DEVELOPMENT OF CHEMOMETRIC CALIBRATION TOOLBOX AND ITS APPLICATION FOR DETERMINATION OF SALEP ADULTERATION

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Chemistry

by
Gün Deniz AKKOÇ

December 2018
İZMİR

We approve the thesis of **Gün Deniz AKKOÇ**

Examining Committee Members:

Prof. Dr. Durmuş ÖZDEMİR
Department of Chemistry, İzmir Institute of Technology

Prof. Dr. Figen TOKATLI
Department of Food Engineering, İzmir Institute of Technology

Assoc. Prof. Dr. Mehmet Fatih CENGİZ
Department of Agricultural Bioengineering, Akdeniz University

28 December 2018

Prof. Dr. Durmuş ÖZDEMİR
Supervisor, Department of Chemistry
İzmir Institute of Technology

Prof. Dr. Ahmet Emin EROĞLU
Head of the Department of
Chemistry

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Prof. Dr. Durmuş ÖZDEMİR for introducing me to chemometrics, for his guidance and support during my M.Sc. studies.

Additionally, I would like to thank my thesis committee members Assoc. Prof. Dr. Mehmet Fatih CENGİZ for providing me the opportunity to work in the TÜBİTAK project led by him in a very systematic and detailed manner and Prof. Dr. Figen TOKATLI for her valuable inputs.

I am very grateful to my mother Nilgün AKKOÇ and to my father Salih AKKOÇ for their unconditional love. More importantly, they have the greatest influence on my passion for learning and discovering. I am very thankful to my aunt İkbal AKKOÇ as she helped me achieve my best by her endless support and efforts.

I would like express my deepest gratitude and respect to Ayşe Neslihan ŞANBEY and Hasan Tahsin ŞANBEY for their big hearts. Their help allowed me to focus on what I have been passionate about and saved me from a lot struggle. I am also very grateful to Ayşe Coşkun and her friends for all their contributions during my education.

I am very thankful to Ayten Ekin MEŞE as she was the one keeping me sane and helping me make the right decisions. She was always there for me to pick me up when I was down.

I want to thank Gokhan ÖZTARHAN for all valuable discussions. I also want to thank him and Caner MENGÜ for all the fun nights we had.

The case study in this work, that is determination of salep adulteration, was supported by the Scientific and Technological Research Council (TÜBİTAK), Grant No: 115O058.

ABSTRACT

DEVELOPMENT OF CHEMOMETRIC CALIBRATION TOOLBOX AND ITS APPLICATION FOR DETERMINATION OF SALEP ADULTERATION

A chemometric calibration toolbox, which contains Inverse Least Squares (ILS) regression, Principle Components Regression (PCR), Partial Least Squares Regression (PLSR), Genetic Inverse Least Squares (GILS) regression, and Ridge regression, was developed in MATLAB environment. During the development, multiple strategies to improve the calculation speed, namely vectorization and parallelization, were employed. Besides these programmatic strategies, efficient cross-validation (CV) procedures were implemented that are specifically tailored for parameter tuning of PCR and PLSR. For GILS, by constructing CV matrices in advance, the computational cost was further reduced. Additionally, a Graphical User Interface (GUI), which also includes baseline correction and variable range selection capabilities, was developed. For increased convenience, regardless of the chosen model, the toolbox returns a single vector of regression coefficients that accounts for centering and scaling of variables along with variable selection.

Using the developed toolbox, quantitative determination of salep adulteration was carried out through chemometric calibration methods on Mid-IR data obtained from FTIR-ATR which is a fast and easy-to-use spectroscopic instrument. The main motivation was the lack of an established method for determination of adulteration of salep which can be quite common due to very high price of pure salep, despite the strict legal regulations. Using 365 samples covering a wide range of adulteration scenarios with 20 adulterants, calibration models were obtained and evaluated. Ensemble model, obtained by averaging GILS and Ridge, yielded the best RMSEP of 6.82 (w/w %). To cope with the unspecific adulterant problem, SIMCA was employed to provide an qualitative insight about the presence of such compounds.

ÖZET

KEMOMETRİK KALİBRASYON YAZILIM PAKETİ GELİŞTİRİLMESİ VE SALEP TAĞŞİŞİNİN BELİRLENMESİNDE KULLANILMASI

Ters En Küçük Kareler (ILS) Regresyonu, Temel Bileşen Regresyonu (PCR), Kısmi En Küçük Kareler Regresyonu (PLSR), Genetik Ters En Küçük Kareler (GILS) Regresyonu ve Ridge Regresyonu uygulamalarına olanak sağlayan bir kemometrik kalibrasyon paketi MATLAB üzerinde geliştirildi. Geliştirme aşamasında yazılımın hızlandırılması amacıyla vektörizasyon ve paralelizasyon yöntemleriden faydalanıldı. Bunun yanı sıra, PCR ve PLS metotlarında parametre optimizasyonu için verimli hesaplama prosedürleri kullanıldı. GILS metodunda ise CV matrislerinin tek sefer hesaplanması yoluyla hız konusunda iyileştirmeler gerçekleştirildi. Ek olarak geliştirilen Grafiksel Kullanıcı Arayüzü (GUI) üzerinden hem zemin kayması düzeltmesi ve değişken çıkarma hem de kolay bir kullanıcı deneyimi hedeflendi. Ortalama, normalizasyon ve değişken çıkarımının yanı sıra, kullanılan kalibrasyon tekniğinden de bağımsız olarak regresyon katsayılarının tüm bu işlemleri kapsayacak ve boyutu girdi boyutuna eşit tek bir vektör ile ifadesi sağlandı.

Geliştirilen bu yazılım paketi vasıtasıyla, hızı ve kolay kullanımı ile bilinen FTIR-ATR spektrometresi kullanılarak elde edilmiş Mid-IR verileri üzerinden, salepte sahteciliğin kantitatif olarak belirlenmesine yönelik kemometrik kalibrasyon çalışmaları gerçekleştirildi. Bu çalışmanın temel motivasyonu, yasal düzenlemelere rağmen yüksek fiyatı nedeniyle salebin sahteciliğe maruz kalma potansiyelinin yüksekliği ve bu konuda halihazırda bir metodolojinin bulunmamasıdır. Bu amaçla, 20 adet tağşiş malzemesini farklı oranlarda bulunduran 365 örnek hazırlanarak olabildiğince çok tağşiş senaryosunun kapsanması hedeflenmiştir. Bu örneklerin Mid-IR spektrumları ile çalışılan kalibrasyon metotları arasında, GILS ve Ridge regresyonlarının ortalaması ile elde edilen kombine model 6.82 (w/w %) RMSEP başarımı ile en uygun model olarak belirlendi. Modele dahil edilmemiş tağşiş malzemelerinin varlığının belirlenmesinde ise SIMCA yönteminin kalitatif bir önbilgi için kullanılabileceği gösterildi.

TABLE OF CONTENTS

LIST OF FIGURES ix

LIST OF TABLES xi

CHAPTER 1. INTRODUCTION 1
1.1. Structure and Scope of the Thesis 4
1.2. Literature Review 4

CHAPTER 2. CALIBRATION METHODS 6
2.1. Beer-Lambert's Law 6
2.2. Univariate Calibration 7
2.3. Multivariate Calibration 10
2.3.1. Overfitting, Underfitting and Validation 11
2.3.2. Least Squares 13
2.3.3. Classical Least Squares 14
2.3.4. Inverse Least Squares 16
2.3.5. Principle Components Regression 19
2.3.5.1. Choosing Number of PCs 22
2.3.6. Partial Least Squares 24
2.3.7. Ridge Regression 25
2.3.8. Genetic Inverse Least Squares 26
2.3.8.1. Selection of Initial Genes 27
2.3.8.2. Evaluation of the Population 28
2.3.8.3. Selection of Parents for Breeding 29
2.3.8.4. Crossover and Mutations 29
2.3.8.5. Replacing Parents with Offspring 30
2.3.8.6. Options and Algorithmic Procedure 31
2.3.8.7. Remarks 32
2.3.9. Ensemble Models 32

CHAPTER 3. DEVELOPMENT OF CHEMOMETRIC CALIBRATION TOOLBOX 33
3.1. Speed Concerns 34
3.1.1. Vectorization 34
3.1.2. Parallelization of Genetic Algorithm 35
3.1.3. Calculation of CV Matrices in Advance for Genetic Algorithm 35
3.1.4. Parallelization of Cross-Validation 36
3.1.5. Efficient Cross-Validation for PCR 37
3.1.6. Efficient Cross-Validation for PLSR 38
3.2. Preprocessing Techniques 38
3.2.1. Mean-centering and Scaling 38
3.2.2. Removing Variables 39
3.2.3. Baseline Correction 39
3.3. Providing Simple Models 41
3.3.1. Accounting For Mean-Centering and Scaling 41
3.3.2. Accounting For Unused Variables 43
3.3.3. Averaging GA Models 44
3.4. User-Friendliness 44
3.4.1. Graphical User Interface 44
3.4.2. Producing Graphs 46
3.5. Compatibility 46

CHAPTER 4. CASE STUDY: DETERMINATION OF SALEP ADULTERATION 48
4.1. Experimentation 48
4.1.1. Sample Collection and Sample Preparation 48
4.1.2. Instrumentation 49
4.2. Results and Discussion 49
4.2.1. FTIR Results 49
4.2.2. Chemometric Studies 50
4.2.2.1. Data Processing 51
4.2.2.2. ILS Regression Results 52
4.2.2.3. PCR Results 52
4.2.2.4. PLSR Results 53
4.2.2.5. Ridge Regression Results 55
4.2.2.6. GILS Regression Results 56

4.2.2.7. GILS+Ridge (Ensemble Model) Results 57
4.2.3. Summary and Comparison of the Calibration Models 59
4.2.3.1. Estimating Prediction Performance For Unknown Adulterants 60
4.2.3.2. Single Class Modeling of Pure Salep 62

CHAPTER 5. CONCLUSION 74

REFERENCES 75

APPENDIX A. GENERATED ADULTERATION SCENARIOS 78

LIST OF FIGURES

Figure **Page**

Figure 1.1. Working principle of ATR 3
Figure 2.1. A visual representation of roulette wheel used in GA 29
Figure 2.2. An illustration of cross-over of two selected genes 30
Figure 3.1. A screenshot of the GUI that is developed as a part of the toolbox 45
Figure 4.1. Mid-IR spectra of all samples 51
Figure 4.2. Spectra of pure salep and pure adulterant samples 52
Figure 4.3. Actual salep concentrations vs. ILS predicted salep concentrations 53
Figure 4.4. Actual salep concentrations vs. corresponding ILS prediction residuals . 54
Figure 4.5. Number of PCs vs. PRESS plot for selecting the optimal number of PCs 55
Figure 4.6. Actual salep concentrations vs. PCR predicted salep concentrations 56
Figure 4.7. Actual salep concentrations vs. corresponding PCR prediction residuals 57
Figure 4.8. Number of LVs vs. PRESS plot for selecting the optimal number of LVs 58
Figure 4.9. Actual salep concentrations vs. PLS predicted salep concentrations 59
Figure 4.10. Actual salep concentrations vs. corresponding PLS prediction residuals 60
Figure 4.11. λ vs. PRESS plot for selecting the optimal shrinkage parameter 61
Figure 4.12. Actual salep concentrations vs. Ridge predicted salep concentrations ... 62
Figure 4.13. Actual salep concentrations vs. corresponding Ridge prediction residuals 63
Figure 4.14. Actual salep concentrations vs. GILS predicted salep concentrations ... 64
Figure 4.15. Actual salep concentrations vs. corresponding GILS prediction residuals 65
Figure 4.16. Overlay of wavenumber selection frequency for GILS model and average spectra 66
Figure 4.17. Actual salep concentrations vs. GILS+Ridge predicted salep concentrations 67
Figure 4.18. Actual salep concentrations vs. corresponding GILS+Ridge prediction residuals 68

Figure 4.19. GILS, Ridge and GILS+Ridge predictions of 14 pure salep validation samples 69
Figure 4.20. Performance estimate of calibration techniques in the presence of an unknown adulterant 70
Figure 4.21. Mid-IR spectra of pure salep, SKA, VNL, TRC, and CMC samples 71
Figure 4.22. Scree plot for PCA model of pure salep samples 71
Figure 4.23. Distance to model vs. distance to group center plot obtained by SIMCA model of pure salep samples 72
Figure 4.24. A closer look at boundaries of SIMCA plot 73

LIST OF TABLES

Table **Page**

Table 4.1. Region and city information of acquired pure salep samples 49

Table 4.2. List of adulterants along with their abbreviations 50

Table 4.3. Performance evaluation of all calibration models 67

Table A.1. List of generated adulteration scenarios (w/w %) 79

CHAPTER 1

INTRODUCTION

The term chemometrics, which was first mentioned by Svante Wold in 70s, can be defined as extracting information from chemical data. This particular decade corresponds to first uses of computational methods for solving scientific problems as well as many interdisciplinary studies hence it was quite suitable for the birth of a such new area (Brereton 2007). The early applications were quite limited by the computational power and the chemical instruments at that time. The most common applications were as simple as determination of multiple compounds from UV-VIS spectra. Chemometrics draw a lot of attention after its success on quantitative modeling of NIR data by PLS, which was proven to be very useful in food chemistry. Since then, the enormous amount of (higher quality) data produced by new instruments, exponentially increasing computational power at lower costs and advances in theory fueled the development of chemometrics as a seperate and comprehensive discipline. Similar to many data-driven sciences, chemometrics contains many elements from other disciplines such as applied mathematics, statistics, and computer science. Today, chemometrics is widely used to deal with (mostly multivariate) calibration, classification, design of experiment and signal processing problems (Brereton 2003).

The classification tasks can be divided into two parts: supervised classification and unsupervised classification. In supervised classification, the aim of a chemometric model is to predict classes, which are defined at the stage of model building, of new samples. The applications range from determining botanic origins of honey samples using H-NMR (Schievano, Peggion, and Mammi 2009) to cancer diagnosis from serum samples through Mid-IR spectrum (Gazi et al. 2003). On the other hand, in unsupervised classification, no class information is provided during modeling step. It is often employed for determination of possible clustering and for outlier detection.

Rather then predictive or exploratory purposes, design of experiment methods aims to provide the relation between experimental conditions and the results. This relation is then can be used to optimize the conditions to shift the outcomes in desired way(s) such as maximizing yield and minimizing cost (Brereton 2003). Often, what varies among design of experiment methods is not how the relation is revealed but it is how the experiments are designed so that the relation can be properly captured and the experiment

conditions can be manipulated accordingly.

Signal processing can be considered as a complimentary strategy that is employed along with calibration or classification methods and does not strictly fall under the area of chemometrics. For spectral data, while smoothing and derivatives can be used to either increase data quality or to reveal more information, chemometrics based methods such as OPLS-DA exist for increased interpretability (Wold et al. 1998; Trygg and Wold 2002) even thought it offers no performance improvement.

Chemometric calibation methods refer to quantitatively relating chemical data and desired information. Beside the least-squares methods, PCA and particularly PLS was the first methods that were capable of revealing underlying structure of highly correlated data outputted by chemical instruments. Not only this allowed the modeling of many challenging data such as NIR spectra, but it also brought interpretability and performance improvements over traditional calibration techniques. Since this thesis study mainly focuses on calibration, detailed information about applications and methods are given seperately in Chapter 2.

For application of chemometric techniques, various software alternatives are available. The programming language R, provides probably the most comprehensive list of methods through various packages since it is development motivation was dealing with data-driven problems in the first place. Python, as more general purpose programming language, offers the popular scikit-learn that is capable of performing not only chemometric but also machine learning methods. On the commercial side, MATLAB offers toolboxes for performing only a few chemometric methods. While it is quite fast and very handy for algebraic operations, the price and the lack of flexibility of the included toolboxes can become limiting.

Fourier Transform Infrared Spectroscopy (FTIR), which takes advantage of Michelson Interferometer to collect absorbance values at different wavenumbers simultaneously, was great leap forward and granted faster and more accurate acquisition of IR spectra compared to the conventional IR spectrometers (Skoog, Holler, and Crouch 2017).

Another advancement was the invention of Attenuated Total Reflectance (ATR). Before ATR gained popularity, in order to obtain IR spectra of a liquid sample, the IR transparent cells were employed where the cleaning of the these cells having very small diameters were quite time consuming for multiple measurements. Dealing with solid samples were even more problematic as they must be mixed with IR-inactive material, most popularly KBr, to produce a ”KBr pellet” whose spectrum is then to be obtained. For preparation of the pellet not only external equipment and material were required, but

also any inconsistencies in this step would result in baseline shifts. ATR, nearly removes the need for sample preparation as it allows acquisition of spectrum from the sample's surface through an ATR crystal having a high refractive index. Furthermore, working principles of ATR is illusturated in Figure 1.1 below.

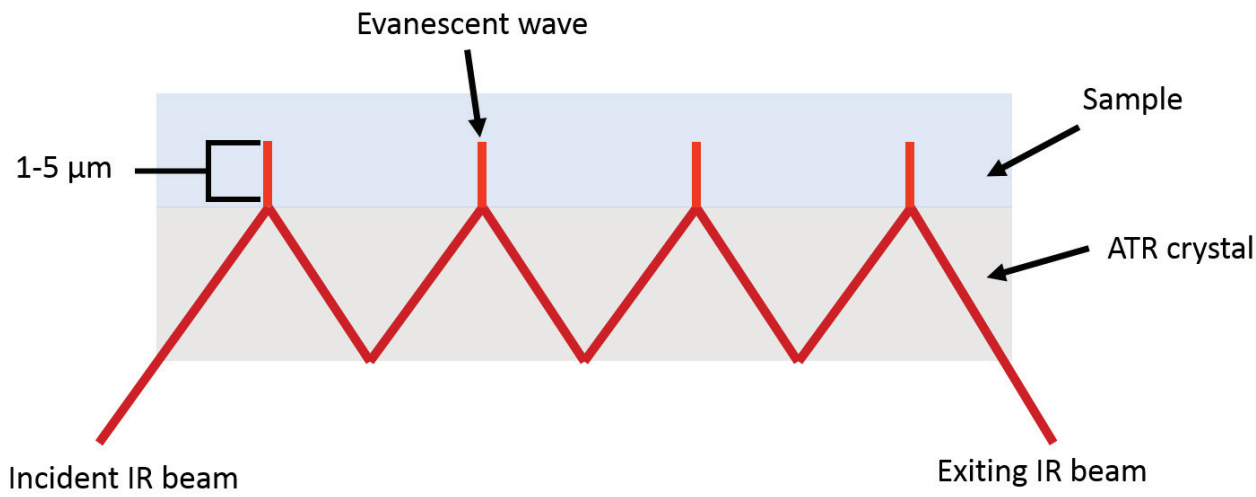

Figure 1.1. Working principle of ATR

As shown in Figure 1.1, the incident infrared beam first travels through the crystal to form an evanescent wave which penetrates the sample with a typical depth of 0.5μ to 5.0μ depending on incidence angle and wavenumber of the beam as well as the refractive index of the whole medium. Depending on the crystal, the attenuated beam either leaves crystal at once or is reflected back and forth several times before reaching the detector (Skoog, Holler, and Crouch 2017).

Given all the practical advantages and improvements of FTIR-ATR over the last decades, it is no surprise that many chemistry related questions are now being answered with this instrument. Determination of adulteration is in fact one of these cases where the conventional methodologies include the use of less practical techniques such as HPLC, GC and MS. Furthermore, these methods involve instrument with high initial-buying and maintenance cost. Usually, there is also need for an expert operating on the instrument and not only sample preparation but also the analysis itself can be time-consuming. Moreover, fingerprint region, that is a part of the Mid-IR region of the electromagnetic spectrum ($4000cm^{-1}$ - $600cm^{-1}$) is suitable for determination of adulteration and for authentication purposes as it corresponds to vibrations of specific functional groups which can carry unique information for each species (Skoog, Holler, and Crouch 2017). Also, the linear

correlation of band intensities with the concentration of this functional groups allows quantitative analyses.

Salep, which is obtained from tubers of orchid, is a quite popular hot beverage in Turkey. Despite the legal regulations of strictly prohibiting exportation, the lack of an established analysis technique that is designed for identification of salep renders the customs unable. Additionally, since the mass production of salep is non-existent, pure salep remains to be a very expensive and encountering adulteration is very likely. For these reasons, there is a need for development of a methodology to quantitatively determine salep adulteration. Combining the powers of Mid-FTIR-ATR and chemometrics promises a fast and cheap method for such purpose.

1.1. Structure and Scope of the Thesis

In the first chapter of thesis, the most common chemometric calibration methods are given along with their motivation, mathematical background and application guidelines. In the second chapter, the implementation challenges to achieve fast and elegant calibration toolbox were discussed and practical strategies of development were shared. Additionally, a case study of determination of salep adulteration were provided in the third chapter in which different calibration techniques were compared and the several metrics are established specifically to account for real-world scenarios.

1.2. Literature Review

There are only a few articles introducing chemometrics toolbox for calibration. TOMCAT toolbox (Daszykowski et al. 2007) provides several calibration techniques such as PCR, PLS, Continuum Power Regression, Partial Robust M-Regression and Radial Bases PLS while also providing classical and robust PCA. SAISIR toolbox (Cordella and Bertrand 2014) provides a more comprehensive toolbox with more than 200 functions which provide calibration, classification, curve resolution and preprocessing capabilities. Furthermore, these toolboxes are either designed for a general purpose or too specialized. Additionally, both of them lacks parallelization and no implementation of genetic algorithm based calibration were offered.

In literature, many studies can be found which combines chemometric methods with FTIR-ATR for adulteration determination. A recent review shows that the use of

this combined method is very popular for food related targets thanks to its extendibility to larger scales, high speed and non-destructive nature (Rodriguez-Saona and Allendorf 2011). It was also shown that both NIR and MIR are used in various of adulteration problems. Furthermore, some of these studies not only provide quantitative results which can be limited by the number adulterants involved, but also authentication considerations to account for unspecific adulteration cases. Recently, this approach is supported by some authors while application of discriminant analysis alone were criticized (Rodionova, Titova, and Pomerantsev 2016).

In one study, NIR spectroscopy combined with GILS provided a SEP values below 3% (v/v) for the determination of extra virgin olive oil adulteration with single (sunflower) and with two (corn oil and sunflow oil) adulterants (Özdemir and Öztürk 2007). In a similar study in which MIR was used, adulteration of extra virgin olive oil with binary mixture of corn and sunflower, rapeseed, and cottonseed oils were determined qualitative through PCA and quantitatively by PLSR with detection limits of 5% and 10%, respectively (Gurdeniz and Ozen 2009). In another study for authentication of fruits, MIR spectroscopy along with SIMCA were employed and 100% classification rate was achieved by extraction (Shao and He 2007). In a more recent study, Mid-FTIR-ATR were used for determination of beefburger adulteration with offal adulterants (kidney, heart, liver and lung), where the adulteration scenarios were generated using D-optimal experimental design (46 samples) and results indicated that SIMCA provides 100% classification performance for both fresh and freezed-thawed sample (Zhao, Downey, and O'Donnell 2014).

CHAPTER 2

CALIBRATION METHODS

In the context of chemistry, calibration refers to relating predictor variable(s), which is often a collected response from an instrument or a measurement, to response variable(s) that is compound's property of interest. This relation is in the form of a mathematical model and is obtained through regression techniques (Brereton 2003). The ultimate goal of a regression technique can be defined as providing a model which yields minimal prediction error.

While there are many calibration methods each attacking to the problem of providing minimal prediction error from different perspectives, there is no single best-performing method in all cases. The performance of the calibration method depends heavily on the nature of the data (Brereton 2003; Friedman, Hastie, and Tibshirani 2001). The characteristics of a data include number of samples, number of variables, amount of noise, correlation of variables, and the type of relation (i.e. linear, quadratic, exponential) (Friedman, Hastie, and Tibshirani 2001). Furthermore, even if a calibration method offers hypothetically the best solution for a specific data, it is also expected to be feasible in terms of memory consumption and computational cost.

In addition to the prediction, the interpretability of the model is also quite favorable. For instance, in spectral data, a calibration model can reveal which peak or combination of peaks are relevant to the concentration of a compound. However, it should be noted that the interpretation may become misleading in presence of multicollinearity that is high correlation among predictor variables (Friedman, Hastie, and Tibshirani 2001).

2.1. Beer-Lambert's Law

Beer-Lambert's law states that absorbance of a species is linearly proportional with the concentration of that species (Skoog, Holler, and Crouch 2017). The derived formulation for this relation is given in Equation (2.1).

$$A_\lambda = \epsilon l c \tag{2.1}$$

In Equation (2.1), A_λ is absorbance at the wavelength λ, ϵ is molar absorptivity,

l is path length, and c is concentration. The wavelength at which the absorbance is used for modeling is usually chosen at the maximum of the corresponding peak to reduce the effect of non-monochromatic radiation and for higher signal-to-noise ratio. Moreover, if the path length and molar absorptivity of a species is known, the concentration can be calculated using only the absorbance value.

Despite its simplicity, the applicability of Beer-Lambert law is limited by the instrument and/or chemical properties (Skoog, Holler, and Crouch 2017). The common limitations are listed below.

- Concentrations above 0.01M (due to increased Coulomb interactions).
- Radiation of sample (i.e. fluorescence or phosphorescence)
- Non-monochromatic radiation
- Stray light
- Scattering

2.2. Univariate Calibration

While the molar absorptivity is available for a wide range of compounds, there are also many instances where it is not, i.e., for a newly synthesized compound. Similarly, if the absorbance readings are carried out using a cell for liquid samples or if a spacer is used for solid samples, then the path length is known whereas for instruments such as FTIR-ATR, the path length is unknown and can only be approximated (Averett, Griffiths, and Nishikida 2008). Therefore, the terms l and c in Equation (2.1) can be combined to a single coefficient β to be found and the resulting equation is given as Equation (2.2).

$$A_\lambda = \beta c \tag{2.2}$$

By carrying out several experiments, where solutions having a range of concentrations are prepared and spectral readings are performed, β can be calculated for prediction of concentration of a species using absorbance values (Brereton 2003; Skoog, Holler, and Crouch 2017). The calibration equation can be represented as in Equation (2.3).

$$a_{n\times 1} = c_{n\times 1} \cdot \beta_{1\times 1} + e_{n\times 1} \tag{2.3}$$

in Equation (2.3), n is the number samples, a is a vector containing absorbance reading at a specific wavelength for each sample, c is a vector containing concentration of each sample, β is a scalar that relates concentration to absorbance value, and e is the errors(residuals) which can not be accounted by the model. In order to find β, the most common method is to use least-squares approach (more details in Multivariate Calibration section). Least squares closed form solution to find $\hat{\beta}$ is given below in Equation (2.4).

$$\hat{\beta} = (c'c)^{-1}c'a \tag{2.4}$$

At this point, small alterations to this equations are necessary for several reasons. Firstly, while univariate calibration is proven to be very useful for the application of Beer-Lambert's law, it is used for a much wider range of chemical problems. To name a few, a chromotographic peak area can be related to the concentration of species and similarly florescence intensity at a specific wavelength may be linearly proportional to an enzymatic activity (Brereton 2003). Therefore, beyond this point, a is replaced with x and c is replaced with y for the sake of convenience. Furthermore, in nearly all cases, the aim is to predict chemical property from instrumental response and not vice-versa. Additionally, having residuals in terms of property of interest can more favorable for interpretability. Thus, the variables are also switched accordingly and final equation is shown below as Equation (2.5)

$$y = x \cdot \beta + e \tag{2.5}$$

The least square solution is shown in Equation (2.6) below. It should be noted that for univariate case regardless of the representation both solutions will yield the same predictions.

$$\hat{\beta} = (x'x)^{-1}x'y \tag{2.6}$$

It is clear from Equation (2.3) and Equation (2.5) that the estimation of $\hat{\beta}$ cor-

responds to the slope of a plot of absorbance vs. concentration. Addition of an intercept term may be favorable to account for the baseline absorbance caused by interfering species with constant concentration or by the instrument (Brereton 2003). The addition of intercept term is possible in two different ways. The first one is to mean-centering both absorbance and concentration values before solving least-squares problem. Mean-centering is shown in Equation (2.7) and Equation (2.9) for absorbance and concentration values, respectively.

$$x_c = x - \bar{x} \tag{2.7}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.8}$$

and

$$y_c = y - \bar{y} \tag{2.9}$$

where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.10}$$

Therefore Equation (2.6) should be altered as shown in Equation (2.11)

$$\hat{\beta} = (x_c' x_c)^{-1} x_c' y_c \tag{2.11}$$

The concentration predictions ($\hat{y}$) from absorbance values of new samples (x_{new}) must now account for mean-centering as shown in Equations (2.12), (2.13) and (2.14).

$$x_{newc} = x_{new} - \bar{x} \tag{2.12}$$

$$\hat{y}_c = x_{newc} \cdot \beta \tag{2.13}$$

$$\hat{y} = \hat{y}_c + \bar{y} \tag{2.14}$$

Another and equivalent option to account for intercept is addition of a column of ones to the absorbance vector to obtain a new matrix as shown in Equation (2.15).

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{2.15}$$

Thus, the calibration problem now becomes Equation (2.16)

$$y_{n\times1} = x\beta_1 + \beta_0 + e = X_{n\times2} \cdot \beta_{2\times1} + e_{n\times1} \tag{2.16}$$

The least-square solution is shown by Equation (2.17)

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.17}$$

This solution, rather than a scalar β, now provides a β vector containing two coefficients, the first being the intercept term (β_0) and the second being the slope (β_1). For the prediction, vector of ones should be concatenated with the absorbance vector or any other type of predictor variable.

2.3. Multivariate Calibration

There are many instances where a specific absorption wavelength also contains absorbances caused by interfering species. This is common, for example, when a concentration of a single compound is to be determined in solution containing many other species. While this problem can be dealt with simply by obtaining molar absorptivity for every interfering species, this approach may and usually does require a lot of effort since it includes seperation, identification and molar absorptivity determination of each interfering compound (Brereton 2003). Additionally, vibrational spectroscopy methods such as Near-Infrared (NIR) spectroscopy inherently yields very wide and usually overlapping peaks (Skoog, Holler, and Crouch 2017). Often for complex compounds, not a single peak but multiple peaks can be relevant to the concentration. In these cases, using a single variable may yield poor prediction performance.

Multivariate calibration addresses the mentioned limitations of univariate calibration by providing the ability to relate multiple predictor variables to response variable (Massart et al. 1997). For the case of concentration determination from absorbance data, multivariate calibration can be considered as an extension of Beer-Lambert's law where the whole or partial spectra can be used for calibration. Furthermore, even though a single absorbance value at the maxima of the peak may provide the most useful information, multivariate calibration allows harvesting of complimentary information around the peak maxima hence can provide increased prediction accuracy (Brereton 2003). There are also cases in which multiple results from several instruments can be combined to provide better models.

Multivariate calibration can refer not only to multiple variables but also multiple responses such that a spectra can be used for prediction of multiple properties.

Besides the advantages over univariate calibration, multivariate calibration has its own drawbacks to be dealt with. In this section, widely used multivariate calibration methods, along with their advantages, drawbacks and practical application strategies will be discussed.

2.3.1. Overfitting, Underfitting and Validation

Many mathematical methods to solve calibration problems involves strong assumptions about the data and whether this assumptions hold or not is not always im-

meditiately apparent (Friedman, Hastie, and Tibshirani 2001). Although there are ways to work-around the limitations within the data to approximate to the best solution, this workarounds involve fine tuning of free parameters(Friedman, Hastie, and Tibshirani 2001; Massart et al. 1997; Brereton 2003).

One big risk with mathematical assumptions and parameter tuning is that a model may appear to be very optimistic for the data used for modeling, even though the actual performance is far worse (Brereton 2003). This phenomenon is called overfitting. On the other hand, rarely, the opposite case can also be encountered. In this case, the model fails to capture relevant information and therefore provides a poor fit and low prediction performance. Thus, for determination of the model performance, it is crucial to validate the model against a data which is not used for modeling. This data set is referred to as validation set or an independent data set or a test set whereas the data used for modeling is called calibration set or training set.

Rather than using the entire data for the calibration, the common practice is splitting the data into calibration and validation sets. While the splitting ratio is optional, the calibration set should cover as much variance as possible and validation set should represent realistic scenarios.

The most common performance metric is the root-mean-squared-error (Brereton 2003). Root-Mean-Squared-Error of Calibration (RMSEC) is defined in Equation (2.18) and is used for calibration set.

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}} \tag{2.18}$$

Above in Equation (2.18), y_i is the actual response variable of ith observation belonging to calibration set (i.e., concentration, activity or another property) and $\hat{y}_i$ is the predicted response variable that is obtained by the model. For validation set, Root-Mean-Squared-Error of Prediction (RMSEP) is defined in Equation (2.19)

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{2.19}$$

In Equation (2.19), y_i now refers to the actual response variable of a sample from

the validation set and $\hat{y}_i$ is its model prediction. The overfitting can now be defined as having very low RMSEC and high RMSEP value whereas underfitting is evident with low RMSEC value. In a perfect scenario, both RMSEC and RMSEP values are expected to be small and close.

2.3.2. Least Squares

As mentioned in univariate calibration, least squares (LS) method can be used to calculate coefficient(s) that relates predictor variables to response variable (Friedman, Hastie, and Tibshirani 2001). For multivariate case, first the definition of the regression equation is defined in Equation (2.20) for i^{th} sample

$$y_{(i)} = x_{(i,1)}\beta_1 + x_{(i,2)}\beta_2 + x_{(i,3)}\beta_3 + \cdots + x_{(i,p)}\beta_p + e_{(i)} \tag{2.20}$$

Above in Equation (2.20), $y_{(i)}$ refers to the response variable belonging to i^{th} sample, p refers to the number of variables, $x_{(i,p)}$ refers to the ith sample's p^{th} variable and $e_{(i)}$ refers to the error. This equation can be represented in a more compact and algebraic manner as in Equation (2.21) for all samples.

$$y_{n\times 1} = X_{n\times p} \cdot \beta_{p\times 1} + e_{n\times 1} \tag{2.21}$$

In Equation (2.21), X is the matrix containing predictor variables in which each column corresponds to a variable (i.e., absorbance at a wavelength) and each row corresponds to a observation (or sample), y is the vector containing response variables (i.e. concentrations), β is the vector of coefficients for each variable and e is the residuals. Since the aim is to find the best coefficients (β), a criterion for "the best" must be defined. For the least square case, the best coefficients (β) are the ones minimizing the function that is the sum of squared residuals (Friedman, Hastie, and Tibshirani 2001). This function is called a loss or a cost function and is given in Equation (2.22).

$$L(\beta) = \sum_{i=1}^{n} (x_i\beta - y)^2 \tag{2.22}$$

Minimizing squared residuals rather than aiming to minimize, for example, absolute residuals is not a coincidence and has good mathematical properties which are being differentiable and having a single minimum (Friedman, Hastie, and Tibshirani 2001). This allows derivation of closed form solution so that the iterative solutions can be avoided. The closed form solution to least squares problem is given in Equation (2.23).

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2.23}$$

2.3.3. Classical Least Squares

As mentioned in univariate calibration, for spectral data, the aim is to predict concentration from absorbance values. While the results are equivalent regardless of the representation for univariate case, for multivariate case, switching predictor and response variables has a big impact on the obtained model.

Classical Least Squares (CLS), in a spectral context, is a direct extension of Beer-Lambert's Law so that not a single absorbance value but multiple ones are related to concentration values hence the naming classical (Brereton 2003). CLS equation to be solved is given in Equation (2.24).

$$X_{n\times p} = Y_{n\times m} \cdot \beta_{m\times p} + e_{n\times p} \tag{2.24}$$

in Equation (2.24), X is the matrix of predictor variables and Y is the matrix containing response variables where m is the number of response variables to be related to predictor variables. This response variables can be the concentrations of multiple chemical species and/or any other properties of interest. It is also apparent that solving this

equation for β allows construction of spectra using the concentrations.

The limitation of CLS is that in order to obtain a successful model, the concentration of each interfering species must be known. Having errors (e) in terms of predictor variables implies that the instrument is the only source of error. Still, use of CLS where appropriate may provide successful models without overfitting (Brereton 2003).

The least squares solution for CLS is given in Equation (2.25)

$$\hat{\beta} = (Y'Y)Y'X \tag{2.25}$$

For most cases, a matrix relating the predictor variables to response variables are desired. For this purpose, a new matrix is derived as in Equation (2.28) by leaving Y alone.

Starting with

$$X = Y\beta \tag{2.26}$$

multiplication of both sides with β' yields to

$$X\beta' = Y\beta\beta' \tag{2.27}$$

followed by multiplication of both sides with $(\beta\beta')^{-1}$

$$Y = X\beta'(\beta\beta')^{-1} \tag{2.28}$$

and a more compact solution can be obtained as in Equation (2.29)

$$Y = XK \tag{2.29}$$

where K is defined as

$$K = \beta'(\beta\beta')^{-1} \tag{2.30}$$

2.3.4. Inverse Least Squares

While a successful model can be obtained by CLS approach when concentration of every interfering species is known, this is not always possible or feasible. Additionally, over the last decades, the instruments are significantly improved in terms of precision and accuracy while, for instance, preparing a set of solutions having various concentrations is still prone to human errors and still limited to precision of containers which have remained nearly same over the years. Thus, there are many instances to assume that the source of error is in the concentration (Brereton 2003).

Inverse Least Squares (ILS), which can also be referred as multivariate extension of inverse Beer-Lambert's Law, addresses this issues by switching predictor and response variables. below in Equation (2.31), ILS calibration equation, which is also used for prediction, is given (Massart et al. 1997; Brereton 2003).

$$Y_{n\times m} = X_{n\times p} \cdot \beta_{p\times m} + e_{n\times m} \tag{2.31}$$

The least square solution for ILS is given in Equation (2.32).

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.32}$$

The problem with ILS is the initial mathematical assumptions that does not always hold (Brereton 2003; Friedman, Hastie, and Tibshirani 2001). Here, calculation of $\hat{\beta}$ in Equation (2.32) involves calculation of left inverse of X that is X^+ in Equation (2.33).

$$X^+ = (X'X)^{-1}X' \tag{2.33}$$

so that

$$I = X^+X \tag{2.34}$$

However, the inversion of $X'X$ requires X having rank r that must be equal to or greater than n. Additionally, the columns of X must be linearly independent. If these conditions are not met, then the inversion of $X'X$ becomes problematic. This is a big issue for spectral data, where having hundreds of variables (columns) and only a few dozens of experiments (rows) is very common ($n << p$ problem). Gaussian and/or possibly symmetric shape of peak(s) introduces dependency among variables (Brereton 2003). Therefore, β estimate obtained by ILS may contain very large coefficients that relates not the relevant information but the noise in the predictor variables to the response variables. This makes the model very sensitive to small changes in the data such as small random fluctuations in spectra (Brereton 2003). These issues makes ILS models very prone to overfitting (Massart et al. 1997; Hoerl and Kennard 1970).

This inversion problem is also apparent with Singular Value Decomposition (SVD) (Golub and Reinsch 1970). First, SVD is shown in Equation (2.35)

$$X = U\Sigma V' \tag{2.35}$$

Above in Equation (2.35), U is the matrix containing the eigenvectors of XX', Σ is the matrix whose diagonal elements are square root of eigenvalues of XX' which are also equal to eigenvalues of $X'X$ and the rest of the elements are zeros, and V is the eigenvectors of $X'X$. The inversion of X can be carried out as in Equation (2.36).

$$X^{+} = V\Sigma^{+}U' \tag{2.36}$$

It is clear from Equation (2.36) that SVD converts the inversion problem to the calculation of Σ^{+} (Golub and Reinsch 1970) where Σ is defined as following:

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{bmatrix} \tag{2.37}$$

The inversion of Σ is carried out by inverting each non-zero element as shown below:

$$\text{For each } \sigma,\ \sigma^{-1} = \begin{cases} 1/\sigma, & \text{if } \sigma \neq 0 \\ 0, & \text{if } \sigma = 0 \end{cases} \tag{2.38}$$

All the properties of data matrix X which makes inversion of $X'X$ problematic is also evident in the Σ as very small diagonal elements. In fact, the inversion of these elements may yield numbers so big that the machine precision may become limiting. To cope with this issue, pseudo-inversion function provided within programming language environments usually round the σ values to zero below a certain threshold so that Equation (2.38) is altered as following:

$$\text{For each } \sigma,\ \sigma^{-1} = \begin{cases} 1/\sigma, & \text{if } \sigma \neq 0 \\ 0, & \text{if } \sigma < \text{threshold} \end{cases} \tag{2.39}$$

By using this approach, numeric inaccuracies as well as very big coefficients in least-squares solution can be partially avoided hence a model with less overfitting can be obtained (Golub and Reinsch 1970). There is, however, no single robust way to determine the best threshold. If the threshold is chosen to be too large, the model may underfit to the data.

2.3.5. Principle Components Regression

Principle Components Analysis (PCA) is a dimension reduction method that can be used prior to calibration (Brereton 2003) and for classification (Brereton 2009) purposes. PCA aims to find the orthogonal directions which maximizes the variance (Friedman, Hastie, and Tibshirani 2001). Application of PCA decomposes the data matrix in two parts namely scores and loadings (Wold, Esbensen, and Geladi 1987). Below, PCA decomposition is given in Equation (2.40).

$$X_{n\times p} = T_{n\times p} \cdot P'_{p\times p} \tag{2.40}$$

in Equation (2.40), T refers to the scores matrix whose columns are sorted in descending order by the variance explained and P is the matrix whose each column is a principle axes and the columns of P are also sorted in the same manner. In order to illustrate the dimension reduction properties of PCA, Equation (2.41) is given below.

$$X_{n\times p} = T_{n\times h} \cdot P'_{h\times p} + E_{n\times p} \tag{2.41}$$

In Equation (2.41), h refers to the number of first principle components (PCs) retained and E refers to the reconstruction error. This projection allows PCA to explain most of the variance in the data matrix X by fewer dimensions (columns). By omitting the directions (PCs) which correspond to small variances, most of the information is preserved (Friedman, Hastie, and Tibshirani 2001). Especially when the variables are highly correlated, the first few principle components are sufficient to account for nearly all variance in the data (Brereton 2009). Furthermore, for spectra data, the instrumental

noises corresponds to a small part of variance and removal of the PCs arguably has noise reduction properties (Brereton 2003).

While there are multiple ways for the application of PCA such as NIPALS algorithm (Wold, Esbensen, and Geladi 1987) and eigendecomposition of covariance matrix (Friedman, Hastie, and Tibshirani 2001), the most common method is using SVD. The complete PCA procedure starts with column-wise mean-centering (and optionally standardizing) the data matrix X to obtain X_c. This is followed by SVD and given in Equation (2.42).

$$X_c = U\Sigma V' \tag{2.42}$$

In Equation (2.42), V corresponds to principle axes P, Σ is a diagonal matrix carrying the square root of eigenvalues. The scores matrix T can be obtained by projection of data on V.

$$T_{n\times h} = X_{c(n\times p)} \cdot V_{p\times h} \tag{2.43}$$

or alternatively,

$$T_{n\times h} = U_{n\times n} \cdot S_{n\times h} \tag{2.44}$$

It is also desired to know how much variance each PC accounts for, so that the number of significant PCs can be determined. For this purpose, first, the total variance should be calculated. This can be achieved either by using scores or by using eigenvalues as shown in Equation (2.45)

$$V_{total} = \sum_{i=1}^{n}\sum_{j=1}^{l} t_{ij}^2 = \sum_{i=1}^{l} \alpha_i^2 \tag{2.45}$$

where l is the number of all PCs, t_{ij} is the ith sample's j^{th} PC score and α_i is the ith singular value obtained from SVD that is the square root of corresponding eigenvalue. Obtaining V_{total} allows calculation of percent variance explained for each individual PC as in Equation (2.46)

$$V_i = \frac{\sum_{j=1}^{l} t_{ij}^2}{V_{total}} \times 100\% = \frac{\alpha_i^2}{V_{total}} \times 100\% \tag{2.46}$$

Principle Components Regression (PCR) refers to relating PCA scores of predictor variables to response variables (Geladi and Kowalski 1986; Brereton 2003). Obtaining this relation by the means of least square is now a more appropriate option for several reasons. The decreased number of dimensions can ease the $n << p$ problem and thanks to the orthogonality of the principle axes, the variables projected on these axes become uncorrelated which is a desired property when solving LS (Friedman, Hastie, and Tibshirani 2001). The LS equation to be solved for PCR is shown in Equation (2.47)

$$Y = T\beta + E \tag{2.47}$$

and the LS solution is given in Equation (2.48)

$$\hat{\beta} = (T'T)^{-1}T'y \tag{2.48}$$

For the prediction of new samples, first the same PCA transformation must be applied to the mean-centered data as in Equation (2.49)

$$T_{new} = X_{new}V \tag{2.49}$$

Then, the prediction is done as shown in Equation (2.50)

$$y_{new} = T_{new}\beta \tag{2.50}$$

Additionally, PCA transformation and prediction steps can be merged and can be carried out using a single predictor K whose calculation is given in Equation (2.51).

$$K = V\beta \tag{2.51}$$

2.3.5.1. Choosing Number of PCs

To construct a PCR model, the number of PCs to be retained must be chosen. One strategy is using a number of PCs which accounts for most of the variance in data and discard the rest. This is usually done by a *scree plot* which is the plot of each PC versus the variance it explains, usually given in percentage (Brereton 2009, 2003). Furthermore an ”elbow” shape is usually observed and evaluated to select number of PCs. In other words, the number of PCs is chosen where the following PCs does not contain relatively significant information that is apparent by values approaching to zero.

Another and more common method for the optimization of number of PCs is using error estimation techniques (Massart et al. 1997; Brereton 2003). This alternative allows the evaluation of potential model performance for each PCs (Brereton 2003). Among the most popular methods that are bootstrapping, jackknife and cross-validation (CV), only CV will be mentioned in this study since it is relatively more common and also included in developed software package.

The basic principle of CV is removing a certain number of samples from calibration set while using removed samples as validation set and repeating this procedure until all samples are used for validation. Moreover, CV can be performed in several ways each refers to the number of samples to be removed in CV steps. The first way is using K-fold CV and as the name suggests each time n/k samples are removed. Another popular alternative is leave-one-out-cross-validation (LOOCV) where a single sample is left out in each CV step. Thus, LOOCV is exactly the same procedure when k is chosen to be the

number of samples n in k-fold CV. The steps for performing LOOCV is given below:

1. Remove the first sample from calibration set
2. Assign the removed sample to validation set
3. Apply preprocessing (mean-centering, standardizing, baseline correction etc.) to the calibration set
4. Construct a calibration model
5. Apply the same preprocessing procedures to validation set
6. Predict the response variables using the obtained model
7. Put the removed sample back into calibration set
8. Remove the next sample unless all the samples are removed once
9. Go to 2nd step

Using CV provides predictions for each sample which are referred to as CV predictions which can be used to calculate the error. The most common error measure is the Predicted Residual Error Sum of Squares (PRESS) whose calculation is given below in Equation (2.52)

$$\text{PRESS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2.52}$$

For PCR, a plot of number of PCs vs. corresponding PRESS values can then be used to determine the optimal number of components (Brereton 2003). In this case, similar to scree plot, an elbow shape is expected where the PRESS values drop sharply for a few number of PCs followed by stabilization or increase. Furthermore, the number of component is also selected with same principles as in a scree plot; where no significant decrease in PRESS value is observed or before the PRESS value begin increasing.

CV results are also useful in terms of providing an initial insight about the potential performance of the final model. Usually, the CV error is expected to be similar to that of final model and big fluctuations that are observed in a number of components vs. PRESS plot usually points to a problem such as day-to-day differences in measurements.

It is important to note that choosing a large number of PCs increases the risk of overfitting. In fact, if all the components are used, PCR solution becomes equivalent to ILS solution which is undesired (Brereton 2009).

2.3.6. Partial Least Squares

Originated from econometrics, Partial Least Squares (PLS) is a valuable and a must-have tool for every chemometrician. Since the invention by Herman Wold in 70s, PLS gained a significant popularity thanks to Svante Wold who used PLS for chemical problems and its successful applications especially on NIR spectra. Furthermore, PLS deals well with the multicollinearity while also being simple and interpretable (Geladi and Kowalski 1986).

Algorithm-wise, there are two flavors of PLS namely NIPALS (or NIPALS-PLS) (Geladi and Kowalski 1986) and SIMPLS (De Jong 1993). NIPALS is the first developed algorithm and unlike least-square method there are no defined objectives to be optimized. In 1993, De Jong (1993) has published SIMPLS method, an alternative algorithm which is faster, more straightforward and yields more interpretable results while also proving that NIPALS-PLS actually follows the objective of maximizing covariance between predictor variables and a single response variable. SIMPLS provides the aforementioned advantages over NIPALS by deflation of covariance matrix in each iteration in contrast to NIPALS which deflates X and Y blocks.

Moreover, in this thesis, the PLS variant that deals with a single response variable (PLS1) were used and explained. Among the two alternative algorithms, SIMPLS was chosen mainly due to its speed.

Similar to PCR, PLS decomposes X matrix but also decomposes y vector (or Y matrix for PLS2 case) as shown in Equation (2.53) and 2.54).

$$X_{n\times p} = T_{n\times h} \cdot P'_{p\times h} + E_{n\times p} \tag{2.53}$$

$$y_{n\times 1} = U_{n\times h} \cdot Q'_{1\times h} + f_{n\times 1} \tag{2.54}$$

In Equation (2.53) and 2.54), h refers to the number of components (for PLS these components are also called Latent Variables, LVs), E refers to the residuals of X, f refers to the residuals of y, U is the y scores and Q is the y loadings. Unlike the PCA loadings,

X loadings (P) in PLS are not used for obtaining X scores, instead SIMPLS algorithm provides ”weights” matrix R, whose columns are normalized and orthogonal. Therefore, the scores are obtained as shown in Equation (2.55).

$$T_{n\times h} = X_{n\times p} \cdot R_{p\times h} \tag{2.55}$$

Multiplication of X scores with y loadings (Q) is then used for prediction as shown in Equation (2.56)

$$\hat{y} = TQ \tag{2.56}$$

Furthermore, it is also possible to define a single vector of regression coefficients for prediction and remove the need for storing R and Q as shown below in Equation (2.57)

$$\hat{\beta} = RQ' \tag{2.57}$$

Selecting number of LVs for PLS Regression (PLSR) is similar to the procedure explained in PCR (Brereton 2003). Generally, a plot of number of LVs vs. PRESS values are evaluated and as more LVs are used, the model becomes more prone to overfitting as it approaches to least-squared solution. Often, the number of significant LVs is related to the number of spectrally active compounds when the aim is to determine concentration from absorbance and this type of information can be exploited when, for instance, the aforementioned graph shows slightly different PRESS values for similar number of LVs.

2.3.7. Ridge Regression

While PCR and PLS copes with the multicollinearity and resulting overfitting problem by truncating the directions of low variance, ridge regression provides an alternative by penalizing the size of regression coefficients (Hoerl and Kennard 1970; Friedman,

Hastie, and Tibshirani 2001). As mentioned in ILS section, as a matrix approaches to singularity, the resulting regression coefficients becomes abnormally large and yields a model with low prediction performance. In ridge regression, another term that is the sum of squares of regression coefficients are added to the least squares cost function and this new function to be minimized is given below.

$$L(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.58}$$

$$= \sum_{i=1}^{n}(x_i\beta - y)^2 + \lambda\beta\beta' \tag{2.59}$$

The closed form solution for minimizing this loss function turns out to be very similar to least-squares solution with addition of a constant to the diagonal elements of the X's covariance matrix $X'X$ to be inversed as shown below (Golub, Hansen, and O'Leary 1999; Hoerl and Kennard 1970; Friedman, Hastie, and Tibshirani 2001).

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'Y \tag{2.60}$$

In Equation (2.60), λ is an constant scalar and and I is the identity matrix matching the size of $X'X$ that is $p \times p$. Fine-tuning of λ is crucial for obtaining a successful model and it can be done by CV. Unlike PLS and PCR, it is usually not possible to exploit prior knowledge about the chemical system hence a large scale of λ values should be scanned and the one that corresponds to the smallest PRESS value can be chosen for final modeling.

Ridge regression falls into category of shrinkage methods and is widely used for machine learning problems (Friedman, Hastie, and Tibshirani 2001) while it is not very popular in chemometrics. It can be considered as a continuous "regularization" method in contrast to the discrete regularization provided by PCR and PLS.

2.3.8. Genetic Inverse Least Squares

Genetic algorithm (GA) is an global search and optimization technique inspired by Charles Darwin's theory of evolution. GA has been introduced by John Holland in 1960 and its been very successful in many areas such as engineering and learning problems (Holland 1992). GA is particularly useful when an analytical solution does not exist and an exhaustive search is computationally unfeasible (Holland 1992).

For the multivariate calibration task, GA adds another level of optimization that is to find best combination of variables in addition to the optimization of the regression coefficients for them, that minimizes the prediction error (in a least-squares or any other manner) (Goldberg and Holland 1988; Özdemir and Dinc 2004). The possible number of combinations grow exponentially with the number of variables and for spectral data it is nearly impossible to test each and every one of variable combination in a reasonable amount of time. Nature, on the other hand, came a long way in a much harder optimization goal of maximizing fitness even though it took billions of years (Holland 1992). Hence, it is no surprise that today GA is one of the most popular nature inspired algorithm for solving difficult problems that we encounter.

Genetic Inverse Least Squares (GILS) is the calibration technique which combines GA and ILS (Özdemir and Dinc 2004). GILS consists of 5 main parts as listed below:

1. Selection of initial genes
2. Evaluation of the population
3. Selection of parents for breeding
4. Cross-over and mutations
5. Replacing parents with offspring

The terminology used in GILS is borrowed from biology and will be explained as mentioned.

2.3.8.1. Selection of Initial Genes

GA starts with populating an initial gene pool. A gene refers to a combination of variables whose count is randomized in certain range that is to avoid $n << p$. For spectral data, a gene is shown below for illustration

$$G_1 : [A_{712}, A_{3662}, A_{744}, A_{941}, A_{2003}, A_{3639}]$$

Above, G_1 consists of 6 variables and A_λ represents the variable that is absorbance at specific wavelength. A predefined number of genes are randomly selected with a constrain of having squared Pearson Correlation Coefficient (R^2) greater than a certain threshold, typically 0.5. The calculation of R^2 is given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.61}$$

Above in Equation (2.61), the prediction $\hat{y}_i$ is obtained by CV using this specific variables with ILS.

2.3.8.2. Evaluation of the Population

The optimization task of the nature is to maximize the fitness. Thus, a fitness criteria must be defined that is appropriate for calibration purpose. Although the fitness function can be chosen as any performance metric, it will be defined here as inverse of root-mean-squared-error-of-cross-validation (RMSECV) as RMSE is usually the final metric used for assessing the performance of the final model. Therefore, RMSECV and fitness function is shown below in Equation (2.62) and (2.63), respectively

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.62}$$

in Equation (2.62), $\hat{y}_i$ is the CV prediction of ith sample.

$$\text{Fitness} = F(G) = \frac{1}{\text{RMSECV}} \tag{2.63}$$

2.3.8.3. Selection of Parents for Breeding

In nature, if an individual has relatively higher fitness, it has more chance of producing offspring and transfer its genes to next generation. In this manner, GA is no different. While there are several methods for simulating this behavior, in this study, ”roulette-wheel selection” method is used.

In roulette-wheel selection, each individual has a chance of being selected for breeding depending on how fit it is. This method can be visualized as placing each individual on a virtual wheel where each of these gene takes up a space directly proportional to their relative fitness as shown in Figure 2.1.

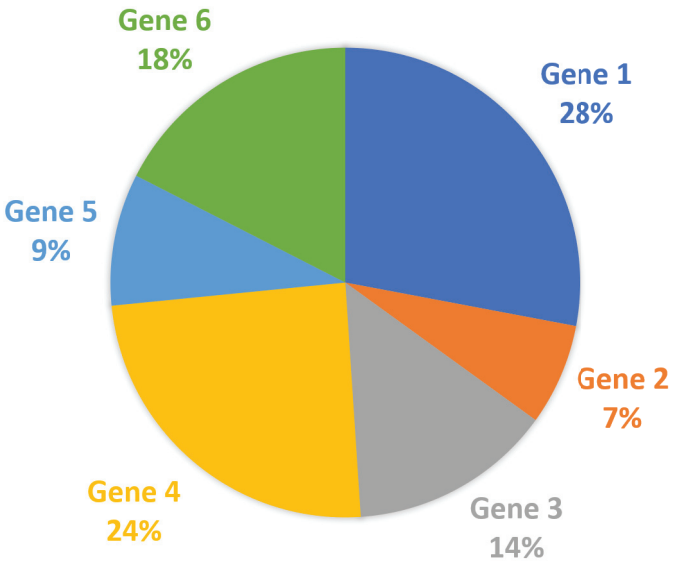

Figure 2.1. A visual representation of roulette wheel used in GA

After the placement of the genes, the wheel is spun that refers to generating a random number corresponding to a gene. After the selection of the gene, it is stored and the procedure is repeated until the number of selected genes are equal to the number of genes.

2.3.8.4. Crossover and Mutations

Perhaps, the most crucial part of GA is the crossover. After the selection of parents for breeding, the matching pair of genes are subjected to a procedure of exchanging

genetic information. In GILS, this is achieved by splitting the both genes from the middle and replacing the halves.

Below 2 hypothetical genes are given:

$$G_1 : [A_{3582}, A_{1653}, A_{2380}, A_{2270}, A_{1988}, A_{1331}]$$
$$G_2 : [A_{2931}, A_{3969}, A_{3999}, A_{1612}, A_{1769}, A_{3008}]$$

The cross-over takes place as shown in Figure 2.2.

$$G_1 : [\boxed{A_{3582}, A_{1653}, A_{2380}} \quad A_{2270}, A_{1988}, A_{1331}\,]$$
$$G_2 : [\,A_{2931}, A_{3969}, A_{3999} \quad \boxed{A_{1612}, A_{1769}, A_{3008}}]$$

Figure 2.2. An illustration of cross-over of two selected genes

The resulting offspring genes are given below:

$$G_1 : [A_{3582}, A_{1653}, A_{2380}, A_{1612}, A_{1769}, A_{3008}]$$
$$G_2 : [A_{2931}, A_{3969}, A_{3999}, A_{2270}, A_{1988}, A_{1331}]$$

In a standard GA implementation, there is small chance of mutation that is removing or adding a random variable. This approach allows some variance within population in a controlled manner while also preventing the population (and therefore the optimization procedure) from being stuck after some iteration (Holland 1992). In GILS implementation, however, mutation procedure is omitted since rather than performing many iterations with the same genes, using multiple small GA runs each with a new population followed by averaging of the resulting models usually yields better results.

2.3.8.5. Replacing Parents with Offspring

After cross-over of all gene pairs, the parents are replaced with their offspring. This effectively causes disappearance of unfit genes while the fitter genes dominate the population.

It should be noted that the even after replacing the parent genes, the one with the highest fitness is always preserved to be used in final model. This is important as GA can suddenly move away from the minimum due to stochastic nature of breeding.

2.3.8.6. Options and Algorithmic Procedure

There exists several options to be set prior to execution of GA. The first option is the number of GA runs. This number refers to how many times a new GA is started. Number of iterations refers to the count of breeding and cross-over in a single GA run. Number of genes is another option that defines how many genes to be used in GA. Number of genes must be an even number since the cross-over takes place on pair-wise selected genes. Furthermore, a range for number of initial variables can also be set and the upper bound typically should not exceed $3 \times n$ and lower boundary should not be less than a few variables. Finally, after many iterations, there is risk of having excessive or deficient number of variables in a gene. Therefore, a maximum and a minimum must be defined and is used after each iteration to ensure the abnormal variable count is avoided. If variable count of a gene exceeds or falls below the defined count, then it is removed and replaced with a new gene.

The algorithmic procedure of GILS is given below:

1. Populate the gene pool by the genes that are initially R^2 values above threshold.
2. Evaluate the fitness of each gene by CV
3. Store the gene having highest fitness
4. Select gene pairs for breeding using roulette-wheel selection.
5. Apply cross-over and optionally mutations and obtain offspring.
6. Replace the parents with offspring
7. Repeat steps 2-6 (a GA iteration) as specified by the number of iterations

8. Repeat steps 1-7 (a GA run) for desired number of times (number of runs).

9. Obtain final model by averaging best genes obtained from each GA run.

2.3.8.7. Remarks

Using GA on top of a calibration method has several advantages specifically for spectral data. The first one is the fact that most of the time not all the variables (absorbance at a wavelength) carry relevant information about the chemical property of interest. GA algorithm effectively scans different combinations and improves the results rapidly to preserve relevant variables while removing the others. Additionally, since each time only a small chunk of variables are used for constructing an ILS model, $n << p$ problem is practically eliminated even though the multicollinearity can still exist. Although it is also possible to use PCR, PLS or Ridge regression with GA, ILS having no parameter to be tuned allows faster constructing of GA models.

2.3.9. Ensemble Models

One strategy to improve model success is to combine the power of multiple models (Dietterich 2000; Friedman, Hastie, and Tibshirani 2001). For the calibration task, while there are many ways of obtaining ensemble models, the most straight-forward approach is averaging predictions of multiple models. Since the calibration methods in this thesis study are all linear models, simply averaging the regression coefficients will provide the same effect for all aforementioned methods except for the CLS. There are many popular alternatives for ensembling including but not limited to bagging, boosting and stacking. They are, however, out of the scope of this study.

Despite not being very popular in chemometrics, machine learning community enjoys the advantages of ensemble model (Dietterich 2000; Friedman, Hastie, and Tibshirani 2001). The main expectation is obtaining a more robust and higher performing model. This can be justified by the fact that each individual calibration technique is designed for different optimization task and deals with the data in a different way. Therefore, using ensembling strategy usually results in a model that is less sensitive to outliers (Friedman, Hastie, and Tibshirani 2001).

CHAPTER 3

DEVELOPMENT OF CHEMOMETRIC CALIBRATION TOOLBOX

Within the context of this thesis, a new toolbox for chemometric regression has been developed in MATLAB environment. One significant motivation of this development is the lack of high quality, easy to use, complete, fast and versatile MATLAB toolbox specifically tailored for spectral data. These properties are only available through highly priced commercial software such as PLS_Toolbox (Eigenvector Research inc.), and additional MATLAB toolboxes. Moreover, the development of this toolbox was carried out with the following concerns in mind:

- Speed
  - Parallelization
  - Vectorization
  - Efficient Cross-Validation
- Coverage of calibration methods
  - ILS
  - PCR
  - PLS
  - GILS
  - Ridge
- Basic (spectral) preprocessing techniques
  - Centering
  - Scaling
  - Baseline correction
  - Variable range selection
- Providing simple models

- User-Friendliness
    - Graphical User Interface
    - Preview of preprocessed data
    - Producing graphs for evaluation of model performance and diagnostics
- Maximum Octave and backward MATLAB compatibility

In this chapter, the aims and methodologies to achieve them will be discussed further in detail.

3.1. Speed Concerns

Even though Moore's Law, which predicts the exponential growth of number of transistors in squared inch accompanied by the decrease in cost, continues to hold to this day, the amount of data produced is also increasing in an exponential way. Therefore, code optimizations are still crucial and help more trial-error approaches to be carried out and may lead to more comprehensive studies. This section mentions the strategies used for speeding up the algorithms during the development of the calibration toolbox.

3.1.1. Vectorization

The algorithms involved in chemometric techniques often consist of repetitive steps applied on data such as dot product, mean-centering and standardization. These types of operations exist not only in chemometrics but also in nearly all computational sciences and 3D applications. Therefore, the CPU manufacturers started implementing Single Instruction Multiple Data (SIMD) concept where, as the name implies, a single instruction is used on multiple parts of data in a predictable way. The use of SIMD concept increases the speed of vectorizable calculations by many folds. While programming languages such as C, C++ and Fortran are compiled languages where the availability of vectorization is often automatically detected and applied by the compiler, interpreted languages such as MATLAB and Python provides this feature via precompiled libraries or functions which can also be embedded in syntax.

Thus, during the development of this toolbox, the use of loops was avoided as much as possible and the use of linear algebraic operands as well as readily available specific functions were favored. Even though interpretability of the codes may have suffered

due to use of such functions, overall code size and therefore the readability is improved along with the run-time.

3.1.2. Parallelization of Genetic Algorithm

Nearly all modern computers now include Central Processing Units (CPUs) that are not only getting faster by each new generation, but also consists of multiple cores each capable of carrying out computations in a parallel manner. Taking advantage of parallelization may provide significant speed gain over serial computations. As mentioned in Chapter 2, our choice of implementation of genetic algorithm involves averaging of multiple models. In fact, each genetic algorithm ”run” is independent from each other hence parallelization is indeed possible. This approach allows number of cores times faster construction of genetic algorithm-based models.

For the purpose of parallelization, MATLAB's Parallelization Toolbox was employed. This toolbox creates multiple CPU threads and distributes the loop as equally as possible among the threads so that each thread is responsible for completing a specific set of runs. This procedure continues until all the runs are completed. The results of runs have to be stored in matrices where no multiple thread is accessing to the same memory block. This implementation is strictly forced by Parallelization Toolbox as it allows for aggressive optimization by eliminating the need for thread-safety measures.

Unfortunately, Octave does not provide a parallelization solution as of writing of this thesis. GA runs and parallelized CV subroutines works as if there is no parallelization therefore no further code alteration is necessary for Octave.

3.1.3. Calculation of CV Matrices in Advance for Genetic Algorithm

In addition to the parallelization, there exists a computational shortcut to reduce computational cost of cross-validations carried out for the evaluation of genes. The minimal number of CVs can be calculated as following:

$$\begin{aligned}\text{Minimum Number of CVs} &= \text{Number of Genes} \\ &\quad \times (1 + (\text{Number of Iterations} \times \text{Number of Runs})) \qquad (3.1)\end{aligned}$$

A typical genetic algorithm run employed in our studies consists of 30 genes, 50 iterations and is repeated (number of runs) 100 times which adds up to 150,030 CVs. As mentioned in Chapter 2.3.5.1, ideally, each time a specified number of samples are removed from calibration set and assigned to validation set, the mean and standard deviation of both predictor and response variables change and need to be recalculated. This calculation is, when repeated 150,030 times, accounts for a significant chunk of computational expense. Additionally, given the number of variables that is typically in the range of 1000 to 3000, each variable is very likely to be used multiple times.

Because of the aforementioned reasons, an approach of precomputing cross-validation matrices where appropriate is proposed to speed up the algorithm. Since the number of samples to be left out is determined by the user and remains constant during runtime, it is possible to calculate mean and standard deviation of each variable for each cross-validation scenario. For instance, considering LOOCV, where each time a single sample is removed until every sample is left out once, the number of ”sub-calibration sets” and ”sub-validation sets” is equal to the number of samples.

At this point, there are two options. The first option is to store each variable mean-centered and optionally standardized for each CV scenario. For the response variable, which is usually a single column concentration vector, this approach yields the optimal memory and computational cost trade-off thanks to its small size. Thus, 3 matrices are calculated and stored; one for mean-centered and optionally standardized response variables to be used as sub-calibration set, one for mean values and one for standard deviation values both of which are necessary for calculation of squared error.

On the other hand, application of this approach on the predictor variables having many variables such as spectral data may exhaust memory. Therefore, the second option, that is only storing mean and standard deviation of predictor variables for sub-calibration sets, is employed. This corresponds to calculation of centered and optionally standardized predictor variables on-the-fly, but in a faster manner. However, the sub-validation sets, thanks to their smaller size, can be calculated in advance and stored as explained in first option.

3.1.4. Parallelization of Cross-Validation

Since each CV step, which is removal of sample(s) followed by modeling and prediction, is independent the parallelization can be exploited. Although parallelization of CV within GA is possible, the general practice is to parallelizing the outer loop. This is

because of the ”overhead” of initiating thread that includes creating a copy of all variables that is accessed within the thread. Therefore, parallelization of CV is only used for PCR, PLS and Ridge methods.

There are two ways of constructing the loop for CV. The first way is looping over CV folds and in each of these loops scanning a certain range of parameters to be tuned that is number of PCs for PCR, number of LVs in PLS and λ in ridge regression. The second way is to looping over parameters where now the inner loop is responsible for removal of samples. While two ways are expected to perform similar, the former way is a lot more favorable as there are ways of hastening parameter scanning procedures. In the following section this property is discussed in detail for PCR and PLS.

3.1.5. Efficient Cross-Validation for PCR

While performing CV for PCR, after removal of sample(s) in a CV step, mean-centering followed by eigendecomposition of covariance matrix is carried out by the means SVD or directly applying eigen-decomposition to covariance matrix. Fortunately, this computationally most expensive step needs to be carried out only once for each CV step. This is because when the decomposition is applied, all the eigenvectors are obtained at once, hence computing eigenvectors for number of PCs is redundant. Furthermore, the algorithmic order to efficient CV for PCR is given below:

1. Remove k samples
2. Mean-center (and apply other desired processing)
3. Apply SVD to obtain eigenvectors ($n \times h$)
4. For each number of PCs
   (a) Calculate scores (T) using the first m number columns of eigenvector matrix where m is the number of PCs
   (b) Calculate regression coefficient (β)
   (c) Predict the left-out samples
   (d) Store the results
5. Put the removed samples back
6. Go to step 1 unless all the samples are left out once.

Upon completion, for calibration task with a single response, the resulting CV prediction matrix has the size of $n \times h$ where each column corresponds to CV predictions of m components and this matrix can then be used to evaluate the success of each PC.

It should be noted that even with algorithms that calculates the eigenvectors one by one such as NIPALS, this looping strategy is still superior since there is no need to recalculate the first m components while $m - 1$ components will be available.

3.1.6. Efficient Cross-Validation for PLSR

Despite SIMPLS and NIPALS (the PLS version) calculating LVs one at a time, having the inner loop for parameter scanning is still advantageous as stated above. For SIMPLS, there is, again, no need to calculate all m LVs as calculating next LV results in addition of a column to the weights and loadings matrices. The efficient CV procedure for PLSR is given below.

1. Remove k samples
2. Mean-center (and apply other desired processing)
3. Use SIMPLS to obtain weights (R) and Y loadings (Q) for all h components which are to be evaluated.
4. For each number of LVs
   (a) Calculate regression coefficient (β) as in Equation (2.57) using first m columns of both R and Q
   (b) Predict the left-out samples
   (c) Store the results
5. Put the removed samples back
6. Go to step 1 unless all the samples are left out once.

3.2. Preprocessing Techniques

Before constructing a calibration model, preprocessing of the data is usually performed to not only address apparent or potential problems but also to improve interpretability.

3.2.1. Mean-centering and Scaling

The developed toolbox allows mean-centering and scaling of predictor and response variables independently. While mean-centering is required by default for many methods either to serve as intercept or as a mathematical need as in PLS, scaling is optional. When the scale of variables contain large differences, the resulting coefficients may lose their comparability. One advantage of scaling is, as the name suggests, transforming all the variables so that they are in a very similar range. This is achieved by dividing each variable by its standard deviation as shown in equation Equation (3.2).

$$x_s = \frac{x}{s_x} \tag{3.2}$$

where standard deviation s_x for variable x is calculated as shown below in Equation (3.3).

$$s_x = \sqrt{\sum_{i=1}^{n} \frac{x_i - \bar{x}}{n}} \tag{3.3}$$

3.2.2. Removing Variables

In many cases, even though a wide range of variables are obtained, some of them may be redundant. A common example is a full Mid-IR data where it may be known in advance that only fingerprint region is relevant to the property of interest. Additionally, some parts of the data may contain large and apparent noisy parts. Removal of such redundant parts may improve model performance. Therefore, the developed toolbox provides a simple removal of variables capability where the user can select the region to be removed by the help of pointer.

3.2.3. Baseline Correction

Despite the improvements in spectrometers over many years, inherent problems regarding the samples and the instrument may still cause baseline shift especially in vibrational spectroscopy. The thermal detectors in these devices are prone to fluctuations caused by change in ambient temperature and humidity. While working with solid samples through an ATR accessory, the non-uniform particle sizes and the change in distance between the sample and the ATR crystal is another common reason for the baseline shift.

One simple strategy, that is proven to be successful and widely used, is fitting a line to account for baseline followed by the subtraction. The whole procedure is referred to as baseline correction. The procedure starts with constructing the matrix of wavelengths optionally also containing their polynomials as illustrated in Equation (3.4).

$$X = \begin{bmatrix} 1 & \lambda_1^1 & \lambda_1^2 & \cdots \lambda_1^k \\ 1 & \lambda_2^1 & \lambda_2^2 & \cdots \lambda_2^k \\ 1 & \lambda_3^1 & \lambda_3^2 & \cdots \lambda_3^k \\ \vdots & \vdots & \vdots & \\ 1 & \lambda_n^1 & \lambda_n^2 & \cdots \lambda_n^k \end{bmatrix} \tag{3.4}$$

In Equation (3.4), λ_n^k represents the nth wavenumber (or wavelength) raised to the kth power. The next step is to construct a LS model by regressing X with absorbance values for each spectra. In this manner, this method can be considered unsupervised. Furthermore, the regression takes place as shown in Equation (3.5).

$$a = X \cdot \beta + e \tag{3.5}$$

in Equation (3.5), a is the single vector containing absorbance values for a single spectrum. After solving Equation (3.5) to obtain the LS estimate of $\hat{\beta}$, following step is then used to calculate baseline as in Equation (3.6).

$$b = X \cdot \hat{\beta} \tag{3.6}$$

Above, b refers to the baseline to be subtracted from the spectrum. It is also important to preserve the information in the spectrum when applying baseline correction. For this reason, use of several points from spectrum rather than full wavenumber range can be preferred. This points are usually selected where no significant variance is observed and/or from a spot that carries no relevant information. Similarly, increasing k to very high values may result in loss of information. In many cases, including only first and second powers is sufficient.

In the developed toolbox, an option is provided for baseline correction using desired data points and desired polynomial order.

3.3. Providing Simple Models

So far, all the mentioned calibration methods are linear. This implies that each model can be represented as a single vector having a size of n for each response variable. Having such simple predictive model is advantageous for several reasons. The first reason is the fixed size of model regardless of the parameters or options used in calibration method. Also, while MATLAB like software packages are specifically designed to deal with mathematical operations, such operations on other programming languages may require a lot more effort to achieve the same operations. Having a single vector allows the model to be implemented on any platform with any programming language with minimal coding, most likely through a single loop. Moreover, it is not uncommon to work on a commercial or a interdisciplinary project involving people who has programming knowledge. A single vector of regression coefficients are even more crucial in this cases as it can be easily used for prediction through office programs such as Microsoft Excel.

In this section, the strategies used to achieve a single vector for all calibration methods will be explained.

3.3.1. Accounting For Mean-Centering and Scaling

As mentioned in Chapter 2, the mean-centering of both predictor and response variables is a very common practice. While standardization is optional, it can be also desired since the resulting regression coefficients become comparable. Rather than storing the mean values and optionally standard deviation values for each variable, it is possible to alter the regression coefficients to account for both of them. To clarify this property, firstly the regression equation with inclusion of mean-centering and standardization is given.

$$\frac{y - \bar{y}}{s_y} = \sum_{i=1}^{n} \beta_i (\frac{x_i - \bar{x}_i}{s_{x_i}}) \tag{3.7}$$

In Equation (3.7), y is the response variable, $\bar{y}$ is the mean of the response variable, s_y is the standard deviation of response variable, x_i is the ith predictor variable, $\bar{x}_i$ is the mean of the ith predictor variable, s_{x_i} is the standard deviation of the ith predictor variable and β_i is the regression coefficient corresponding to i^{th} variable. After the first rearrangement that is to leave y alone, the resulting equation is given below.

$$y = \sum_{i=1}^{n} s_y \beta_i (\frac{x_i - \bar{x}_i}{s_{x_i}}) + \bar{y} \tag{3.8}$$

After splitting the summation into two parts the equation then becomes as following.

$$y = \sum_{i=1}^{n} \frac{s_y \beta_i x_i}{s_{x_i}} + (\bar{y} - \sum_{i=1}^{n} \frac{s_y \beta_i \bar{x}_i}{s_{x_i}}) \tag{3.9}$$

The second part of Equation (3.9) is independent of the variables and is a scalar. Therefore, it can be used as an intercept term as shown below.

$$\beta_0 = \bar{y} - \sum_{i=1}^{n} \frac{s_y \beta_i \bar{x}_i}{s_{x_i}} \tag{3.10}$$

Also, in Equation (3.8), the first part shows how the regression coefficients (β) can be adjusted to account for scaling. The calculation of each adjusted regression coefficient (β_{new}) is given below.

$$\beta_{new_i} = \frac{s_y \beta_i x_i}{s_{x_i}} \tag{3.11}$$

Therefore, the prediction can now be carried out as shown below.

$$\hat{y} = X \cdot \beta_{new} + \beta_0 \tag{3.12}$$

For different combinations of centering and scaling, Equation (3.10) and Equation (3.11) can be altered accordingly. For instance, if no centering is applied on predictor variables, $\bar{x}$ can be set to 0. Similarly, if no scaling is applied on response variables, s_y can be set to 1. Furthermore, the intercept term is usually included with the regression coefficients as first element.

In the toolbox, all the methods return a single coefficient vector that accounts for centering and scaling having a size of $n+1$ where $+1$ is the result of addition of intercept term as first element.

3.3.2. Accounting For Unused Variables

In many scenarios, while the instrument produces all variables, using only a part of it can be desirable. In spectroscopic cases, for instance, if the relevant peak in known in advance, the removal of the irrelevant part of the spectrum is quite common as it decreases the number of variables and may provide higher performing models. Similarly, some parts of the spectra may contain apparent noise and often removed. In GA, one of the main

goals is to find the best combination of variables whose count is strictly upper-limited. In order to match the size of the model with the instrumental output's regardless of which variables are used, a simple approach is used.

This approach involves keeping a vector of ones having the size matching with the instrumental output. If some variables are removed by user or by GA, the corresponding elements of that vector is set to zero. At the end of constructing the calibration model, the regression coefficients are placed to a vector of zeros whose size matches the instrumental output's in a way that only the elements corresponding to ones are filled. In other words, the final coefficients are zero for all unused variables.

All the calibration methods in this toolbox are designed to take advantage of this approach and the returned model always matches the size of the input variables.

3.3.3. Averaging GA Models

While each GA run produces a linear model whose results are then to be averaged, the averaging of the models is also possible. To achieve this, however, the size of the models must match. By involving the method mentioned in Section 3.3.2, it can be ensured that each GA run produces the model of same size. Therefore, only averaging these models which is averaging the regression coefficients obtained from each GA run, is sufficient to represent the whole GA based model in a compact single vector. Moreover, after averaging, the mean-centering and scaling can also be accounted as explained in Section 3.3.2. This methodology is also implemented in this toolbox.

3.4. User-Friendliness

3.4.1. Graphical User Interface

In order to enable interacting with the toolbox in a easy and visual way, a Graphical User Interface (GUI) was developed. This GUI involves nearly all functionality of the toolbox through a single window for modeling and its screenshot is given below in Figure 3.1.

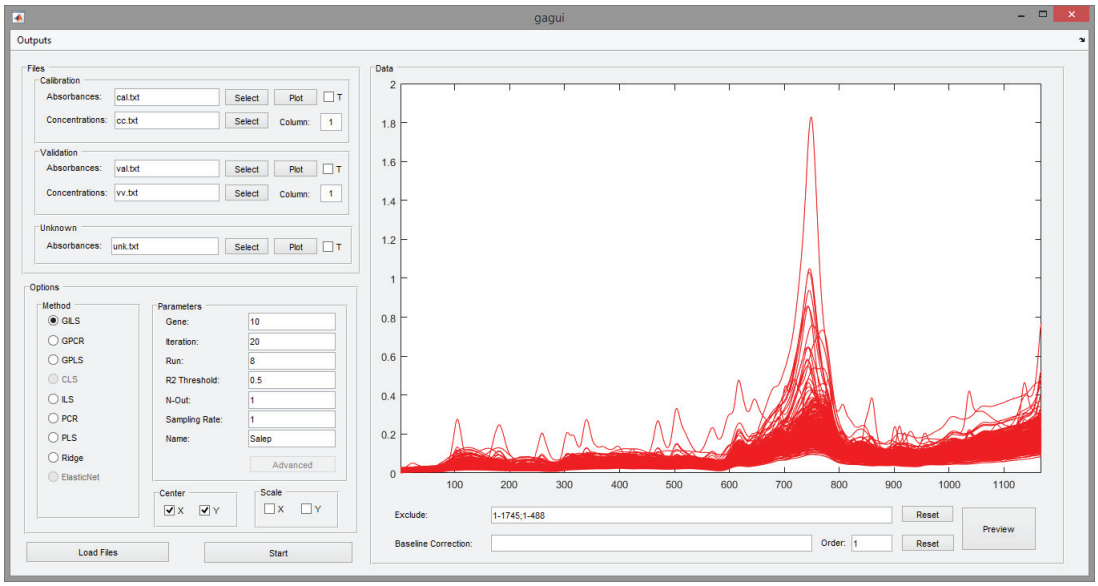

Figure 3.1. A screenshot of the GUI that is developed as a part of the toolbox

As seen in Figure 3.1, the predictor and response variables can be loaded individually through GUI along with the unknown data to be predicted. The toolbox is designed to accept many input formats including character separated files such as ".csv" file format and Microsoft Excel files. The files can be also be loaded in their transposed form.

The main model parameters which can be determined via GUI is listed below:

- Genetic Algorithm Options
  - Number of runs
  - Number of iterations
  - Number of genes
  - R^2 threshold for selection of initial genes
- General Options
  - Number of samples to leave out during CV
  - Sampling rate to omit variables with a given rate
  - Mean-centering of X (predictor variables) and Y (response variables)
  - Scaling of X (predictor variables) and Y (response variables)
  - Selection of component to be modeled (if multiple response variables are given in a single file)

Furthermore, through the ”Preview” button, a preview of the spectra which is baseline corrected with given parameters are shown. If some variables (wavenumbers) are removed, then the preview of spectra will correspond to baseline correction followed by removal of variables. The removal of variables can also be done by clicking on the spectra and selecting the part to be removed. In that case, the resulting spectra are shown immediately.

After setting desired parameters, the target model can be chosen among ILS, PCR, PLS, GILS, and Ridge and the ”Start” button is then used to start the calibration procedure.

3.4.2. Producing Graphs

In the toolbox, several figures are automatically generated for parameter tuning, performance evaluation and diagnostic purposes. The list of these figure are given below:

- Parameter to be tuned vs. PRESS plot (for PCR, PLS and Ridge)
- Actual vs. model predicted values plot
- Residuals plot
- Predictions for unknown samples
- Plot of (scaled) regression coefficients over average variables
- Plot of selection frequency of variables over average variables (GILS)

3.5. Compatibility

Octave is an open-source alternative to MATLAB. While not as elegant and complete, given the price point of MATLAB even without additional toolboxes, the popularity of Octave is not a surprise. Therefore, during the development of this toolbox, use of special functions and abilities that is provided only within MATLAB environment is avoided. By doing so, the aim is to make the toolbox accessible by wider range of academic and commercial users. Unfortunately, GUI is not supported by Octave, yet all the functionality can still be access through the functions and scripts included in the toolbox.

One other concern was the backward compatibility of toolbox to the previous versions of MATLAB. To achieve this, specific functions, which are introduced in new MATLAB versions and provide no improvement, were not used. Furthermore, to decrease the dependency of the toolbox to MATLAB's own toolboxes, most of the functions are re-written i.e. for R^2 calculation and PCA.

CHAPTER 4

CASE STUDY: DETERMINATION OF SALEP ADULTERATION

This chapter of the thesis study contains parts from TUBITAK project (project code: 1003, project number: 115O058) titled as ”Development of fast analytical method(s) for the determination of salep and its adulteration” in which I official worked as a researcher. Furthermore, the collection and preparation of all samples were carried out in Akdeniz University by Assos. Prof. Dr. Mehmet Fatih Cengiz and his students who were also researchers in this project. Mid-FTIR analyses were conducted in our laboratory by Ayten Ekin Mese and Ayse Kevser Bilgin.

4.1. Experimentation

4.1.1. Sample Collection and Sample Preparation

In order to obtain pure salep samples, salep tubers were purchased from various merchants and collectors located in 15 cities which are provided in Table 4.2. The collected tubers were treated in conventional ways. The initially wet tubers were washed and boiled for roughly 15 minutes. Next, the tubers were left to dry under the sun until they are no longer breakable by hand. These tubers are then grinded and filtered through several sieves of different mesh sizes. The final size of particles were ensured to be under 150 μm.

The most encountered 20 adulterants, which are given in Table 4.2, were purchased. Same procedure of grinding and filtering that was applied to salep tubers were also applied to these adulterants. All samples were stored at $4\,^{\circ}\mathrm{C}$ in polyethylene containers.

Additionally, a small algorithm were developed to generate random adulteration scenarios. The algorithm ensures that in each adulteration scenario, there is minimum 1 and maximum 4 adulterant along with certain presence of salep. The amount of adul-

Table 4.1. Region and city information of acquired pure salep samples

Region	Cities
Northern Anatolia	Kastamonu - Tokat - Yozgat - Bartin
Southwest Anatolia	Mugla
Southern Anatolia	Antalya - Mersin
Southeastern Anatolia	Kahramanmaras - Adiyaman - Malatya
Eastern Anatolia	Van - Mus - Siirt - Hakkari - Bitlis

terant(s) and salep is randomized in a way that their sum adds up to 100 since the aim is determining weight percentage (w/w %) of adulteration. Using this algorithm, 290 adulteration scenarios were generated and are provided in Table A.1.

4.1.2. Instrumentation

For collecting FTIR spectra of samples, Perkin Elmer Frontier with Grazing Angle Attenuated Total Reflectance (GATR) accessory was used. The ATR crystal was diamond with 3 reflection. First, spectrum of air was collected to be used as background. Next, each sample was placed on the ATR crystal and the pressure arm was used to apply a certain pressure of 120 units. The application of pressure forces the solid particles to the crystal surface and improves signal. Similarly, the pressure was set to a constant value to obtain consistent results. Since solid samples are known to cause inconsistent baseline-shifts, each sample's measurement were repeated 3 times and averaged to further improve the quality of the spectra to be used for modeling. The spectra were collected in 4000-600 cm^{-1} wavenumber range with 8 cm^{-1} resolution and with 1 cm^{-1} intervals at ambient room temperature.

4.2. Results and Discussion

4.2.1. FTIR Results

Spectra of all 365 samples are given in Figure 4.1.

Table 4.2. List of adulterants along with their abbreviations

Adulterant Group	Name	Abbreviation
Gums	Guar Gum	GUM
	Gum Arabic	GAR
	Locust Bean Gum	KGB
	Konjac Gum	KJG
	Gum Tragacanth	KTR
Starches	Wheat Starch	BNT
	Corn Starch	MNT
	Potato Starch	PAN
	Rice Starch	PIN
Commercial Additives	Sugar	SKR
	Skimmed Milk Powder	YST
	Vanilla	VNL
	Cinnamon	TRC
Others	Grinded Pasta	OMK
	Fine White Flour	GLT
	Grinded Rice	OPR
	Grinded Bulgur	OBL
	Wheat Flour	BUN
	Solid Salep Flavorant	SKA
	Carboxymethyl Cellulose	CMC

As seen in Figure 4.1, there is nearly no baseline shift and no apparent noise except for the small noise around 2000 cm^{-1} caused by the diamond crystal. Moreover, an overlay of pure salep spectra and pure adulterant spectra is given in Figure 4.2 to further observe whether there is a clear spectral difference between salep and adulterants.

In Figure 4.2, while there are a few easily recognizable adulterants, no consistently visible peak(s) for consistently distinguishing pure salep and pure adulterants was observed.

4.2.2. Chemometric Studies

For the determination of salep adulteration, calibration models were established using ILS Regression, PCR, PLSR, Ridge Regression and GILS Regression methods. Even though the concentration of of all adulterants and salep were available for all samples, CLS model were excluded given the complexity of the salep matrix and the purpose of determining only salep's concentration rather than the adulterants. The parameters

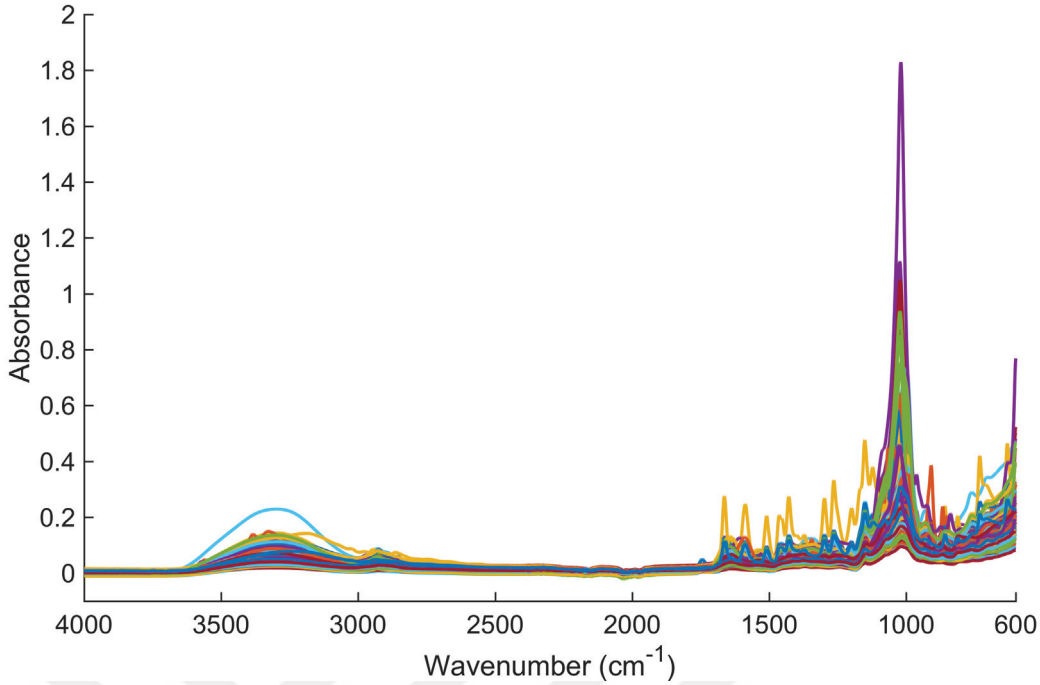

Figure 4.1. Mid-IR spectra of all samples

used for calibration technique are given in their corresponding sections below.

The success of each model can be evaluated through given RMSEC and RMSEP values. The residuals corresponding to each concentration can be observed in residuals plot for additional comments on the model such as error range, behavior at boundary conditions (residuals of pure salep and pure adulterant).

All the calibration studies were carried out using the the calibration toolbox that is developed within the scope of this thesis. Although SIMCA was not included in the toolbox, it was also developed in-house on MATLAB environment.

4.2.2.1. Data Processing

From a total of 365 samples, 280 of them were assigned to calibration set and the remaining 85 samples were assigned to validation set. While designing the calibration set, spectra of all the pure adulterants (20 samples) and 41 spectra of pure salep samples were included in order to make the resulting model cover boundary conditions and account for most of the variance whereas the remaining samples were chosen randomly among

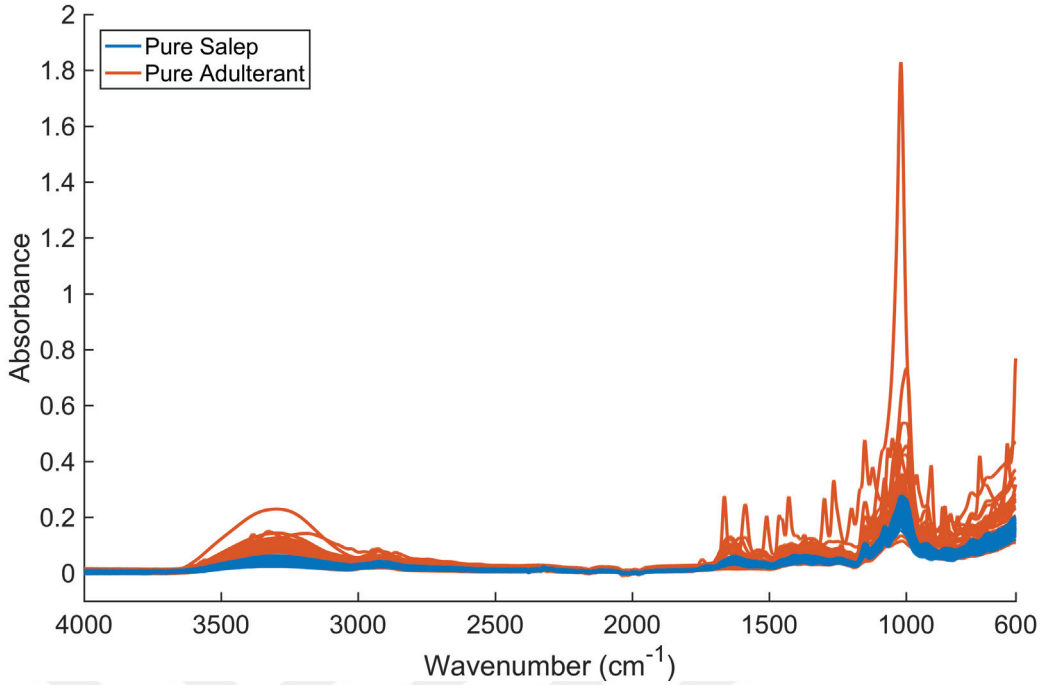

Figure 4.2. Spectra of pure salep and pure adulterant samples

samples of adulteration scenarios. Moreover, validation set consists of 14 pure salep spectra and 71 spectra of adulteration scenario that is randomly chosen.

4.2.2.2. ILS Regression Results

After mean-centering of both spectra (predictor variables) and concentration (response variable), ILS model was established. In Figure 4.3, actual salep concentration and their ILS regression predictions are given.

ILS model, as seen in Figure 4.3, clearly suffers from overfitting since the model provided predictions with nearly no error for the calibration set while the prediction errors are very large. Furthermore, RMSEC and RMSEP values are calculated as 0.00 (w/w %) and 11.00 (w/w %), respectively. The R^2 value is found as 1.00 for calibration set and 0.88 for validation set.

The residuals plot is given in Figure 4.4.

While there is nearly no residual of calibration set predictions, the residuals of validation set are very high overall.

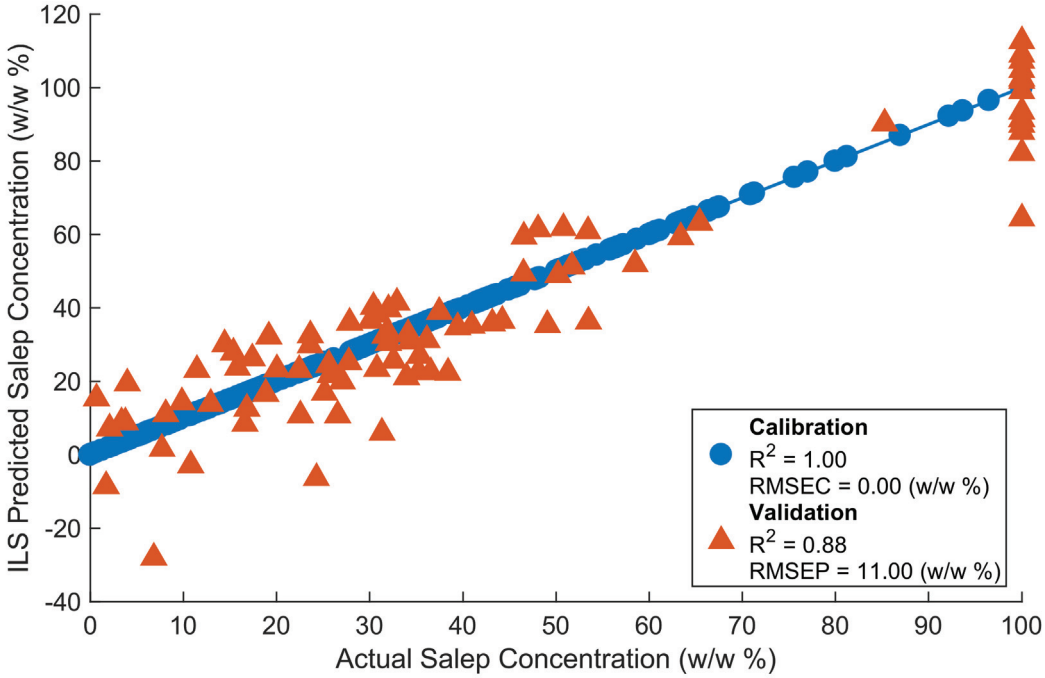

Figure 4.3. Actual salep concentrations vs. ILS predicted salep concentrations

4.2.2.3. PCR Results

In order to determine optimal number of PCs, 40-fold CV (that is leaving 7 samples out in each CV iteration) was applied with mean-centering of both absorbance and concentration variables. In Figure 4.5, the obtained PRESS values for each number of PCs is given.

By using Figure 4.5, a PCR model was constructed using 25 PCs and the results are reported in Figure 4.6.

As seen in Figure 4.6, model's performance on calibration and validation set are not marginally different. Moreover, for PCR model, the RMSEC value was found to be 11.40 (w/w %) whereas RMSEP value is 13.53 (w/w %). The R^2 values are calculated as 0.87 and 0.80 for calibration and validation set, respectively.

The residuals for PCR model is given in Figure 4.7 for both validation and calibration sets.

As apparent in Figure 4.7, almost all residuals fall in the range of ± 40 (w/w %) which is quite high, rendering PCR unusable when high RMSEC and RMSEP values are also considered.

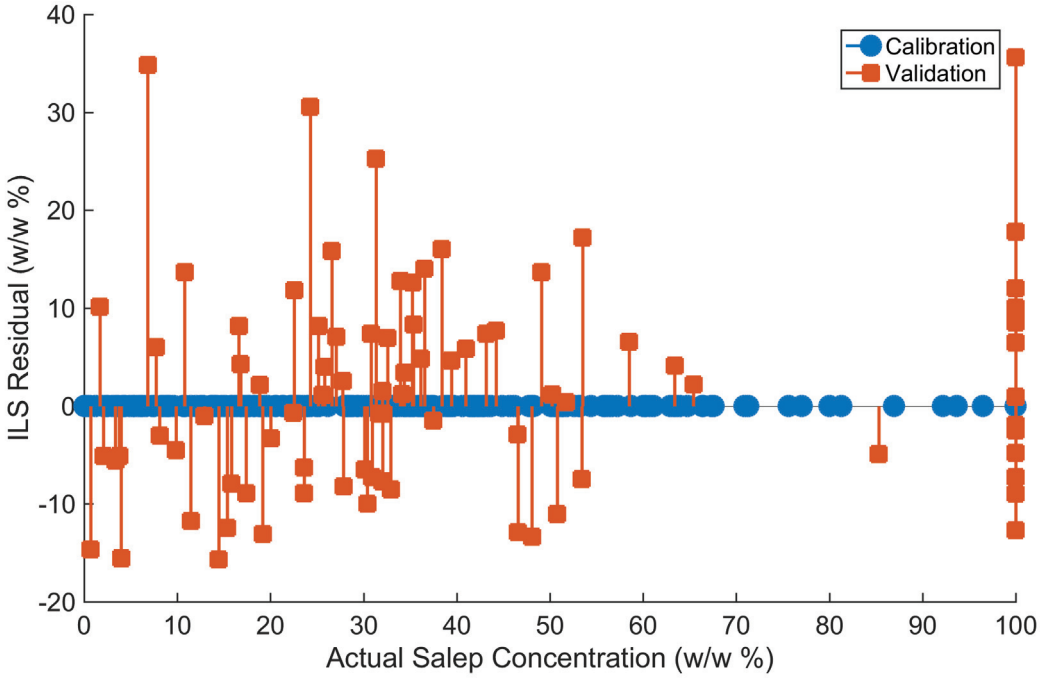

Figure 4.4. Actual salep concentrations vs. corresponding ILS prediction residuals

4.2.2.4. PLSR Results

For finding optimal number of LVs, PRESS values were calculated using 40-fold CV predictions with mean-centering for the first 30 LVs and the results are given in Figure 4.5.

By using Figure 4.5, a PLSR model with 28 LVs were obtained after mean-centering of the data. Figure 4.9 is then obtained to evaluate the model performance.

As seen in Figure 4.9, the model performance is quite close for calibration and validation set predictions with calibration performance being only slightly better indicating no significant overfitting. Additionally, RMSEC and RMSEP values are found to be 5.13 (w/w %) and 7.50 (w/w %), respectively. The R^2 value for calibration set predictions are calculated as 0.97, and the R^2 value for validation set is 0.94. For the determination of error range along with possible residual trends, the residuals plot is given in Figure 4.7.

While most of the residuals are in the range of ± 15 (w/w %), there are exceptions especially for the prediction of pure salep samples, as apperant in Figure 4.7. Even though PLSR model is relatively more successful, this can be problematic for real-life applications as the highest errors are on negative side and can mislead the end-user about

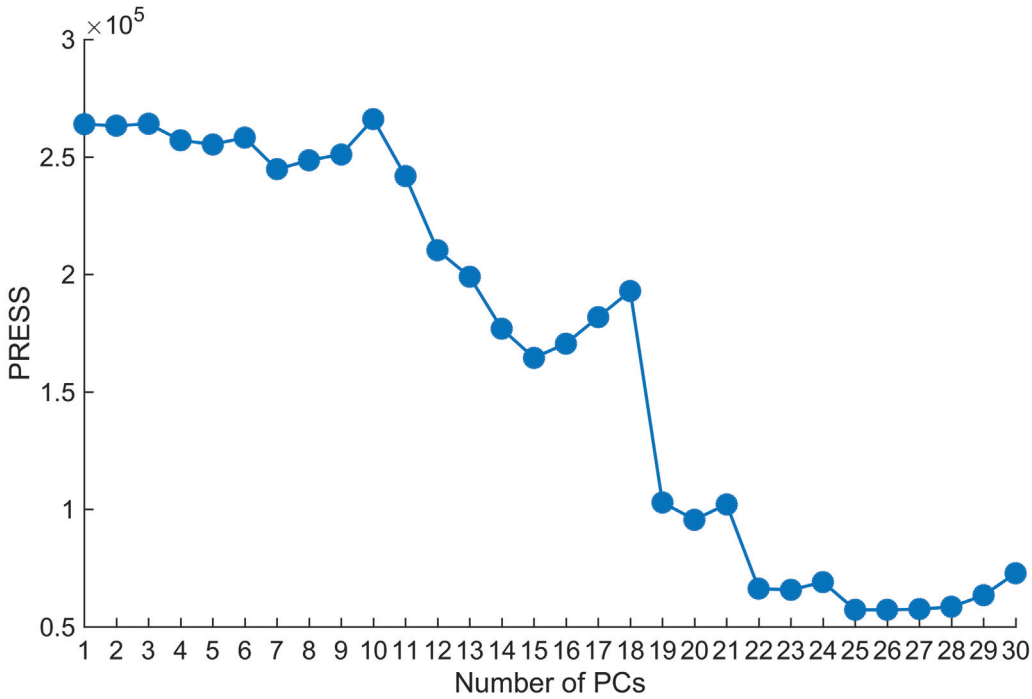

Figure 4.5. Number of PCs vs. PRESS plot for selecting the optimal number of PCs

the samples which are actually pure salep but predicted to be otherwise. One reason for this behavior may be the very high number of components, yet it can be justified by the complexity of the salep and resulting large spectral variance even for different origins.

4.2.2.5. Ridge Regression Results

Unlike PCR and PLSR where the parameter optimization is discrete, in Ridge, the shrinkage parameter λ is a continuous variable. Therefore a range of different λ values were scanned and PRESS values corresponding to them were obtained by 10-fold CV (leaving 28 samples out) with mean-centering. The results are shown in Figure 4.11

According to Figure 4.11, λ parameter were chosen to be 1.1×10^{-5} and a ridge model was obtained after mean-centering. For this model, actual vs. predicted concentration plot is given in Figure 4.12.

As seen in Figure 4.12, while the model has quite low prediction errors for calibration set, it also relatively more successful in predicting validation set compared to the other calibration methods. Furthermore, RMSEC value was calculated as 2.99 (w/w %)

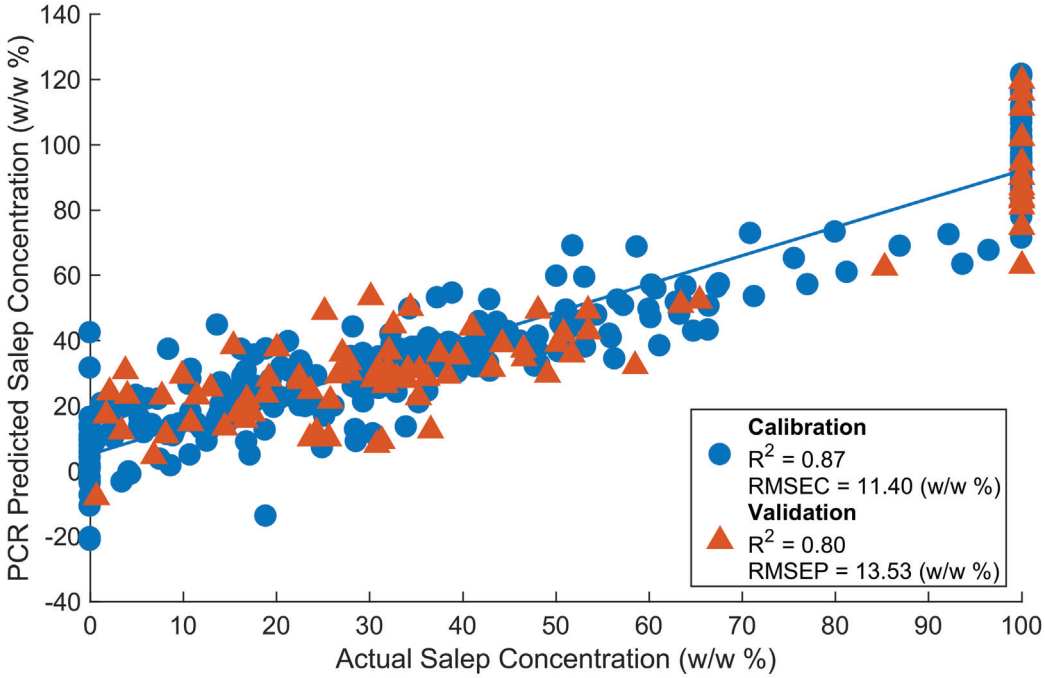

Figure 4.6. Actual salep concentrations vs. PCR predicted salep concentrations

and RMSEP value was found to be 7.20 (w/w %) whereas the R^2 values are 0.99 and 0.95 for calibration and validation set, respectively. To further evaluate the model, the residuals plot is given in Figure 4.13

No apparent residual pattern was observed in Figure 4.13. While a few predictions exceeds 15 % of error, most of them falls in the range of ± 15 % and the overall performance of prediction of pure samples is reasonable.

4.2.2.6. GILS Regression Results

A GILS model was established with 30 genes, 50 iterations and 100 runs where R^2 threshold for selection of initial genes were 0.5 and 10-fold CV was used for determination of fitness. Mean-centering was enabled for both predictor and response variables. In order to determine the success of the model, a plot of actual salep concentrations and their GILS predictions is given in Figure 4.14.

As seen in 4.14, GILS provided one of the most promising model for determination of salep adulteration. In fact, RMSEC was calculated as 4.10 (w/w %) and RMSEP

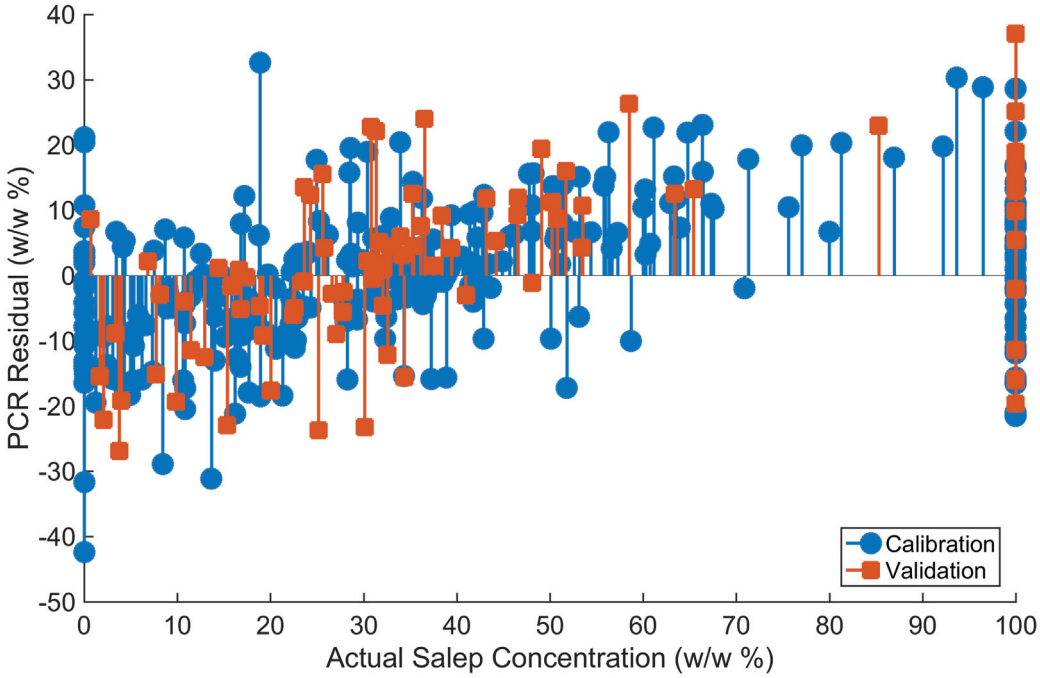

Figure 4.7. Actual salep concentrations vs. corresponding PCR prediction residuals

was calculated as 7.18 (w/w %) which is the lowest prediction error among all individual models followed by Ridge regression. Moreover, R^2 value for calibration set is 0.98 and for validation set it is found to be 0.95. The residuals plot is given in Figure 4.15 to inspect the behavior of errors.

According to Figure 4.15, most residuals are even below ± 15 % and the residuals corresponding to prediction of pure salep samples are in similar range except for a single sample.

One advantage of GILS (or GA in general) is that the selection frequency of variables can be examined and the important wavenumbers for the calibration can be determined. For this purpose, the selection frequency of each wavenumber was given along with the average spectra in Figure 4.16.

As expected, most of the frequently selected wavenumbers are in the fingerprint region according to Figure 4.16. There are, however, few wavenumbers which probably carries no information yet frequently selected. They might be serving as a correction factor for slight baseline or may be the results of stochastic nature of GA.

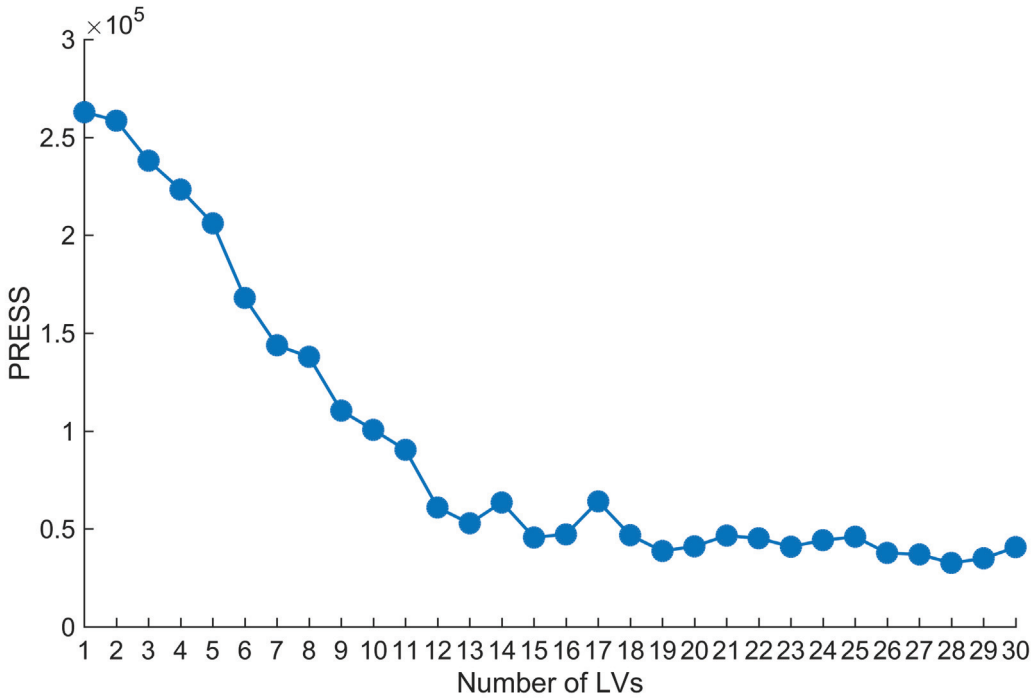

Figure 4.8. Number of LVs vs. PRESS plot for selecting the optimal number of LVs

4.2.2.7. GILS+Ridge (Ensemble Model) Results

While PLS, Ridge and GILS models provided the best results individually, close inspection of residuals indicates that different models perform good on different samples. Hence, employing a combined model can be used as a reasonable way to improve overall performance. Among the three best models, not only PLS performs the worst, but its relatively high error for prediction of pure salep is concerning. Therefore, an ensemble model, which is average of only GILS and Ridge, is established. This model will now be referred to as GILS+Ridge. For evaluation of this model, actual vs. prediction plot is given in Figure 4.17.

It is clear from Figure 4.17 that GILS+Ridge approach outperforms all other models in terms of validation performance. In fact, RMSEC was value is found to be 3.39 (w/w %) whereas RMSEP value, which is the more realistic measure of model performance, is 6.82 (w/w %), the lowest prediction error obtained so far. The R^2 values are 0.99 and 0.95 respectively for calibration and validation set predictions, respectively. Furthermore, the residuals plot for GILS+Ridge model is given in Figure 4.18.

As apparent in Figure 4.18, not only overall performance but also the crucial pure

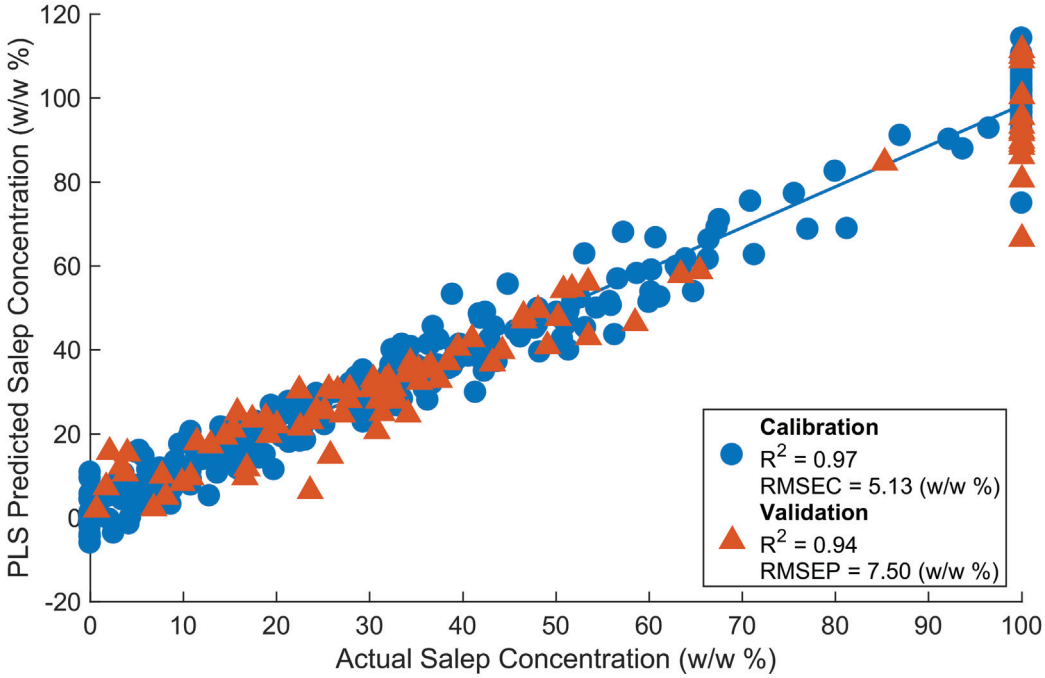

Figure 4.9. Actual salep concentrations vs. PLS predicted salep concentrations

salep prediction performance is improved. The residuals are mostly in the range of ± 10 including most of the pure salep prediction residuals.

4.2.3. Summary and Comparison of the Calibration Models

In Table 4.3, RMSE and R^2 values for all 6 calibration models are provided.

As seen in Table 4.3, despite of the similar RMSEC and RMSEP values indicating no severe overfitting, PCR has the lowest prediction ability. The most probable reason is omitting of PCs which carry relevant information even though these PCs correspond to relatively small variance. This is also apparent by the fact that PCR performed worse compared to ILS. On the other hand, the combination of ILS and GA (GILS) is one of the best performing methods thanks to the variable selection and averaging of multiple runs which reduces the variance. With continuous regularization, Ridge regression provided a RMSEP value similar to that of GILS. Constructing an average model of GILS and Ridge regression improved the RMSEP value even further compared to the individual models.

In addition to the RMSEP, which provides only an overall insight about the model

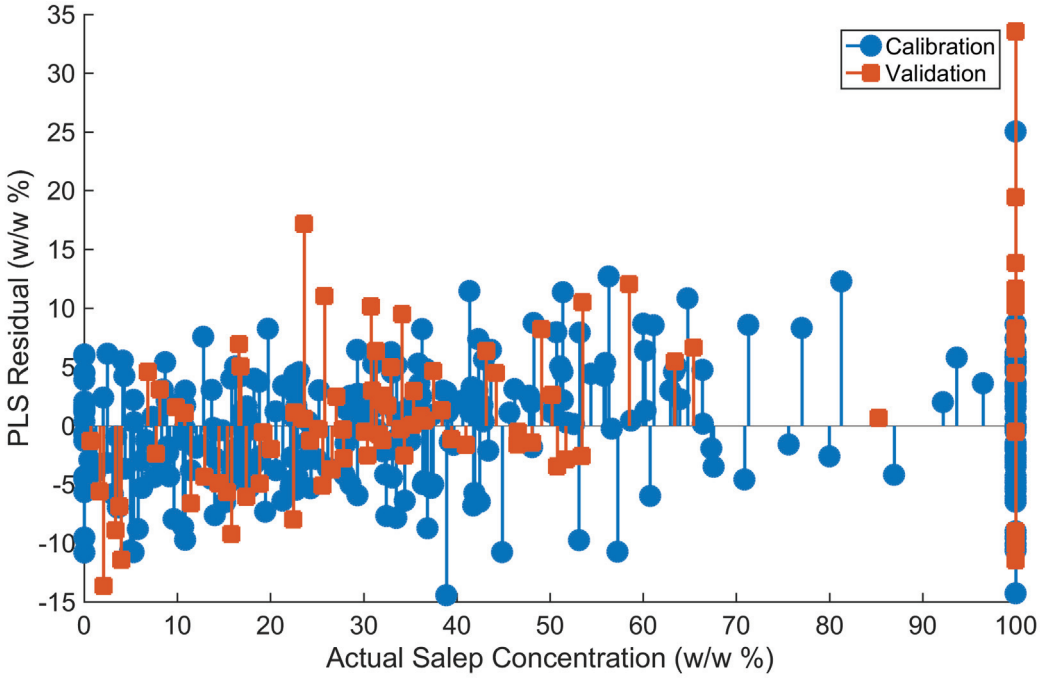

Figure 4.10. Actual salep concentrations vs. corresponding PLS prediction residuals

performance, the ability of the models to correctly determine the pure samples is also significant for adulteration cases. In Figure 4.19, the predictions of pure salep samples within the validation set are given for 3 best models, namely GILS, Ridge, and GILS+Ridge.

Figure 4.19 indicates that for most of the pure salep samples, the predicted values are similar and mostly range between 85%-110% (w/w). Moreover, averaging of GILS and Ridge models provided slightly better predictions since some of the errors appear to be canceled out.

4.2.3.1. Estimating Prediction Performance For Unknown Adulterants

So far, the same 20 possible adulterants were used to generate various adulteration scenarios which were then split into calibration and validation set uniformly. The reported RMSEP values, thus, represent the model performance only for the scenarios involving known adulterants. It is, however, also crucial to evaluate prediction performance of models when a sample involves unknown adulterant(s). For this purpose, a CV-like procedure

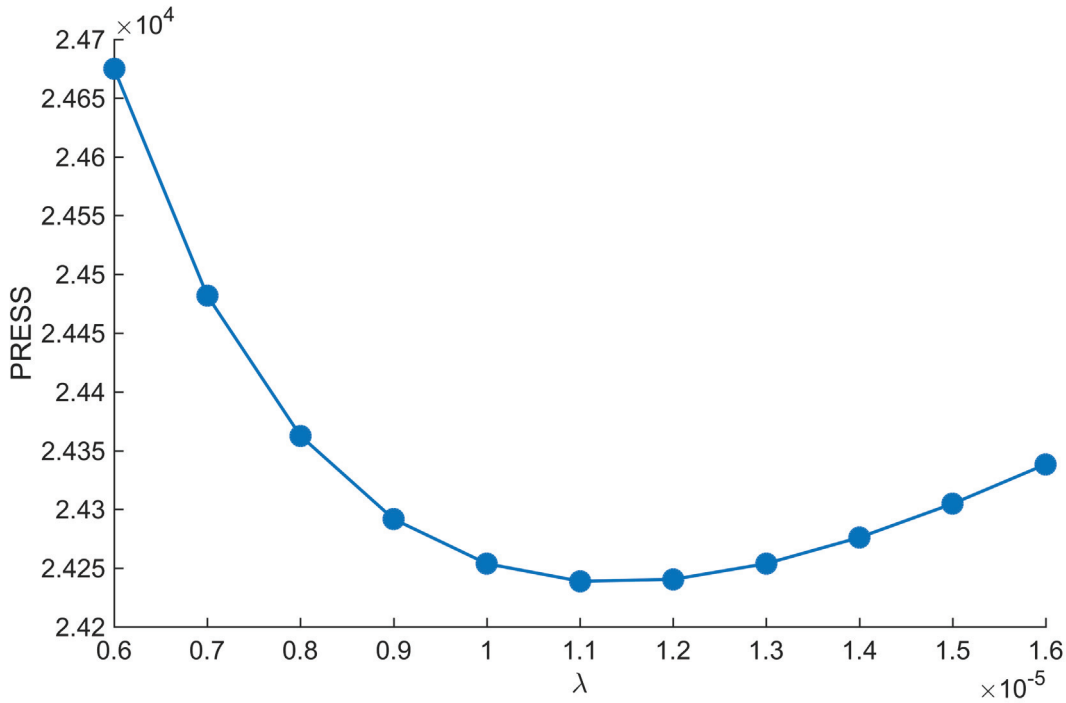

Figure 4.11. λ vs. PRESS plot for selecting the optimal shrinkage parameter

whose steps are given below is carried out to estimate the prediction performance.

For each adulterant:

1. Obtain a calibration set that consists of samples which do not contain that adulterant
2. Assign remaining samples, which are strictly the ones containing that adulterant, to validation set.
3. Construct a model with each calibration technique
4. Calculate and store RMSEP
5. Continue with next adulterant

By this procedure, RMSEP value for each adulterant were obtained using ILS, PCR, PLS, GILS and Ridge regression. The parameter tuning for PCR and PLS were carried out manually using the results from 40-fold CV. GILS model were constructed using 10-fold CV, 30 genes, 20 iterations and 20 runs with 0.5 R^2 threshold. For ridge regression, λ parameters ranging from 10^{-6} to 10^{-2} were evaluated with 10-fold CV and the λ value yielding the least PRESS value were chosen. The results are reported in Figure 4.20

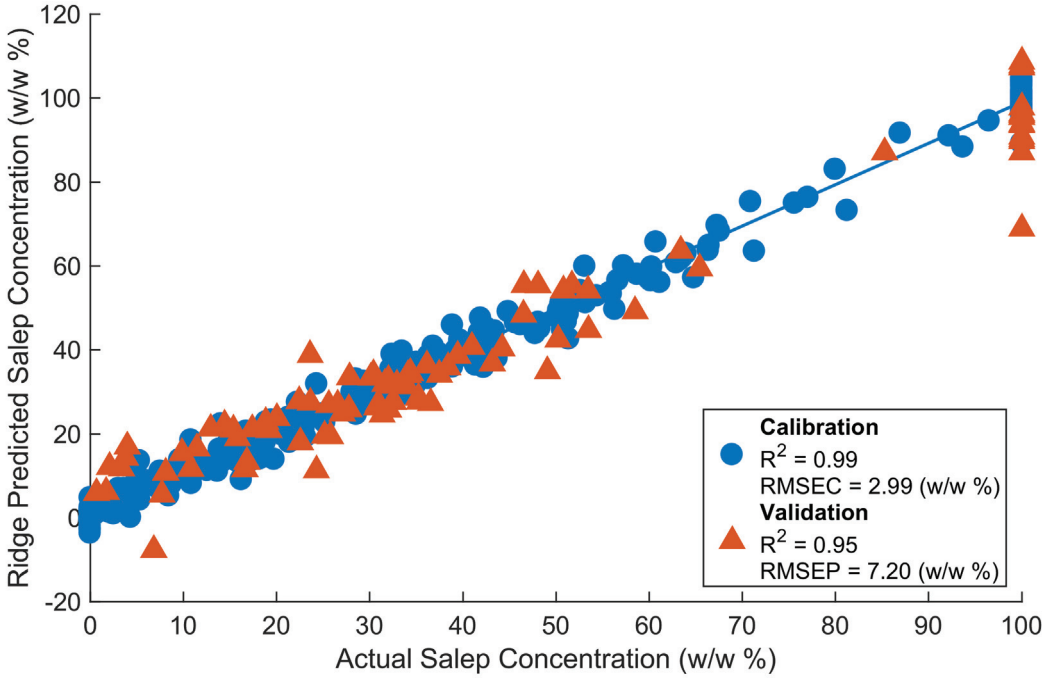

Figure 4.12. Actual salep concentrations vs. Ridge predicted salep concentrations

As seen in Figure 4.20, the results show that RMSEP values are significantly higher for many adulterants, specifically VNL, TRC, SKA, and CMC, when they are not included in the calibration set. For further inspection, the spectra of these adulterants along with spectra of pure salep samples are given in Figure 4.21

As seen in in Figure 4.21, spectra of all these adulterants are significantly different compared to spectra of pure salep samples especially in fingerprint region. Therefore, the low prediction performance of the calibration models for this adulterants can be explained by this high variance that is not introduced to the models within the calibration set.

In these type of cases, predictions can be not only invalid (large values below 0% and above 100%) but also misleading. While accounting for every single possible adulterant may be unfeasible, using a data set with more adulterants should, in theory, help making the model more specific to salep even though there is always a risk of encountering an interfering adulterant.

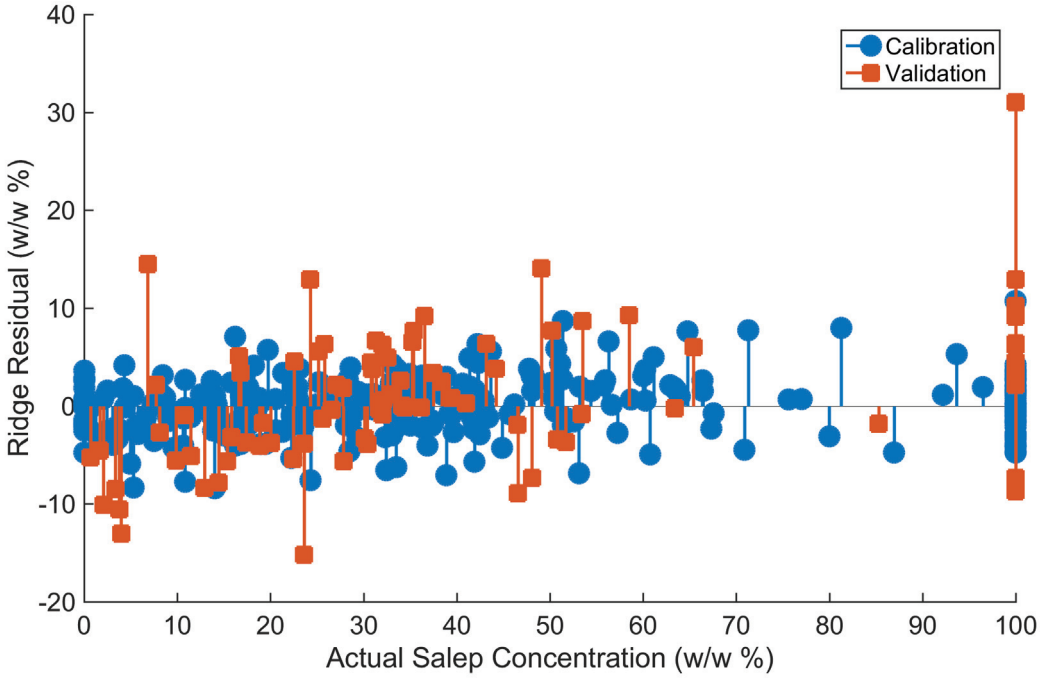

Figure 4.13. Actual salep concentrations vs. corresponding Ridge prediction residuals

4.2.3.2. Single Class Modeling of Pure Salep

Soft Independent Modeling Class Analogies (SIMCA) is a PCA based classification method which allows single class modeling (Wold 1976; Brereton 2009). Using SIMCA is a suitable strategy for determining presence of unknown adulterants since only the variance of pure samples are modeled (Rodionova, Titova, and Pomerantsev 2016). Below, the procedure for SIMCA modeling is given:

1. Apply PCA on the pure samples as in Equation (2.41).
2. Choose the number of PCs by scree plot.
3. Calculate the mean of the scores (group center).
4. Calculate reconstruction error that is $q_i = e_i e_i'$ where e_i corresponds to the PCA residuals of ith sample (as in Equation 2.41).
5. Calculate mahalanobis distance of each sample to the group center to obtain "leverages". For i^{th} sample, the calculation can be done using $d_i = t_i \lambda^{-}1 t_i'$ where λ is

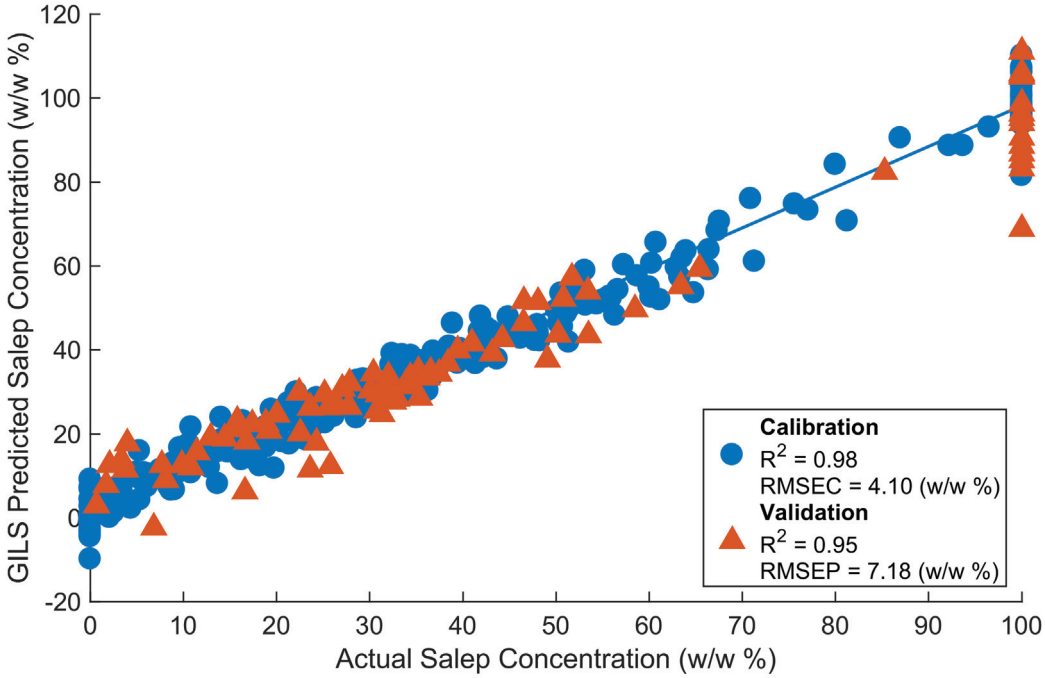

Figure 4.14. Actual salep concentrations vs. GILS predicted salep concentrations

a diagonal matrix whose elements are standard deviations of corresponding score columns.

For prediction of new samples the procedure is given below:

1. Project new samples on the obtained PCA model as in Equation (2.49).
2. Calculate reconstruction error that is $q_i = e_i e_i'$ however this time e_i is calculated as $e_i = x_i - TV'$ where x_i is the predictor variables of ith sample.
3. Calculate mahalanobis distance to group center (calculated during modeling) using the scores obtained by PCA projection of new samples.

Furthermore, since the PCA aims to find directions of maximum variance for a given data set, reconstruction error of samples that does not belong to the class are expected to be higher compared to the ones belonging to the group. A sample's distance to group center also increases as its similarity to the class members decreases. Therefore, plotting this two measures of class-belonging along with their statistically determined decision boundaries (SIMCA plot) can be employed for classification.

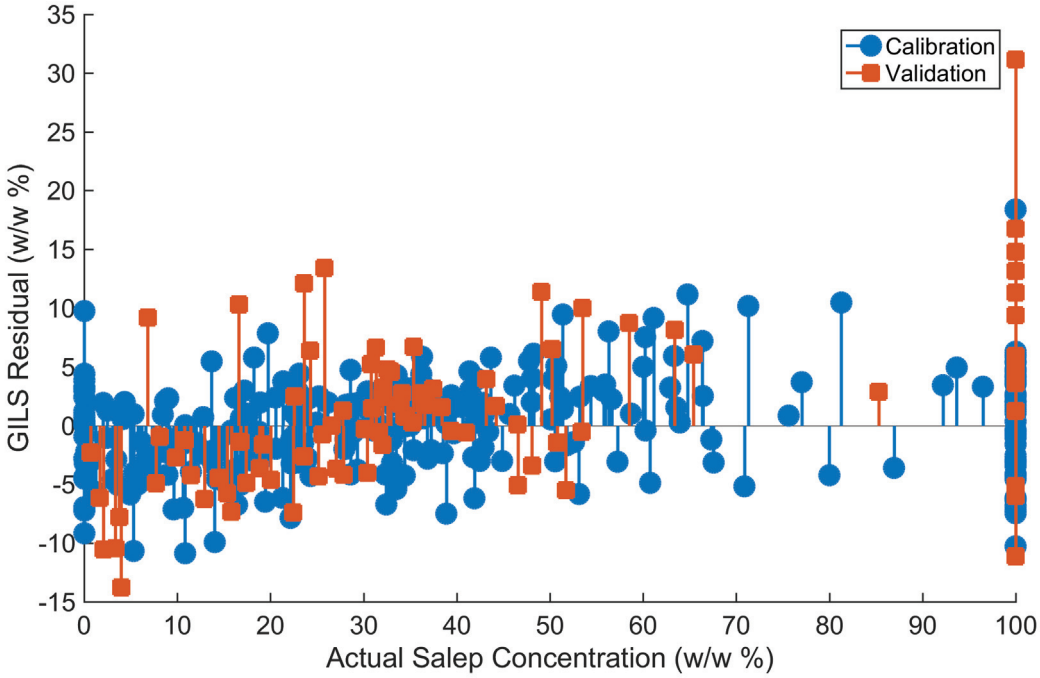

Figure 4.15. Actual salep concentrations vs. corresponding GILS prediction residuals

For determination of salep adulteration, out of 55 pure salep samples, 44 of them were used to construct PCA model whereas 11 pure salep samples and remaining 290 adulterated samples are used for validation. For determination of number of PCs, scree plot were obtained and given in Figure 4.22.

Using 2 PCs, a SIMCA model was constructed and a total of 301 validation samples were predicted. The results are given in Figure 4.23

Due to presence of samples whose predictions yielded very large numbers and does not allow inspection of predictions near the decision boundary, an alternative plot is provided in Figure 4.24 that specifically focuses on the boundaries.

As evident in Figure 4.24, all validation samples were correctly classified. On the other hand, two training set samples were partially misclassified. The most probable reason is PCA (and thus SIMCA) being unable to account for variance among salep samples. It should also be noted that these samples were still correctly classified either with distance to class or with distance to model measures. Furthermore, the level of adulteration and the distance measures are quite correlated which can be useful for providing an quantitative insight.

The use of SIMCA along with calibration predictions can be very useful as SIMCA

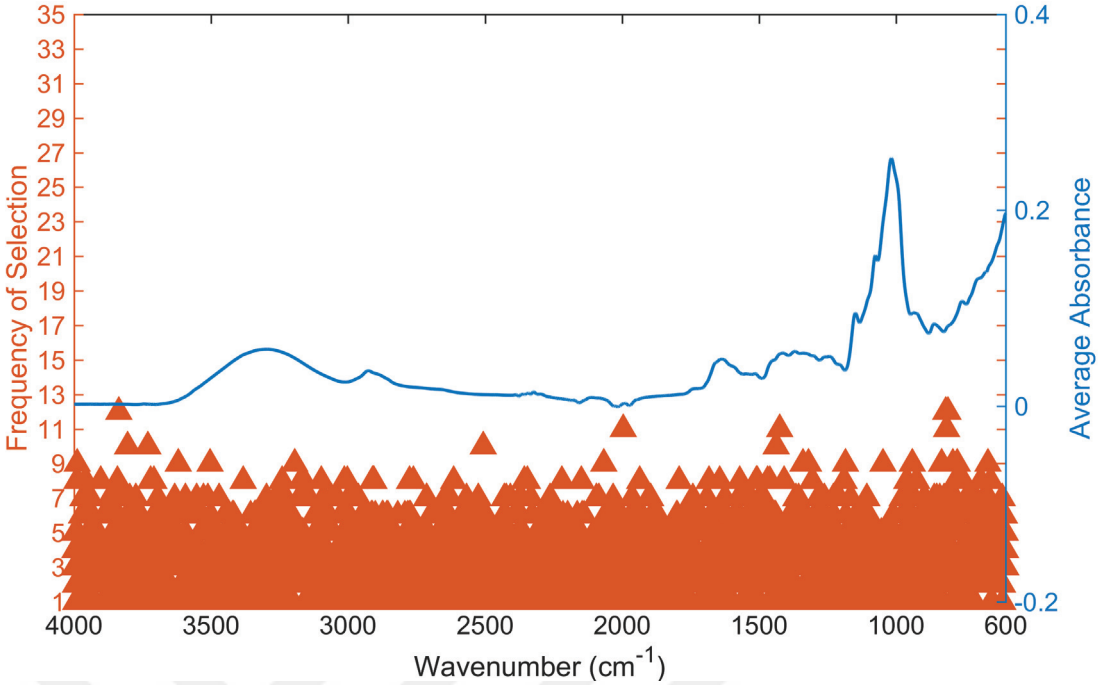

Figure 4.16. Overlay of wavenumber selection frequency for GILS model and average spectra

modeling does not depend on adulterants and can successfully account for virtually any adulteration scenario. While SIMCA was able to classify all validation samples correctly, more samples with lower adulteration levels are necessary for determination of true detection limit.

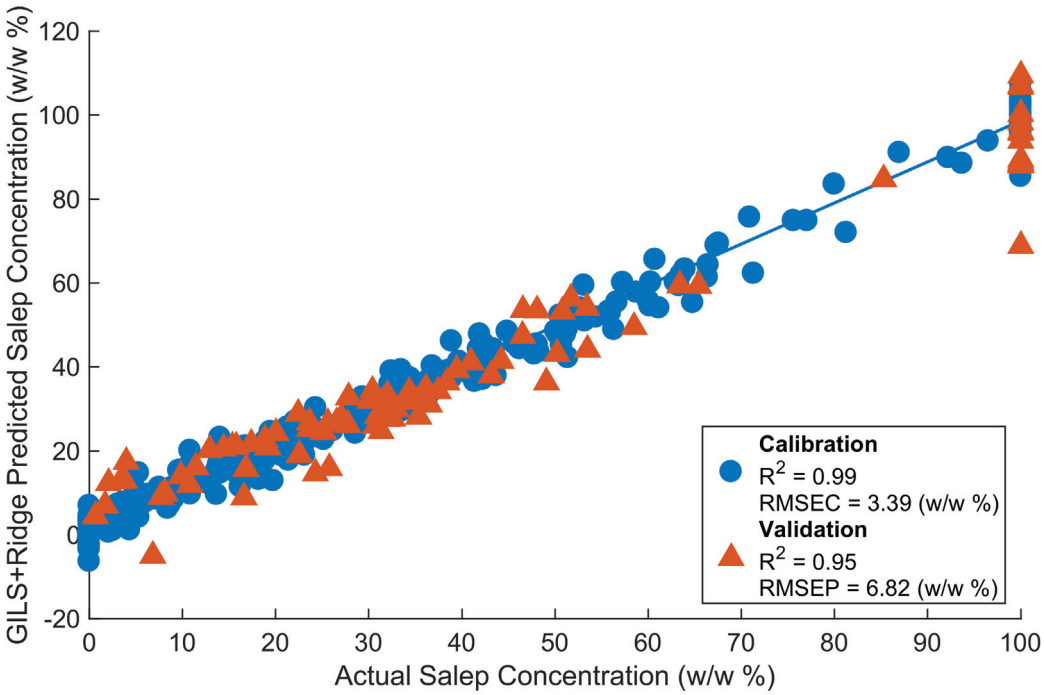

Figure 4.17. Actual salep concentrations vs. GILS+Ridge predicted salep concentrations

Table 4.3. Performance evaluation of all calibration models

Model	**RMSEC (w/w %)**	**Calibration** R^2	**RMSEP (w/w %)**	**Validation** R^2
ILS	0.00	1.00	11.00	0.88
PCR	11.40	0.87	13.53	0.80
PLS	5.13	0.97	7.50	0.94
GILS	4.10	0.98	7.18	0.95
Ridge	2.99	0.99	7.20	0.95
GILS+Ridge	3.39	0.99	6.82	0.95

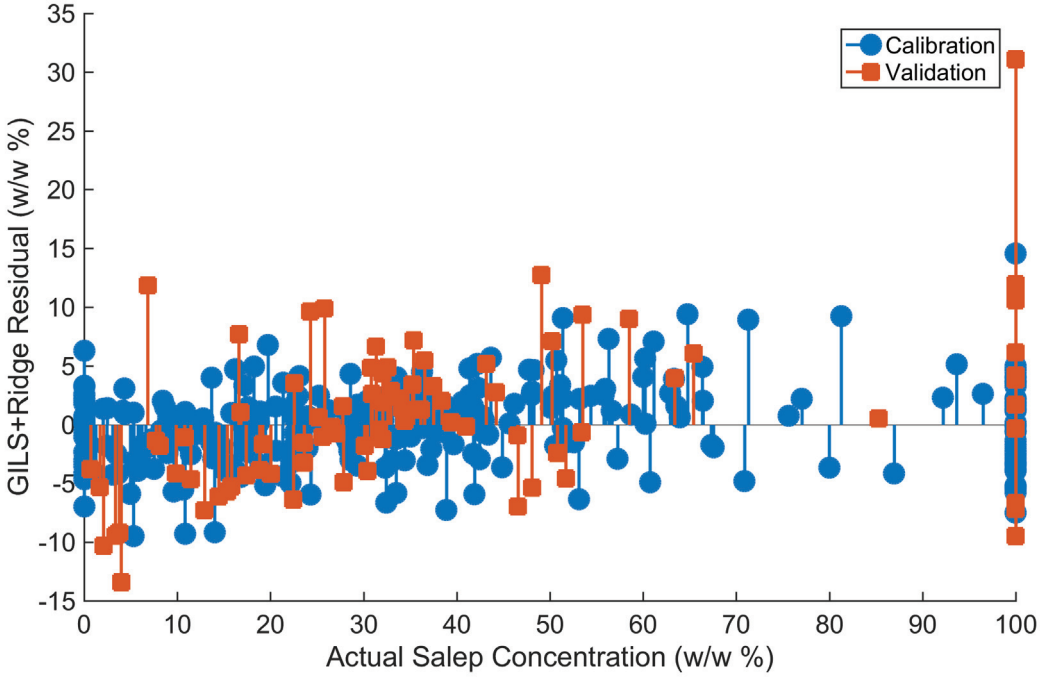

Figure 4.18. Actual salep concentrations vs. corresponding GILS+Ridge prediction residuals

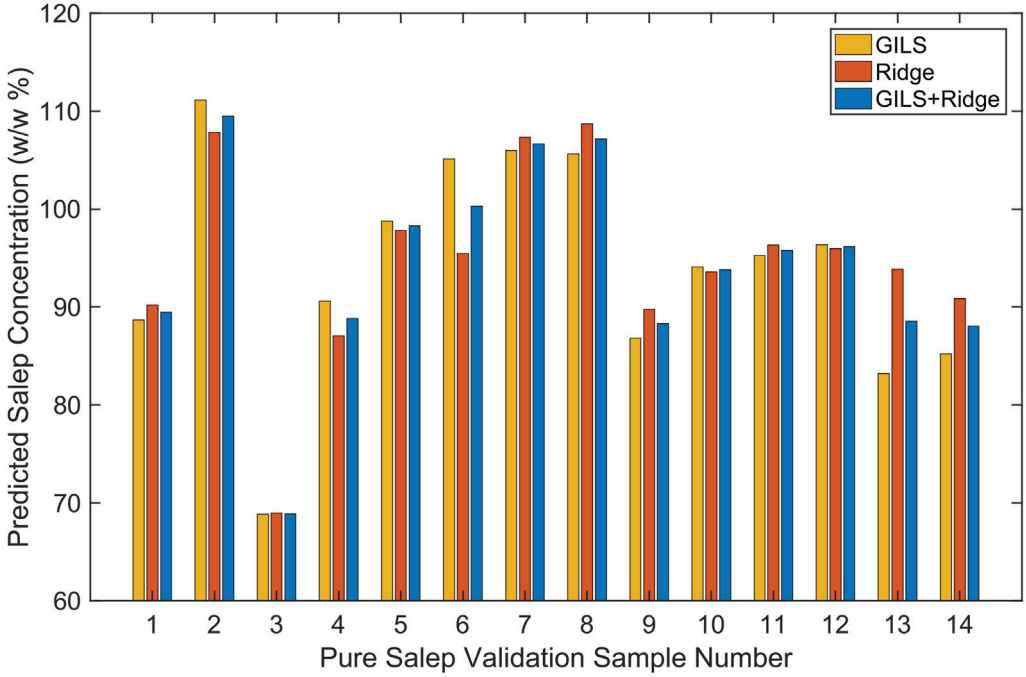

Figure 4.19. GILS, Ridge and GILS+Ridge predictions of 14 pure salep validation samples

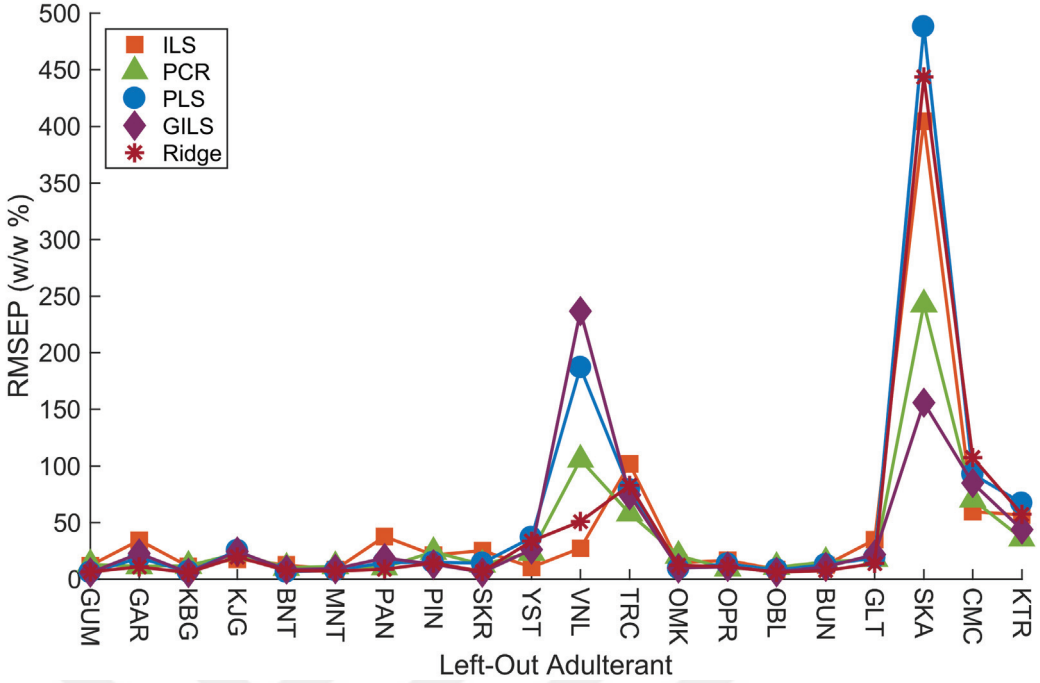

Figure 4.20. Performance estimate of calibration techniques in the presence of an unknown adulterant

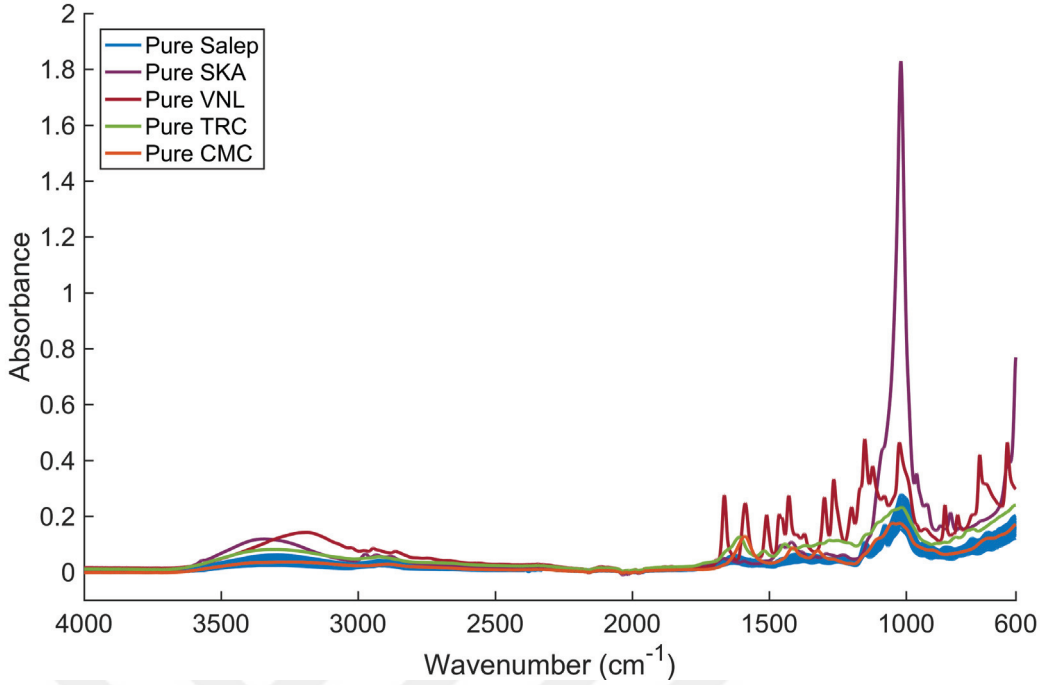

Figure 4.21. Mid-IR spectra of pure salep, SKA, VNL, TRC, and CMC samples

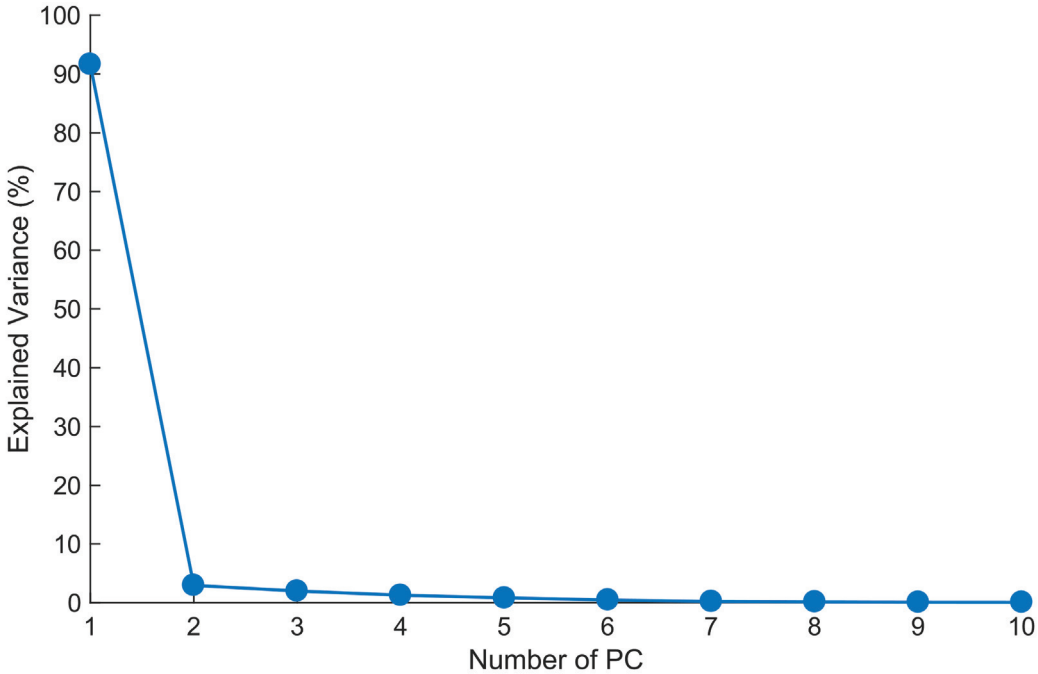

Figure 4.22. Scree plot for PCA model of pure salep samples

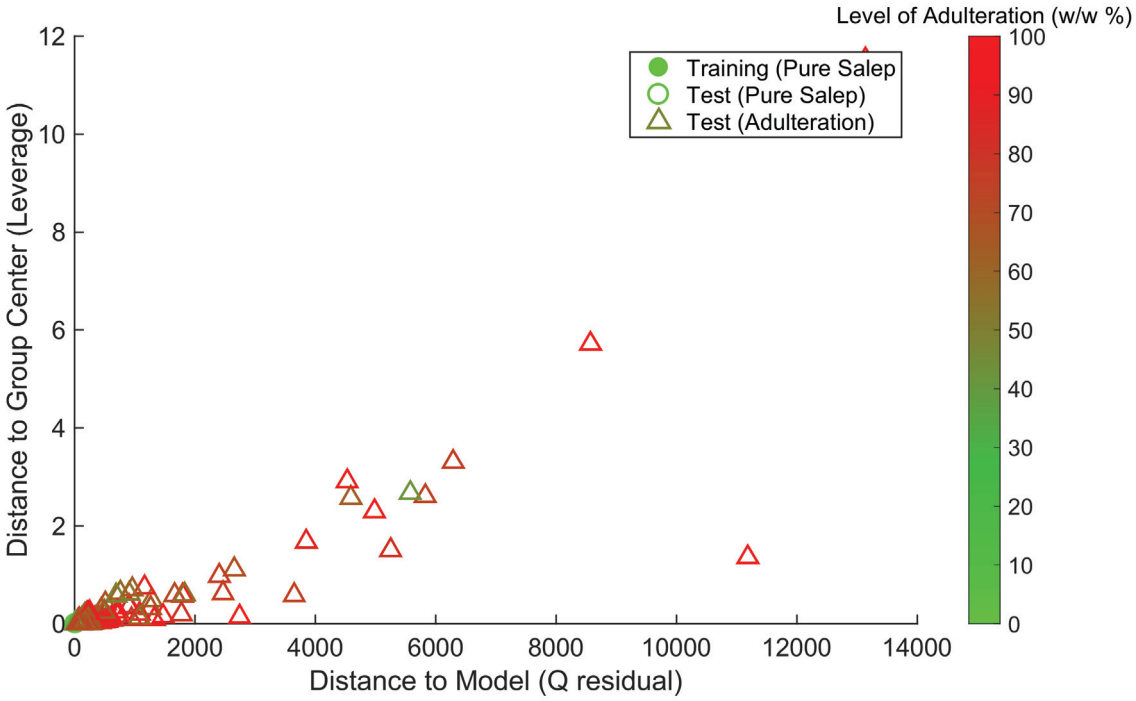

Figure 4.23. Distance to model vs. distance to group center plot obtained by SIMCA model of pure salep samples

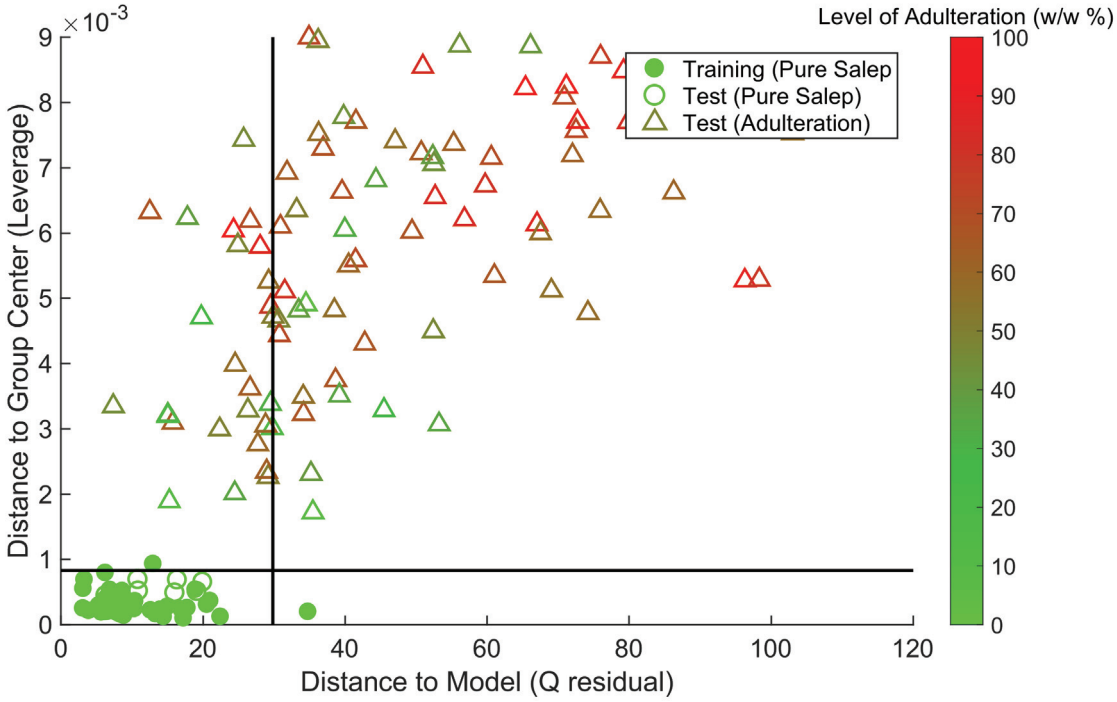

Figure 4.24. A closer look at boundaries of SIMCA plot

CHAPTER 5

CONCLUSION

In this thesis study, a new chemometric calibration toolbox, that includes ILS, PCR, PLS, GILS and Ridge methods, was developed in MATLAB programming environment. Parallelization, vectorization, and possible shortcuts for the calculations were heavily exploited to speed up both modeling and parameter tuning procedures. By involving GUI, the software was made easily accessible especially to users with no programming experience. Additionally, by delivering resulted models in the form of a single vector of regression coefficients, the prediction of new samples was simplified and made possible even without the aid of any programming language.

Despite the spectral complexity and wide variety of salep, Mid-FTIR-ATR, when empowered with chemometric tools, offers a fast and effortless way of quantitatively determining salep adulteration. Among the tested chemometric calibration techniques, GILS and Ridge regression were the best performing individual methods whereas ensemble model obtained by their average (GILS+Ridge) yielded even better results with 6.82 (w/w %) RMSEP and 0.95 R^2 value. Further evaluation of all models shows that, regardless of the calibration method of choice, reliability of a model depends heavily on whether adulterant(s) that is present in the suspected sample was included in the samples of calibration set. While 20 adulterants were involved in this study, extending the adulteration scenarios further to cover more adulterants can improve robustness as well as overall prediction performance of the model with the cost of increased sample count. Alternatively, inspecting SIMCA predictions prior to quantitative evaluation can be used to partially mitigate the unknown adulterant problem as it can at least reveal the presence of such adulterant(s) in the sample.

REFERENCES

Averett, Lacey A, Peter R Griffiths, and Koichi Nishikida. 2008. "Effective path length in attenuated total reflection spectroscopy." *Analytical chemistry* 80 (8): 3045–3049.

Brereton, Richard G. 2003. *Chemometrics: data analysis for the laboratory and chemical plant.* John Wiley & Sons.

———. 2007. *Applied chemometrics for scientists.* John Wiley & Sons.

———. 2009. *Chemometrics for pattern recognition.* John Wiley & Sons.

Cordella, Christophe BY, and Dominique Bertrand. 2014. "SAISIR: A new general chemometric toolbox." *TrAC Trends in Analytical Chemistry* 54:75–82.

Daszykowski, Michał, Sven Serneels, Krzysztof Kaczmarek, Piet Van Espen, Christophe Croux, and Beata Walczak. 2007. "TOMCAT: A MATLAB toolbox for multivariate calibration techniques." *Chemometrics and intelligent laboratory systems* 85 (2): 269–277.

De Jong, Sijmen. 1993. "SIMPLS: an alternative approach to partial least squares regression." *Chemometrics and intelligent laboratory systems* 18 (3): 251–263.

Dietterich, Thomas G. 2000. "Ensemble methods in machine learning." In *International workshop on multiple classifier systems,* 1–15. Springer.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning.* Vol. 1. 10. Springer series in statistics New York, NY, USA:

Gazi, Ehsan, John Dwyer, Peter Gardner, A Ghanbari-Siahkali, AP Wade, J Miyan, Nicholas P Lockyer, John C Vickerman, Noel W Clarke, Jonathan H Shanks, et al. 2003. "Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. A pilot study." *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 201 (1): 99–108.

Geladi, Paul, and Bruce R Kowalski. 1986. "Partial least-squares regression: a tutorial." *Analytica chimica acta* 185:1–17.

Goldberg, David E, and John H Holland. 1988. “Genetic algorithms and machine learning.” *Machine learning* 3 (2): 95–99.

Golub, Gene H, Per Christian Hansen, and Dianne P O’Leary. 1999. “Tikhonov regularization and total least squares.” *SIAM Journal on Matrix Analysis and Applications* 21 (1): 185–194.

Golub, Gene H, and Christian Reinsch. 1970. “Singular value decomposition and least squares solutions.” *Numerische mathematik* 14 (5): 403–420.

Gurdeniz, Gozde, and Banu Ozen. 2009. “Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data.” *Food chemistry* 116 (2): 519–525.

Hoerl, Arthur E, and Robert W Kennard. 1970. “Ridge regression: Biased estimation for nonorthogonal problems.” *Technometrics* 12 (1): 55–67.

Holland, John H. 1992. “Genetic algorithms.” *Scientific american* 267 (1): 66–73.

Massart, Desiré Luc, et al. 1997. *Handbook of chemometrics and qualimetrics.* Elsevier.

Özdemir, Durmus, and Erdal Dinc. 2004. “Determination of thiamine HCl and pyridoxine HCl in pharmaceutical preparations using uv–visible spectrophotometry and genetic algorithm based multivariate calibration methods.” *Chemical and pharmaceutical bulletin* 52 (7): 810–817.

Özdemir, Durmuş, and Betül Öztürk. 2007. “Near Infrared Spectroscopic Determination of Olive Oil Adulteration with Sunflower and Corn Oil.” *Journal of Food and Drug Analysis* 15 (1).

Rodionova, Oxana Ye, Anna V Titova, and Alexey L Pomerantsev. 2016. “Discriminant analysis is an inappropriate method of authentication.” *TrAC Trends in Analytical Chemistry* 78:17–22.

Rodriguez-Saona, LE, and ME Allendorf. 2011. “Use of FTIR for rapid authentication and detection of adulteration of food.” *Annual review of food science and technology* 2:467–483.

Schievano, Elisabetta, Evaristo Peggion, and Stefano Mammi. 2009. “1H nuclear magnetic resonance spectra of chloroform extracts of honey for chemometric determination of its botanical origin.” *Journal of Agricultural and Food Chemistry* 58 (1): 57–65.

Shao, Yongni, and Yong He. 2007. “Nondestructive measurement of the internal quality of bayberry juice using Vis/NIR spectroscopy.” *Journal of Food Engineering* 79 (3): 1015–1019.

Skoog, Douglas A, F James Holler, and Stanley R Crouch. 2017. *Principles of instrumental analysis.* Cengage learning.

Trygg, Johan, and Svante Wold. 2002. “Orthogonal projections to latent structures (O-PLS).” *Journal of Chemometrics: A Journal of the Chemometrics Society* 16 (3): 119–128.

Wold, Svante. 1976. “Pattern recognition by means of disjoint principal components models.” *Pattern recognition* 8 (3): 127–139.

Wold, Svante, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. 1998. “Orthogonal signal correction of near-infrared spectra.” *Chemometrics and Intelligent laboratory systems* 44 (1-2): 175–185.

Wold, Svante, Kim Esbensen, and Paul Geladi. 1987. “Principal component analysis.” *Chemometrics and intelligent laboratory systems* 2 (1-3): 37–52.

Zhao, Ming, Gerard Downey, and Colm P O’Donnell. 2014. “Detection of adulteration in fresh and frozen beefburger products by beef offal using mid-infrared ATR spectroscopy and multivariate data analysis.” *Meat science* 96 (2): 1003–1011.

APPENDIX A

GENERATED ADULTERATION SCENARIOS

Table A.1. List of generated adulteration scenarios (w/w %)

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
1	17.2	0.00	0.00	0.00	0.00	0.00	21.0	28.2	0.00	23.0	0.00	0.00	0.00	0.00	10.6	0.00	0.00	0.00	0.00	0.00	0.00
2	36.0	11.3	0.00	0.00	21.2	0.00	0.00	0.00	0.00	0.00	2.90	28.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.05	0.00	0.00	0.00	32.0	0.00	0.00	28.6	0.00	0.00	0.00	0.00	0.00	0.00	14.5	0.00	0.00	0.00	24.8	0.00	0.00
4	43.7	0.00	0.00	0.00	56.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	34.1	0.00	0.00	9.82	0.00	0.00	0.00	0.00	0.00	0.00	15.9	0.00	26.1	0.00	0.00	0.00	0.00	0.00	14.1	0.00	0.00
6	63.3	0.00	0.00	36.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	24.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	45.2	0.00	1.69	0.00	0.00	0.00	15.1	0.00	0.00	0.00	13.1
8	15.8	24.8	0.00	7.60	0.00	0.00	0.00	0.00	14.6	0.00	0.00	37.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	12.9	0.00	0.00	0.00	31.4	0.00	0.00	55.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10.8	0.00	0.00	0.00	0.00	0.00	23.7	0.00	0.00	0.00	0.00	0.00	0.00	33.9	0.00	0.00	0.00	0.00	0.00	31.6	0.00
11	9.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	29.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.3	0.00	30.5	0.00
12	18.8	0.00	0.00	40.8	0.00	0.00	0.00	0.00	0.00	40.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	17.6	0.00	27.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26.2	0.00	0.00	0.00	28.6	0.00	0.00	0.00	0.00	0.00	0.00
14	16.8	0.00	0.00	0.00	0.00	40.3	0.00	0.00	0.00	0.00	0.00	0.00	3.94	0.00	0.00	0.00	0.00	0.00	0.00	39.0	0.00
15	46.2	0.00	0.00	0.00	0.00	16.6	0.00	0.00	0.00	26.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.7
16	18.8	20.9	0.00	0.00	0.00	0.00	11.0	0.00	49.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	13.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.4	31.8	0.00	0.00
18	28.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.9	0.00	0.00	0.00	0.00	38.1	12.8	0.00	0.00	0.00	0.00	0.00	0.00
19	35.2	0.00	7.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.3	32.2	0.00	0.00	0.00	0.00	0.00
20	35.9	0.00	0.00	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.95	0.00	0.00	57.8
21	56.3	0.00	0.00	0.00	0.00	0.00	43.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	42.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	57.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
23	16.2	0.00	0.00	7.52	0.00	0.00	0.00	0.00	13.4	0.00	0.00	0.00	0.00	62.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	45.6	0.00	0.00	0.00	0.00	8.29	0.00	0.00	0.00	0.00	0.00	0.00	34.7	0.00	0.00	11.3	0.00	0.00	0.00	0.00	0.00
25	22.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.4	12.2	0.00	0.00	17.0	0.00	26.6	0.00	0.00	0.00	0.00	0.00	0.00
26	53.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.6	0.00	0.00	0.00
27	60.7	0.00	0.00	0.00	0.00	0.00	2.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.5	0.00	0.00	0.00	0.00	0.00
28	32.3	0.00	0.00	47.0	0.00	20.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29	23.5	20.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	27.6	0.00	0.00	0.00	0.00	0.00	0.00	28.4	0.00
30	25.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26.8	0.00	0.00	0.00	0.00	0.00	47.4	0.00	0.00
31	33.5	0.00	0.00	0.00	0.00	0.00	0.00	56.5	0.00	0.00	0.00	0.00	0.00	6.69	3.32	0.00	0.00	0.00	0.00	0.00	0.00
32	42.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.9	25.4	0.00	0.00	0.00	0.00
33	66.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.6	0.00
34	6.20	0.00	0.00	3.12	30.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.0	24.9
35	67.5	32.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
36	22.1	0.00	0.00	0.00	0.00	0.00	0.86	0.00	29.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.1	28.8	0.00
37	39.1	0.00	0.00	11.8	0.00	11.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.8	0.00
38	57.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	42.7	0.00	0.00
39	18.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.0	0.00	15.2	0.00	15.9	0.00	0.00	0.00	0.00	29.0	0.00	0.00
40	20.6	0.00	0.00	0.00	0.00	0.00	0.00	8.18	0.00	0.00	0.00	0.00	0.00	0.00	2.91	0.00	0.00	30.6	0.00	0.00	37.7
41	51.4	42.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
42	16.8	0.00	0.00	16.9	0.00	15.1	0.00	0.00	34.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.3	0.00
43	48.0	0.00	43.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.97	0.00	0.00
44	31.0	7.29	0.00	0.00	0.00	0.00	0.00	17.4	19.1	0.00	0.00	0.00	0.00	25.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45	43.2	0.00	0.00	0.00	0.00	31.2	0.00	0.00	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.2

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
46	27.8	24.9	0.00	0.00	0.00	0.00	0.00	0.00	24.5	0.00	0.00	0.00	22.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	14.2	0.00	0.00	28.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	32.4	0.00
48	16.4	0.00	0.00	0.00	0.00	0.00	0.00	27.0	0.00	0.00	0.00	0.00	28.9	16.9	0.00	0.00	10.8	0.00	0.00	0.00	0.00
49	77.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	39.2	0.00	34.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.59	0.00	0.00	19.5
51	50.8	0.06	0.00	0.00	0.00	0.00	0.00	23.5	0.00	0.00	0.00	25.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
52	41.9	58.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53	86.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
54	28.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	41.8	0.00	0.00	0.00	5.85	23.5	0.00	0.00	0.00	0.82	0.00
55	31.3	0.00	27.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.0	0.00	16.3	0.00	0.00
56	31.0	0.00	0.00	0.00	3.44	0.00	0.00	0.00	0.00	0.00	12.3	0.00	0.00	0.00	15.5	0.00	0.00	0.00	0.00	37.8	0.00
57	5.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.3	0.00	20.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.4	0.00
58	34.6	0.00	0.00	13.1	0.00	0.00	9.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.0	0.00	0.00	0.00	0.00	32.0
59	38.5	0.00	0.00	0.00	5.42	46.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.15	0.00	0.00	0.00	0.00	0.00
60	14.0	39.7	0.00	0.00	0.00	25.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00
61	85.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
62	34.3	0.00	0.00	0.00	65.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
63	26.2	0.00	0.00	38.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.66	0.00	0.00	0.00	0.00	29.9	0.00	0.00	0.00	0.00
64	32.6	0.00	0.00	0.00	28.4	0.00	0.00	21.4	12.5	0.00	0.00	0.00	0.00	5.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
65	17.5	0.00	0.00	41.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	41.5	0.00	0.00	0.00
66	32.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.64	0.00	0.00	0.00	9.00	0.00	31.3	20.7	0.00	0.00	0.00	0.00	0.00
67	58.7	0.00	0.00	2.13	35.7	0.00	0.00	0.00	0.00	3.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
68	48.1	0.00	0.00	0.00	0.00	0.00	0.00	51.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
69	8.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	47.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	43.7	0.00	0.00
70	35.4	0.00	0.00	0.00	0.00	0.00	64.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
71	55.9	0.00	0.00	0.00	0.00	44.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
72	37.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	62.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00
73	3.98	0.00	0.00	0.00	0.00	31.5	0.00	0.00	0.00	0.00	0.00	0.00	29.8	0.00	0.00	0.00	0.00	0.00	0.00	34.7	0.00
74	53.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.5	0.00	0.00	0.00	0.00
75	24.0	0.00	14.0	0.00	0.00	21.5	0.00	0.00	0.00	0.00	0.00	27.4	0.00	0.00	13.1	0.00	0.00	0.00	0.00	0.00	0.00
76	32.0	0.00	0.00	0.00	0.00	0.00	26.5	0.00	0.00	0.00	0.00	0.00	0.00	9.24	0.00	0.00	0.00	0.00	0.00	0.00	32.2
77	41.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	29.2	0.00	29.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
78	41.4	0.00	0.00	0.00	0.00	0.00	0.00	1.52	32.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.4	0.00	0.00	0.00	0.00
79	6.62	0.00	0.00	65.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.0	0.00	0.00	0.00	0.00
80	18.9	0.00	0.00	0.00	28.9	0.00	16.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.3	0.00	0.00	0.00	0.00	0.00
81	41.0	0.00	0.00	59.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
82	28.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.3	0.00	1.83	0.00	0.00	30.6	0.00	0.00	0.00	0.00	0.00	4.11
83	60.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.17	0.00	0.00	30.1	6.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00
84	75.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.4	0.00	0.00	0.00	0.00	0.00	0.00
85	20.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.5	0.00	0.00	0.00	0.00	0.00	9.86	0.00	27.4	23.7	0.00	0.00	0.00
86	31.4	0.00	0.00	0.00	0.00	0.00	0.00	14.1	0.00	0.00	32.1	0.00	0.00	0.00	0.00	0.00	22.4	0.00	0.00	0.00	0.00
87	60.0	0.00	0.00	40.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
88	2.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.26	58.4	0.00	0.00	0.00	0.00	0.00	0.00	37.3
89	67.3	0.00	19.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
90	41.9	0.00	0.00	0.00	0.00	0.00	0.00	2.03	0.00	0.00	0.00	0.00	0.00	0.00	45.1	10.9	0.00	0.00	0.00	0.00	0.00
91	37.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	62.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
92	37.5	0.00	32.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.5
93	63.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
94	1.23	0.00	25.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	32.8	0.00	0.00	0.00	0.00	0.00	0.00	19.4	21.1
95	29.3	0.00	0.00	0.00	16.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.8	22.8	0.00	0.00
96	53.1	0.00	0.00	46.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
97	26.6	0.00	0.00	15.1	0.00	25.5	0.00	0.00	32.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
98	2.49	0.00	0.00	0.00	32.3	0.00	0.00	0.00	0.00	0.00	0.00	28.1	0.00	23.1	0.00	0.00	0.00	0.00	14.1	0.00	0.00
99	13.9	0.00	38.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.5	0.00	0.00	0.00	34.6	0.00	0.00	0.00
100	1.69	0.00	0.00	0.00	0.00	27.7	0.00	20.5	0.00	0.00	30.6	0.00	0.00	0.00	19.5	0.00	0.00	0.00	0.00	0.00	0.00
101	66.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.6	0.00	0.00	0.00	0.00	0.00	0.00
102	51.7	22.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
103	9.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	47.1	0.00	22.1	0.00	0.00	20.9	0.00	0.00	0.00	0.00	0.00
104	7.31	0.00	0.00	0.00	0.00	0.00	0.00	22.7	26.8	0.00	0.00	0.00	0.00	11.1	0.00	32.1	0.00	0.00	0.00	0.00	0.00
105	22.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.2	0.00	0.00	0.00	28.1	0.00	9.89	0.00	0.00	23.9	0.00	0.00
106	39.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	60.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
107	28.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.5	0.00	0.00	0.00	0.00	0.00	51.0	0.00	0.00	0.00	0.00
108	11.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.9	0.00	0.00	26.6	0.00	0.00	0.00	28.1	18.0
109	23.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	38.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	38.5	0.00
110	70.9	0.00	0.00	19.2	4.26	0.00	5.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
111	43.4	0.00	0.00	56.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
112	12.0	39.1	0.00	37.5	0.00	0.00	0.00	0.00	0.00	0.00	11.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
113	36.7	0.00	12.6	0.00	28.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.4	0.00
114	63.9	0.00	31.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
115	65.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	34.6	0.00
116	7.71	0.00	0.00	0.00	26.9	0.00	0.00	15.3	0.00	0.00	0.00	21.1	0.00	29.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
117	37.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40.4	0.00	22.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
118	19.7	0.00	0.00	0.00	0.00	0.00	0.00	28.8	22.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.3	9.81	0.00	0.00	0.00
119	32.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.8	0.00	18.5	0.00	0.00	0.00	0.00	0.00	0.00	32.7
120	22.8	0.00	0.00	5.56	0.00	26.4	0.00	37.3	7.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
121	44.9	37.9	17.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
122	18.3	29.6	15.7	0.00	0.00	0.00	0.00	0.00	36.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
123	14.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	86.0	0.00	0.00	0.00	0.00	0.00	0.00
124	30.1	0.00	0.00	0.00	60.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.54	0.00	0.00	0.00	0.00	0.00
125	17.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	82.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
126	28.6	32.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.57
127	28.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	58.3	0.00	0.00	0.00	0.00	9.42	4.24	0.00
128	4.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	41.5	32.0	0.00	0.00	0.00	0.00	0.00	0.00	22.1	0.00
129	30.4	0.00	0.00	0.00	3.16	0.00	0.00	36.2	0.00	0.00	0.00	30.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
130	19.2	0.00	32.8	0.00	27.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.6	0.00	0.00	0.00	0.00	0.00	0.00
131	34.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.6	0.00	0.00	0.00	0.00	0.00	0.00	35.0	0.00	0.00	0.00	0.00	0.00
132	36.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.61	0.00	0.00	0.00	34.9	27.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
133	62.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.1	0.00	0.00	0.00	0.00	0.00
134	64.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
135	12.5	0.00	7.91	0.00	0.00	0.00	0.00	0.00	0.00	35.2	0.00	21.3	0.00	0.00	0.00	23.0	0.00	0.00	0.00	0.00	0.00
136	80.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
137	41.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.56	49.7	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
138	42.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.7	0.00	41.9	0.00	0.00	0.00	0.00	0.00
139	15.8	0.00	34.8	0.00	0.00	0.00	10.7	0.00	0.00	0.00	0.00	0.00	11.0	27.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
140	36.3	0.00	9.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.8	0.00	0.00	23.5
141	38.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	61.1	0.00	0.00	0.00	0.00	0.00
142	10.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.7	16.5	0.00	0.00	0.00	0.00	29.8	0.00	0.00	0.00	26.2
143	10.8	0.00	0.00	0.00	62.6	0.00	0.00	0.00	26.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
144	15.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	84.2	0.00	0.00	0.00	0.00	0.00
145	92.2	0.00	0.00	7.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
146	38.9	0.00	0.00	0.00	48.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
147	28.8	0.00	0.00	0.00	0.00	0.00	31.4	0.00	0.00	0.00	0.00	0.00	0.00	8.73	31.1	0.00	0.00	0.00	0.00	0.00	0.00
148	51.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	34.2	14.7	0.00	0.00	0.00	0.00	0.00	0.00
149	10.8	0.00	38.9	0.00	0.00	0.00	0.00	0.00	0.00	7.19	0.00	0.00	0.00	0.00	0.00	43.1	0.00	0.00	0.00	0.00	0.00
150	18.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.0	0.00	0.00	0.00	39.2	0.00	0.00	19.6	0.00	0.00	0.00	0.00
151	3.43	0.00	0.00	19.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.9	0.00	0.00	0.00	0.00	33.7	0.00	9.79	0.00	0.00
152	3.35	25.8	4.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.6	0.00	30.2	0.00	0.00	0.00	0.00
153	24.3	42.1	15.6	0.00	8.18	0.00	0.00	9.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
154	4.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	39.6	0.00	0.00	16.7	0.00	0.00	0.00	0.00	38.7
155	47.8	52.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
156	33.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	32.1	0.00	34.0	0.00	0.00	0.00	0.00	0.00	0.00
157	50.7	0.00	0.00	0.00	33.0	0.00	0.00	0.00	0.00	0.00	0.00	16.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
158	22.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.0	56.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
159	49.1	25.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.4	0.00	0.00	0.00
160	29.9	0.00	0.00	1.61	0.00	0.00	0.00	0.00	0.00	0.00	32.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.6	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
161	32.9	0.00	0.00	18.0	0.00	0.00	32.3	11.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.87
162	2.44	0.00	0.00	0.00	0.00	0.00	30.9	29.3	0.00	0.00	24.0	0.00	0.00	13.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
163	8.43	0.00	0.00	0.00	41.4	0.00	0.00	0.00	0.00	0.00	46.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.86	0.00
164	60.3	0.00	39.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
165	21.4	0.00	19.7	32.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.1	0.00	1.79	0.00
166	28.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
167	5.77	0.00	0.25	0.00	0.00	0.00	0.00	0.00	41.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	52.0	0.00
168	22.6	42.0	0.00	0.00	0.00	0.00	0.00	0.00	12.7	0.00	0.00	0.00	16.7	0.00	6.03	0.00	0.00	0.00	0.00	0.00	0.00
169	2.04	0.00	0.00	0.00	0.00	0.00	0.00	22.8	0.00	0.00	0.00	0.00	0.00	18.0	0.00	0.00	42.4	0.00	14.7	0.00	0.00
170	56.6	0.00	0.00	31.8	1.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
171	48.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	51.9	0.00	0.00	0.00	0.00	0.00
172	41.7	24.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.7	0.00	0.00	0.00	0.00	0.00
173	7.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.5	0.00	15.2	0.00	0.00	0.00	24.9	0.00	0.00	0.00	14.9
174	33.5	0.00	8.27	0.00	0.00	0.00	0.00	0.00	0.00	33.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.2	0.00	0.00	0.00
175	27.9	0.00	0.00	19.1	0.00	0.00	0.00	0.00	1.04	0.00	0.00	0.00	0.00	27.2	0.00	0.00	24.7	0.00	0.00	0.00	0.00
176	29.3	0.00	0.00	58.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
177	9.16	0.00	0.00	37.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	34.3	0.00	0.00	0.00	0.00	0.00	18.7	0.00	0.00	0.00
178	50.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.6	26.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
179	4.14	0.00	0.00	0.00	0.00	0.00	0.00	11.0	0.00	0.00	33.4	0.00	0.00	0.00	0.00	51.5	0.00	0.00	0.00	0.00	0.00
180	41.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.8	0.00	0.00	0.00	0.00	0.00	0.00	23.1	21.5	0.00	0.00
181	0.68	0.00	0.00	0.00	0.00	99.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
182	93.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.34	0.00	0.00	0.00	0.00
183	20.0	36.3	0.00	0.00	19.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
184	14.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.6	0.00	0.00	36.4	0.00	13.7	0.00	0.00	0.00	9.80
185	17.2	0.00	25.9	0.00	0.00	0.00	0.00	12.0	0.00	17.8	0.00	0.00	0.00	0.00	0.00	0.00	27.1	0.00	0.00	0.00	0.00
186	23.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.8	0.00	0.00	0.00	0.00	1.11	19.2	31.3	0.00	0.00
187	10.8	0.00	7.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.0	62.9
188	54.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.7	0.00	0.00	13.7	0.00	0.00	0.00	3.25
189	4.31	0.00	0.00	95.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
190	8.68	0.00	0.00	0.00	0.00	0.00	0.00	30.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	60.9	0.00	0.00
191	36.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.71	0.00	0.00	0.00	51.7	0.00	7.02
192	24.0	56.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.65	0.00	0.00	15.4	0.00	0.00	0.00
193	10.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	47.7	0.00	25.5	0.00	16.1
194	33.0	0.00	0.00	0.00	0.00	67.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
195	63.5	0.00	0.00	0.00	0.00	1.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	34.5	0.00	0.00	0.00	0.00	0.00
196	55.8	0.00	0.00	0.00	0.00	0.00	44.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
197	36.5	0.00	16.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	47.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
198	36.8	0.00	0.00	0.00	0.00	0.00	0.00	63.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
199	30.4	0.00	0.00	0.00	0.00	35.4	0.00	0.00	0.00	34.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	27.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.29	0.00	0.00	0.00	0.00	71.0
201	50.1	0.00	0.00	0.00	49.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
202	25.2	0.00	3.07	0.00	12.1	0.00	0.00	0.00	0.00	0.00	0.00	1.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	58.6	0.00
203	22.9	0.00	77.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
204	5.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	49.5	0.00	45.2
205	5.34	0.00	0.00	0.00	0.00	0.00	0.00	27.7	0.00	0.00	0.00	0.00	31.3	0.00	0.00	0.00	35.7	0.00	0.00	0.00	0.00
206	14.8	85.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
207	23.6	0.00	0.00	0.00	0.00	19.0	0.00	0.00	0.00	19.6	0.00	0.00	0.00	0.00	0.00	0.00	30.3	0.00	0.00	7.49	0.00
208	3.63	0.00	0.00	0.00	0.00	0.00	27.5	0.00	0.00	0.00	8.77	0.00	0.00	0.00	29.6	0.00	0.00	0.00	0.00	30.5	0.00
209	19.4	0.00	0.00	38.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.87	0.00	0.00	0.00	32.9	0.00	0.00	0.00	0.00
210	32.1	0.00	0.00	0.00	0.00	0.00	0.00	17.6	0.00	0.00	0.00	0.00	20.9	0.00	0.00	0.00	0.00	0.00	0.00	29.5	0.00
211	42.3	0.00	0.00	0.00	29.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.0	0.00	0.00
212	15.1	28.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
213	28.3	0.00	0.00	0.00	0.00	0.00	3.78	38.8	0.00	0.00	0.00	0.00	0.00	26.4	2.69	0.00	0.00	0.00	0.00	0.00	0.00
214	38.4	0.00	0.00	21.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.4	0.00	27.9	0.00	0.00
215	33.1	0.00	47.5	0.00	0.00	0.00	0.00	0.00	19.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
216	39.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	52.8	0.00	0.00	0.00	0.00	0.00	0.00	7.56
217	51.3	0.00	0.00	0.00	0.00	5.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.7	0.00	0.00	21.9	0.00	0.00	0.00
218	24.3	7.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.2	0.00	0.00	0.00	19.7	0.00	0.00	0.00	0.00	23.0	0.00	0.00
219	30.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.6	0.00	0.00	0.00	41.9	0.00	16.7	0.00	0.00	0.00	0.00	0.00	0.00
220	3.76	0.00	0.00	34.5	21.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.4	0.00	0.00	0.00	0.00	0.00	0.00	21.1	0.00
221	58.5	0.00	0.00	0.00	0.00	0.00	41.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
222	22.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	62.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.3
223	25.2	0.00	0.00	0.00	52.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.9	0.00	0.00	0.00	0.00	0.00
224	8.84	0.00	25.4	0.00	0.00	26.2	0.00	0.00	6.25	0.00	0.00	33.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
225	46.5	0.00	0.00	0.00	0.00	0.00	0.00	26.6	0.00	0.00	0.00	26.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
226	42.9	0.00	0.00	0.00	0.00	17.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	39.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
227	12.8	0.00	0.00	0.00	0.00	10.1	0.00	19.3	0.00	0.00	0.00	0.00	0.00	0.00	28.9	0.00	28.9	0.00	0.00	0.00	0.00
228	10.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	45.2	0.00	0.00	0.00	0.00	0.00	0.00	44.2
229	50.5	0.00	0.00	49.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
230	22.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.9	0.00	0.00	0.00	0.00	0.00	40.2	0.00	0.00
231	25.6	0.00	0.00	0.00	0.00	12.5	0.00	0.00	0.00	0.00	5.64	15.1	0.00	0.00	0.00	0.00	41.2	0.00	0.00	0.00	0.00
232	61.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	38.9	0.00
233	33.5	0.00	11.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.6	0.00	0.00	23.3
234	33.9	0.00	0.00	0.00	0.00	58.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.95	0.00	0.00
235	8.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	49.0	0.00	42.7	0.00	0.00	0.00	0.00
236	42.9	0.00	0.00	0.00	0.00	0.00	0.00	4.44	0.00	0.00	0.00	51.6	0.00	1.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
237	9.03	0.00	0.00	0.00	0.00	0.00	37.5	0.00	53.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
238	16.9	0.00	0.00	0.00	8.51	0.00	0.00	0.00	0.00	20.1	19.1	0.00	0.00	0.00	35.4	0.00	0.00	0.00	0.00	0.00	0.00
239	36.7	0.00	59.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.83	0.00	0.00	0.00	0.00
240	0.55	0.00	0.00	0.00	0.00	53.5	0.00	0.00	0.00	30.1	0.00	0.00	0.00	15.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
241	81.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
242	17.7	0.00	0.00	0.00	0.00	20.4	0.00	43.3	0.00	0.00	18.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
243	33.1	0.00	0.00	0.00	0.00	0.00	11.9	23.7	0.00	18.0	0.00	0.00	0.00	13.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
244	32.0	32.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.87	0.00	0.00	0.00	22.6	4.98	0.00	0.00	0.00	0.00	0.00	0.00
245	21.3	0.00	0.00	0.00	77.0	0.00	0.00	0.00	1.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
246	36.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.00	43.7	0.00	0.00	0.00	0.00	0.00	0.00	14.7	4.71
247	33.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.1	0.00	10.4	0.00	0.00	0.00	0.00
248	17.0	0.00	0.00	0.00	29.8	21.0	0.00	12.6	0.00	0.00	19.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
249	53.2	36.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.3	0.00
250	13.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.14	0.00	0.00	0.00	0.00	66.3	0.00	17.9	0.00	0.00	0.00
251	6.11	0.00	27.1	0.00	0.00	7.49	7.00	0.00	0.00	52.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
252	44.2	0.00	0.00	18.8	0.00	0.00	0.00	0.00	0.00	25.2	0.00	0.00	0.00	11.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
253	36.3	0.00	0.00	0.00	0.00	0.00	36.3	0.00	0.00	0.00	27.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
254	16.6	0.00	0.00	0.00	0.00	0.00	0.00	83.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
255	22.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.4	0.00	0.00	0.00	0.00	55.1	0.00	0.00	0.00	0.00	0.00	0.00
256	8.10	0.00	0.00	0.00	0.00	0.00	8.15	0.00	0.00	0.00	0.00	0.00	0.00	34.6	0.00	0.00	0.00	49.2	0.00	0.00	0.00
257	40.6	0.00	25.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.1	0.00	4.83	0.00	0.00	6.40	0.00	0.00	0.00	0.00	0.00
258	17.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	29.8	52.8	0.00	0.00	0.00	0.00
259	71.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.4
260	36.1	0.00	21.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.2	0.00	0.00	0.00	0.00	0.00	31.4	0.00	0.00	0.00
261	17.7	0.00	0.00	0.00	30.0	0.00	0.00	0.00	0.00	0.00	32.0	0.00	0.00	0.00	0.00	20.4	0.00	0.00	0.00	0.00	0.00
262	34.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.9
263	28.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.0	0.00	0.00	0.02	28.6	0.00	0.00	0.00	17.7	0.00
264	22.4	0.00	25.3	0.00	0.00	3.18	0.00	0.00	0.00	19.8	29.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
265	11.4	0.00	0.00	0.00	0.00	0.00	0.00	30.0	0.00	0.00	0.00	0.00	29.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.8
266	27.1	0.00	0.00	0.00	48.5	0.00	11.2	0.00	0.00	0.00	0.00	0.00	13.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
267	22.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	21.9	0.00	0.00	23.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	32.5	0.00
268	29.4	14.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.62	26.8	0.00	27.8	0.00	0.00	0.00
269	34.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	65.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
270	35.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.9	0.00	0.00	18.9	0.00	0.00	0.00	0.00	0.00	0.00	30.1	0.00
271	48.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	34.0	0.00	0.00	0.00	17.8
272	3.06	0.00	0.00	0.00	0.00	0.00	22.1	0.00	0.00	10.3	0.00	0.00	0.00	0.00	0.00	0.00	26.2	0.00	0.00	0.00	38.3
273	16.8	0.00	0.00	0.00	0.00	83.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
274	15.4	0.00	0.00	0.00	43.0	0.00	0.00	0.00	0.00	41.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
275	35.4	16.4	0.00	0.00	0.00	0.00	0.00	0.00	2.07	0.00	0.00	0.00	21.2	0.00	25.0	0.00	0.00	0.00	0.00	0.00	0.00

(cont. on next page)

Table A.1 – *(cont.)*

#	SLP	GUM	GAR	KBG	KJG	BNT	MNT	PAN	PIN	SKR	YST	VNL	TRC	OMK	OPR	OBL	BUN	GLT	SKA	CMC	KTR
276	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.1	0.00	0.00	31.0	23.6
277	52.6	0.00	0.00	0.00	0.00	0.00	0.00	47.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
278	35.3	0.00	0.00	0.00	0.00	0.00	0.00	35.7	0.00	29.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
279	46.6	48.2	0.00	0.00	4.36	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
280	96.5	0.00	0.00	0.00	0.00	3.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
281	30.9	0.00	35.2	0.00	0.00	10.2	0.00	7.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.7	0.00	0.00	0.00
282	6.83	0.00	9.51	0.00	0.00	0.00	0.00	34.0	0.00	0.00	24.8	0.00	0.00	0.00	0.00	0.00	0.00	24.8	0.00	0.00	0.00
283	32.9	15.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	51.5	0.00	0.00	0.00
284	16.2	0.00	0.00	0.00	0.00	0.00	0.00	16.2	0.00	0.00	0.00	0.00	6.82	0.00	0.00	60.8	0.00	0.00	0.00	0.00	0.00
285	39.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.5	0.00	0.00	0.00	5.50	0.00	0.00	0.00	19.6	0.00	0.00	0.00
286	51.8	0.00	0.00	0.00	48.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
287	31.6	54.5	0.00	13.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
288	22.5	0.00	0.00	0.00	0.00	18.1	0.00	0.00	0.00	0.00	0.00	34.9	0.00	0.00	0.00	0.00	3.11	0.00	0.00	0.00	21.3
289	51.4	0.00	0.00	0.00	0.00	0.00	0.00	24.3	10.8	0.00	0.00	0.00	0.00	5.20	0.00	8.29	0.00	0.00	0.00	0.00	0.00
290	50.2	0.00	0.00	0.00	0.00	11.9	0.00	0.00	0.00	0.00	0.00	37.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00