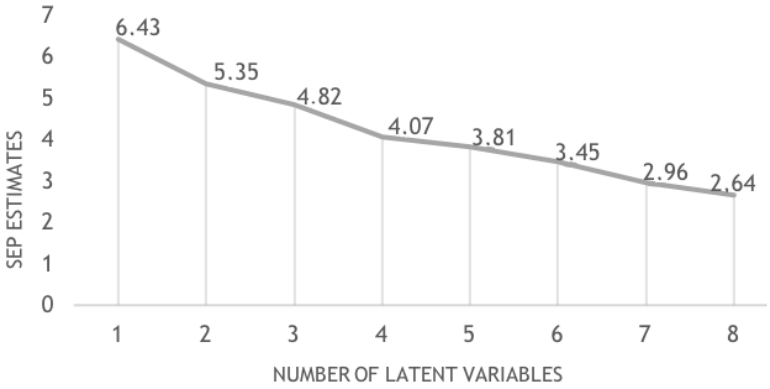

Figure A2.6. SEP estimates as a function of Latent variables for AcrylWtrOpt1.2 model.

Table A2.7. Cumulative variance values for AcrylTiO_2Opt1.2 model for the different number of latent variables

LV	Cum X-Var (%)	Cum Y-Var (%)
1	13.7	88.9
2	21.1	93.1
3	32.1	95.3
4	37.5	96.5
5	45.0	97.4
6	51.1	98.0
7	54.1	98.6
8	57.3	98.9

LV = Latente Variables, Cum = Cumulative, Var = Variance

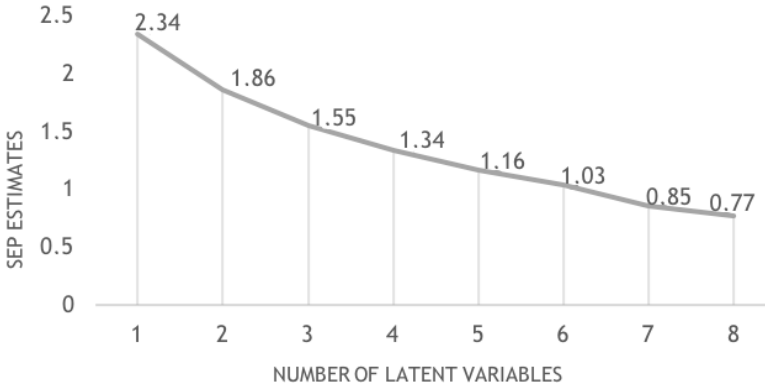

Figure A2.7. SEP estimates as a function of Latent variables for AcrylTiO_2Opt1.2 model.