CHEM 353 INTRODUCTION TO CHEMOMETRICS FINAL EXAM

Nurefşan Nazlı Yüzükırmızı - 280102002

JANUARY 16, 2025

İZMİR INSTITUTE OF TECHNOLOGY

Table of Contents

Article 1: Phenolic Composition Analysis of Wine Samples

Abstract

This study investigates the phenolic composition of 178 wine samples from three distinct geographical regions using Principal Component Analysis (PCA). The dataset comprises 13 phenolic compounds (measured in ppm) that influence the sensory and chemical properties of wines. Randomly selected subsets of 30 samples from each region were analyzed to identify patterns and groupings based on phenolic profiles. PCA effectively reduced data dimensionality while retaining key variability, allowing the visualization of distinct clusters for the regions. While most samples were classified correctly, some overlap between regions was observed, particularly between regions B and C, likely due to similar environmental or viticultural conditions. This study highlights PCA's potential for regional wine profiling and its application in improving quality control and understanding phenolic composition in winemaking.

Nurefşan Nazlı YÜZÜKIRMIZI

Introduction

Phenolic compounds play a critical role in defining the sensory and chemical characteristics of wine, influencing its taste, aroma, color, and potential health benefits. These secondary metabolites are shaped by factors such as grape variety, cultivation methods, and the environmental conditions specific to each region. Analyzing the distribution of phenolic compounds provides valuable insights for viticulture and winemaking, enhancing quality control processes and meeting consumer expectations.

To analyze these complex datasets, multivariate techniques like Principal Component Analysis (PCA) are particularly effective. PCA reduces data dimensionality while preserving the variability critical to interpretation. In this study, PCA is applied to classify wine samples from three geographical regions based on their phenolic profiles, identify key compounds contributing to regional differentiation, and assess whether a reduced dataset could achieve similar classification accuracy. The results offer practical insights for improving wine profiling and optimizing winemaking practices.

Methodology and DATA's:

Phenolic compounds, as secondary metabolites derived from plants, significantly influence the sensory and chemical attributes of wine. Understanding their distribution is essential for profiling wines from different regions and enhancing overall wine quality control.

This study applied Principal Component Analysis (PCA) to a dataset of phenolic composition from wine samples to uncover patterns and classify samples based on their phenolic profiles. The PCA approach facilitated the visualization of key relationships between samples and compounds, enabling the identification of regional differences and potential overlaps.

Data Collection and Sampling

The dataset consisted of phenolic composition data from 178 wine samples distributed as follows:

- Region A: 59 samples
- Region B: 71 samples
- Region C: 48 samples

To ensure balanced representation and statistical reliability, 30 samples were randomly selected from each region using the following steps:

1. Loaded the dataset into Excel.
2. Used the “Data Analysis > Random Number Generator” feature to randomly sort and select 30 samples per region.

No	CLAS	Type	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
59	1	A	13.72	1.43	2.5	16.7	108	3.4	3.67	0.19	2.04	6.8	0.89	2.87	1285
33	1	A	13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97	3.84	1.23	2.87	990
27	1	A	13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45	4.8	0.92	3.22	1195
42	1	A	13.41	3.84	2.12	18.8	90	2.45	2.68	0.27	1.48	4.28	0.91	3	1035
3	1	A	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
51	1	A	13.05	1.73	2.04	12.4	92	2.72	3.27	0.17	2.91	7.2	1.12	2.91	1150
46	1	A	14.21	4.04	2.44	18.9	111	2.85	2.65	0.3	1.25	5.24	0.87	3.33	1080
49	1	A	14.1	2.02	2.4	18.8	103	2.75	2.92	0.32	2.38	6.2	1.07	2.75	1060
34	1	A	13.76	1.53	2.7	19.5	132	2.95	2.74	0.5	1.35	5.4	1.25	3	1235
16	1	A	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
55	1	A	13.74	1.67	2.25	16.4	118	2.6	2.9	0.21	1.62	5.85	0.92	3.2	1060
57	1	A	14.22	1.7	2.3	16.3	118	3.2	3	0.26	2.03	6.38	0.94	3.31	970
7	1	A	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
35	1	A	13.51	1.8	2.65	19	110	2.35	2.53	0.29	1.54	4.2	1.1	2.87	1095
13	1	A	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
56	1	A	13.56	1.73	2.46	20.5	116	2.96	2.78	0.2	2.45	6.25	0.98	3.03	1120
32	1	A	13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95	6.9	1.09	2.88	1515
5	1	A	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
9	1	A	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
18	1	A	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130
11	1	A	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
30	1	A	14.02	1.68	2.21	16	96	2.65	2.33	0.26	1.98	4.7	1.04	3.59	1035
12	1	A	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
39	1	A	13.07	1.5	2.1	15.5	98	2.4	2.64	0.28	1.37	3.7	1.18	2.69	1020
44	1	A	13.24	3.98	2.29	17.5	103	2.64	2.63	0.32	1.66	4.36	0.82	3	680
23	1	A	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.8	1.11	4	1035
6	1	A	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
24	1	A	12.85	1.6	2.52	17.8	95	2.48	2.37	0.26	1.46	3.93	1.09	3.63	1015
17	1	A	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
58	1	A	13.29	1.97	2.68	16.8	102	3	3.23	0.31	1.66	6	1.07	2.84	1270
58	2	B	11.82	1.47	1.99	20.8	86	1.98	1.6	0.3	1.53	1.95	0.95	3.33	495
15	2	B	12.99	1.67	2.6	30	139	3.3	2.89	0.21	1.96	3.35	1.31	3.5	985
6	2	B	12.17	1.45	2.53	19	104	1.89	1.75	0.45	1.03	2.95	1.45	2.23	355
48	2	B	12.25	1.73	2.12	19	80	1.65	2.03	0.37	1.63	3.4	1	3.17	510
11	2	B	12.21	1.19	1.75	16.8	151	1.85	1.28	0.14	2.5	2.85	1.28	3.07	718
18	2	B	13.03	0.9	1.71	16	86	1.95	2.03	0.24	1.46	4.6	1.19	2.48	392
30	2	B	11.64	2.06	2.46	21.6	84	1.95	1.69	0.48	1.35	2.8	1	2.75	680
71	2	B	12.04	4.3	2.38	22	80	2.1	1.75	0.42	1.35	2.6	0.79	2.57	580
54	2	B	11.76	2.68	2.92	20	103	1.75	2.03	0.6	1.05	3.8	1.23	2.5	607
44	2	B	12.34	2.45	2.46	21	98	2.56	2.11	0.34	1.31	2.8	0.8	3.38	438
59	2	B	12.42	1.61	2.19	22.5	108	2	2.09	0.34	1.61	2.06	1.06	2.96	345
40	2	B	12.37	1.07	2.1	18.5	88	3.52	3.75	0.24	1.95	4.5	1.04	2.77	660
33	2	B	12	1.51	2.42	22	86	1.45	1.25	0.5	1.63	3.6	1.05	2.65	450
46	2	B	12.51	1.73	1.98	20.5	85	2.2	1.92	0.32	1.48	2.94	1.04	3.57	672
27	2	B	12.67	0.98	2.24	18	99	2.2	1.94	0.3	1.46	2.62	1.23	3.16	450
25	2	B	13.05	3.86	2.32	22.5	85	1.65	1.59	0.61	1.62	4.8	0.84	2.01	515
41	2	B	12.29	3.17	2.21	18	88	2.85	2.99	0.45	2.81	2.3	1.42	2.83	406
50	2	B	12.22	1.29	1.94	19	92	2.36	2.04	0.39	2.08	2.7	0.86	3.02	312
62	2	B	11.45	2.4	2.42	20	96	2.9	2.79	0.32	1.83	3.25	0.8	3.39	625
32	2	B	12.08	1.83	2.32	18.5	81	1.6	1.5	0.52	1.64	2.4	1.08	2.27	480
38	2	B	11.81	2.12	2.74	21.5	134	1.6	0.99	0.14	1.56	2.5	0.95	2.26	625
67	2	B	12.07	2.16	2.17	21	85	2.6	2.65	0.37	1.35	2.76	0.86	3.28	378
13	2	B	13.86	1.51	2.67	25	86	2.95	2.86	0.21	1.87	3.38	1.36	3.16	410
24	2	B	12.08	1.13	2.51	24	78	2	1.58	0.4	1.4	2.2	1.31	2.72	630
42	2	B	12.08	2.08	1.7	17.5	97	2.23	2.17	0.26	1.4	3.3	1.27	2.96	710
52	2	B	11.46	3.74	1.82	19.5	107	3.18	2.58	0.24	3.58	2.9	0.75	2.81	562
19	2	B	11.84	2.89	2.23	18	112	1.72	1.32	0.43	0.95	2.65	0.96	2.52	500
69	2	B	11.79	2.13	2.78	28.5	92	2.13	2.24	0.58	1.76	3	0.97	2.44	466
7	2	B	12.37	1.21	2.56	18.1	98	2.42	2.65	0.37	2.08	4.6	1.19	2.3	678
53	2	B	12.52	2.43	2.17	21	88	2.55	2.27	0.26	1.22	2	0.9	2.78	325
2	3	C	12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530
21	3	C	13.5	3.12	2.62	24	123	1.4	1.57	0.22	1.25	8.6	0.59	1.3	500
38	3	C	12.82	3.37	2.3	19.5	88	1.48	0.66	0.4	0.97	10.26	0.72	1.75	685
48	3	C	14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560
37	3	C	13.45	3.7	2.6	23	111	1.7	0.92	0.43	1.46	10.68	0.85	1.56	695
3	3	C	12.81	2.31	2.4	24	98	1.15	1.09	0.27	0.83	5.7	0.66	1.36	560
28	3	C	12.45	3.03	2.64	27	97	1.9	0.58	0.63	1.14	7.5	0.67	1.73	880
16	3	C	13.16	3.57	2.15	21	102	1.5	0.55	0.43	1.3	4	0.6	1.68	830
4	3	C	12.7	3.55	2.36	21.5	106	1.7	1.2	0.17	0.84	5	0.78	1.29	600
32	3	C	13.69	3.26	2.54	20	107	1.83	0.56	0.5	0.8	5.88	0.96	1.82	680
12	3	C	13.36	2.56	2.35	20	89	1.4	0.5	0.37	0.64	5.6	0.7	2.47	780
27	3	C	13.84	4.12	2.38	19.5	89	1.8	0.83	0.48	1.56	9.01	0.57	1.64	480
43	3	C	14.16	2.51	2.48	20	91	1.68	0.7	0.44	1.24	9.7	0.62	1.71	660
35	3	C	13.78	2.76	2.3	22	90	1.35	0.68	0.41	1.03	9.58	0.7	1.68	615
1	3	C	12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
5	3	C	12.51	1.24	2.25	17.5	85	2	0.58	0.6	1.25	5.45	0.75	1.51	650
18	3	C	12.87	4.61	2.48	21.5	86	1.7	0.65	0.47	0.86	7.65	0.54	1.86	625
36	3	C	13.73	4.36	2.26	22.5	88	1.28	0.47	0.52	1.15	6.62	0.78	1.75	520
26	3	C	13.17	5.19	2.32	22	93	1.74	0.63	0.61	1.55	7.9	0.6	1.48	725
14	3	C	13.62	4.95	2.35	20	92	2	0.8	0.47	1.02	4.4	0.91	2.05	550
34	3	C	12.96	3.45	2.35	18.5	106	1.39	0.7	0.4	0.94	5.28	0.68	1.75	675
13	3	C	13.52	3.17	2.72	23.5	97	1.55	0.52	0.5	0.55	4.35	0.89	2.06	520
9	3	C	13.49	3.59	2.19	19.5	88	1.62	0.48	0.58	0.88	5.7	0.81	1.82	580
42	3	C	12.77	2.39	2.28	19.5	86	1.39	0.51	0.48	0.64	9.9	0.57	1.63	470
20	3	C	13.08	3.9	2.36	21.5	113	1.41	1.39	0.34	1.14	9.4	0.57	1.33	550
46	3	C	13.27	4.28	2.26	20	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
24	3	C	13.23	3.3	2.28	18.5	98	1.8	0.83	0.61	1.87	10.52	0.56	1.51	675
44	3	C	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
15	3	C	12.25	3.88	2.2	18.5	112	1.38	0.78	0.29	1.14	8.21	0.65	2	855
19	3	C	13.32	3.24	2.38	21.5	92	1.93	0.76	0.45	1.25	8.42	0.55	1.62	650

PCA Procedure

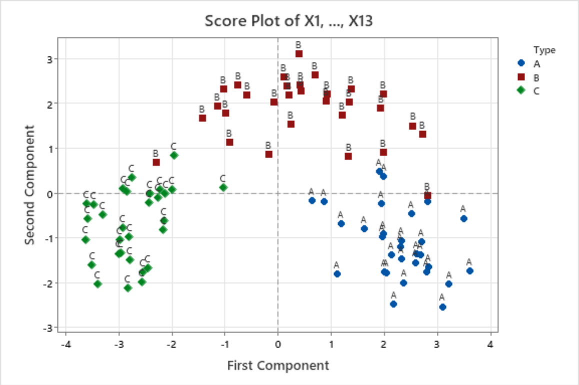
PCA was implemented to reduce dimensionality while retaining most of the data's variability. The steps included:

1. In Minitab, selecting **Stat > Multivariate > Principal Components Analysis**.
2. Using 13 phenolic compounds (e.g., X1 hrough X13) as variables.
3. Generating score plots, loading plots, and explained variance tables.
4. Interpreting eigenvalues to assess variance explained by each principal component.
5. Visualizing score and loading plots, along with 3D scatter plots, for interpretation.

Results and Interpretation

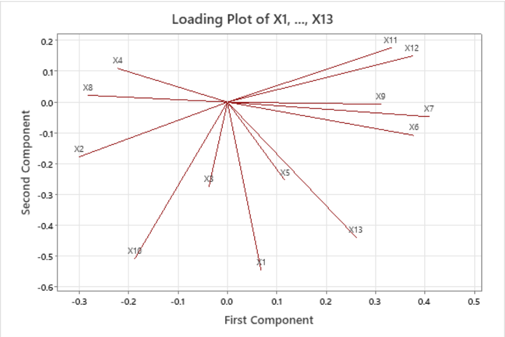
Score Plot Analysis

The PCA score plot revealed distinct clusters for regions A, B, and C. Region A samples were well-separated, indicating unique phenolic characteristics. However, overlap between regions B and C suggested shared phenolic profiles, likely influenced by environmental similarities.



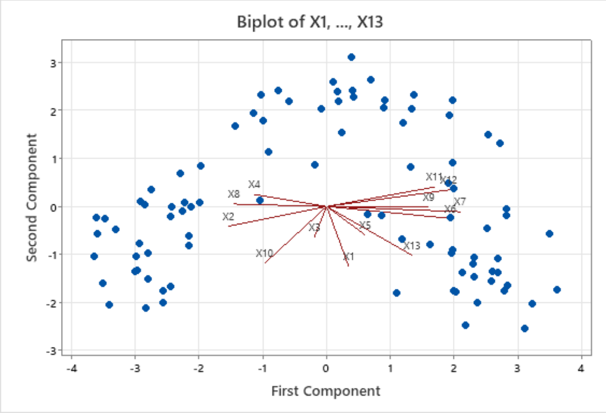
Loading Plot Analysis

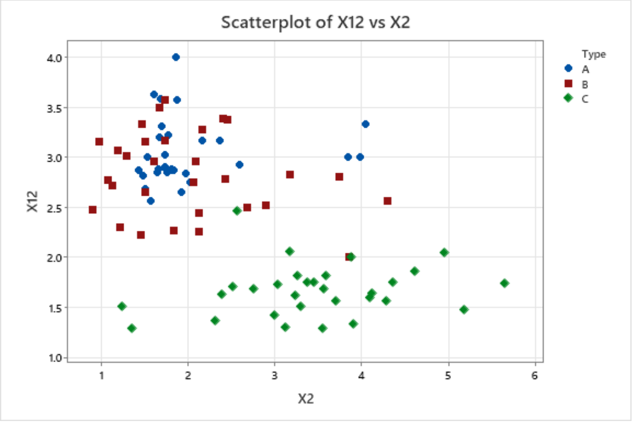
Phenolic compounds X3, X5, and X9 were identified as key contributors to regional separation, based on their high loadings.

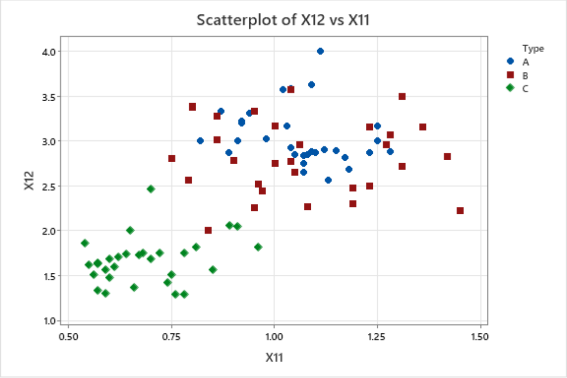


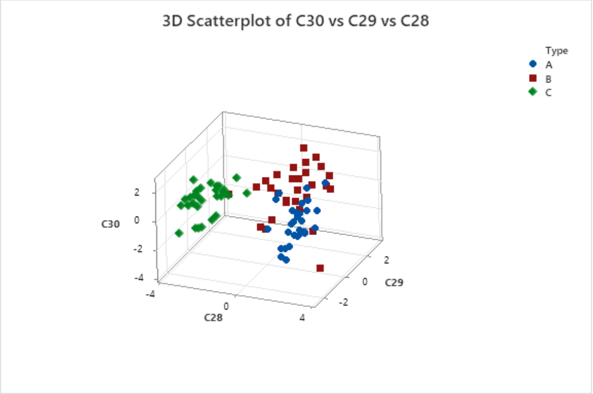
Classification Accuracy

Using the first two principal components, most samples were correctly grouped. However, the overlap between regions A and B resulted in some misclassifications.









Discussion

The results demonstrate the power of PCA in reducing data dimensionality while preserving essential variability. The distinct separation of region A highlights unique phenolic profiles, potentially influenced by environmental or viticultural factors. The overlap between regions B and C suggests similar growing conditions or winemaking practices, which warrants further investigation. Simplifying the model by using fewer phenolic compounds was computationally efficient but reduced classification accuracy, emphasizing the importance of retaining a comprehensive dataset for robust analysis.

Future studies could incorporate additional variables, such as soil composition, climate data, and specific winemaking techniques, to better understand the factors influencing phenolic profiles. Supervised classification methods, like discriminant analysis, might also improve regional classification accuracy.

Conclusion

This study demonstrated the effectiveness of Principal Component Analysis (PCA) in classifying wine samples based on their phenolic compositions. The analysis revealed distinct phenolic profiles for region C, while regions B and A showed significant overlap, likely due to shared environmental or viticultural factors. PCA successfully reduced data dimensionality, enabling the identification of key compounds (X2, X12, and X1) responsible for regional differentiation.

Although simplifying the model by focusing on fewer phenolic compounds increased computational efficiency, it also reduced classification accuracy, emphasizing the importance of maintaining a comprehensive dataset for robust results. The findings underscore the potential of PCA as a valuable tool in wine research, improving quality control and providing insights into regional phenolic profiles. Future work should incorporate additional variables and advanced classification techniques to enhance accuracy and further explore the factors contributing to regional differences.

Article 2: Multivariate Calibration of Biodiesel Mixtures

Abstract

This study explores the application of multivariate calibration models to predict the concentrations of sunflower oil, biodiesel, and petroleum diesel in ternary mixtures using FTIR (Fourier Transform Infrared) absorbance spectra. Partial Least Squares (PLS) regression was utilized to build predictive models, and its performance was compared with simpler linear regression techniques. The analysis demonstrated that PLS regression effectively captures multivariate relationships, achieving low error rates in concentration predictions. Key metrics such as Standard Error of Cross-Validation (SECV) and Standard Error of Prediction (SEP) highlighted the superiority of PLS regression. This approach showcases the potential of advanced chemometric techniques for biodiesel and petroleum diesel quality assurance.

Nurefşan Nazlı YÜZÜKIRMIZI

Introduction

As global energy demands grow, biodiesel has emerged as a sustainable alternative to petroleum-based fuels. Mixtures of biodiesel, petroleum diesel, and plant oils such as sunflower oil are widely used to optimize fuel properties. Accurate determination of component concentrations in these mixtures is critical for ensuring fuel quality and compliance with regulatory standards. Fourier Transform Infrared (FTIR) spectroscopy offers a rapid and non-destructive method for analyzing such mixtures.

This study investigates the use of Partial Least Squares (PLS) regression to predict the concentrations of sunflower oil, biodiesel, and petroleum diesel in ternary mixtures. PLS regression is a robust multivariate technique capable of handling complex datasets and capturing latent relationships between variables. Additionally, the performance of PLS regression is compared with that of simple linear regression to evaluate the benefits of using advanced chemometric approaches.

Methodology

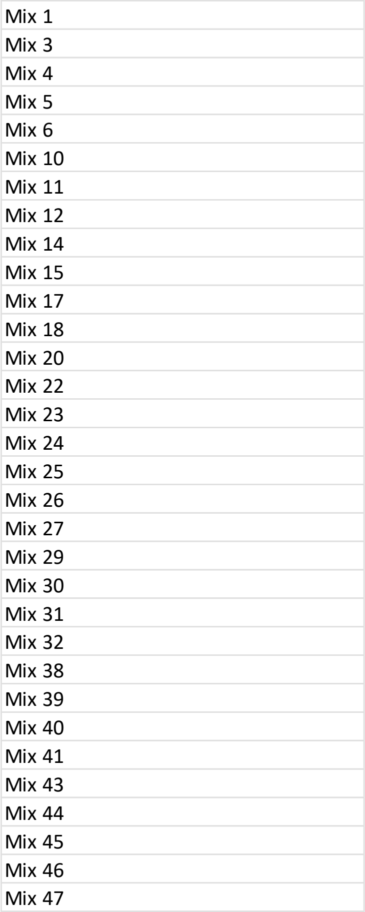
Dataset

The dataset included FTIR absorbance spectra for 47 ternary mixtures and three pure components (sunflower oil, biodiesel, and petroleum diesel). A calibration set of 32 mixtures and pure components was used to build the model, while the remaining 15 mixtures formed the validation set.

Steps to Split the Dataset

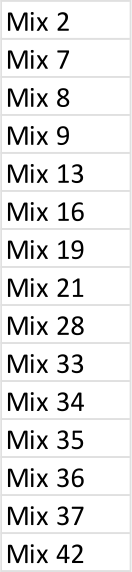
1. The dataset was loaded into Excel.
2. The “Data Analysis > Random Number Generator” feature was used to randomly split the data into calibration and validation sets.
3. Calibration samples were plotted to ensure they spanned the full concentration range of all three components.

Calibration set:



Mix 1
Mix 3
Mix 4
Mix 5
Mix 6
Mix 10
Mix 11
Mix 12
Mix 14
Mix 15
Mix 17
Mix 18
Mix 20
Mix 22
Mix 23
Mix 24
Mix 25
Mix 26
Mix 27
Mix 29
Mix 30
Mix 31
Mix 32
Mix 38
Mix 39
Mix 40
Mix 41
Mix 43
Mix 44
Mix 45
Mix 46
Mix 47

Validation Set:



Mix 2
Mix 7
Mix 8
Mix 9
Mix 13
Mix 16
Mix 19
Mix 21
Mix 28
Mix 33
Mix 34
Mix 35
Mix 36
Mix 37
Mix 42

PLS Regression Procedure

1. In Minitab, **Stat** > **Regression** > **Partial Least Squares** was selected.
2. FTIR absorbance spectra were assigned as predictors, and concentrations of sunflower oil, biodiesel, and petroleum diesel were set as responses.
3. Cross-validation determined the optimal number of components for the model.
4. Residual plots, actual vs. predicted plots, and model summary statistics were generated for evaluation.

Linear Regression for Comparison

A simple linear regression model was applied to predict the concentration of each component using a single spectral variable, providing a benchmark for PLS performance.

Results and Interpretation

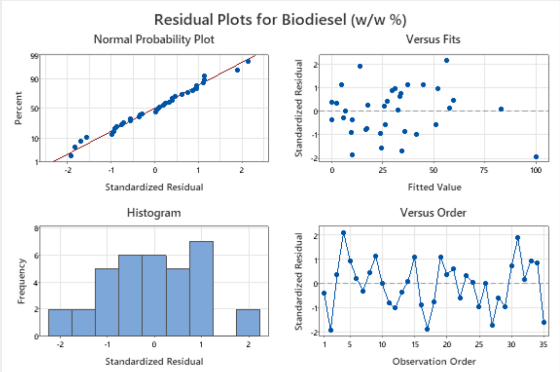
PLS Model Performance

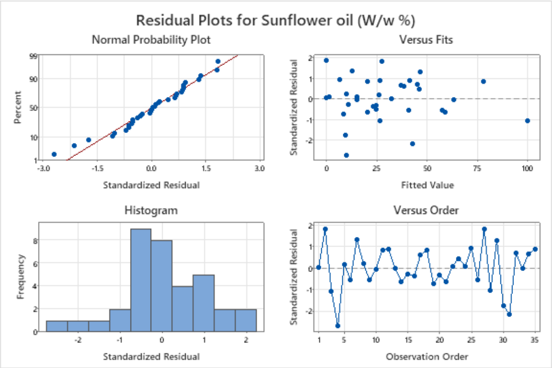
- **Sunflower Oil**: SECV = 2.40, SEP = 1.86
- **Biodiesel**: SECV = 1.86, SEP = 0.73
- **Petroleum Diesel**: SECV = 2.86, SEP = 2.13

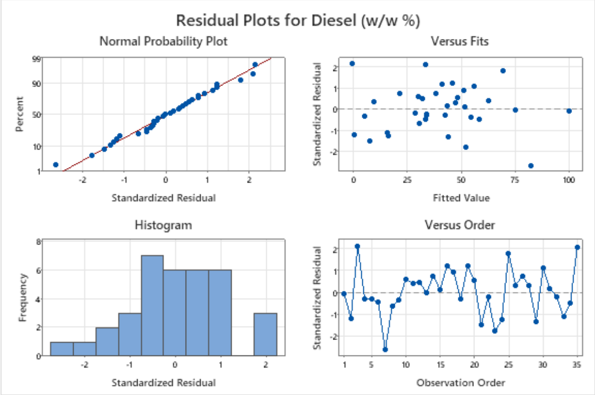
PLS regression provided accurate predictions for all three components, with low SECV and SEP values. Residual plots showed no systematic patterns, confirming model validity. Actual vs. predicted plots indicated strong linear relationships for all components.

Supporting Graphs and Visualizations

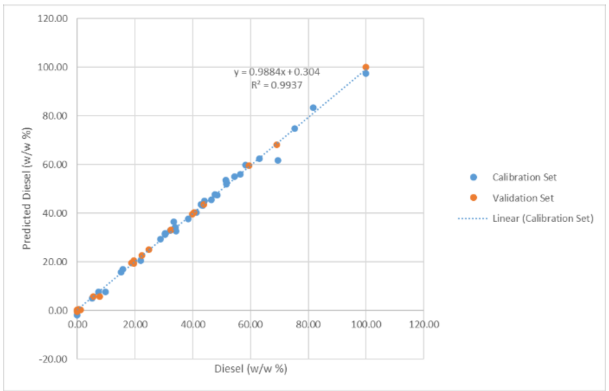
1. *Residual Plots:* Confirmed randomness and absence of bias.

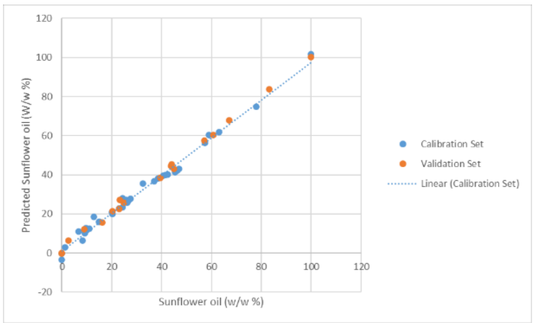


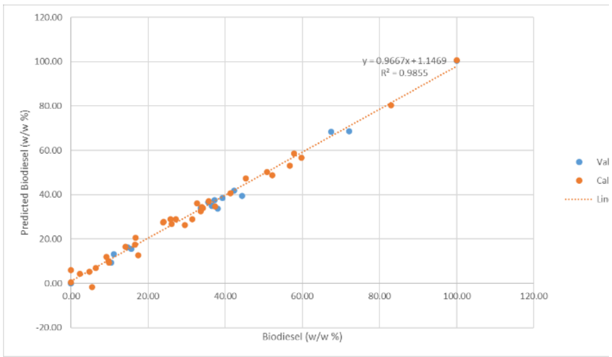




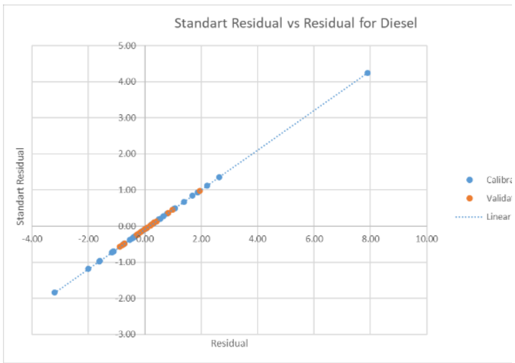
2. *Actual vs. Predicted Plots:* Demonstrated strong prediction accuracy for PLS.

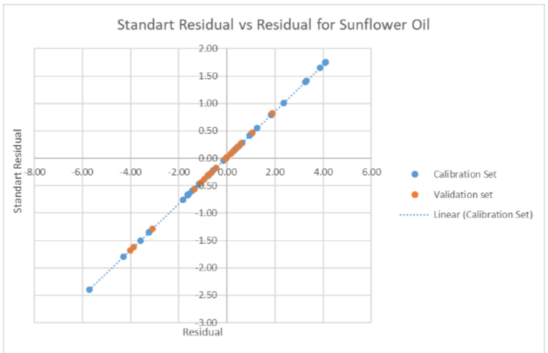


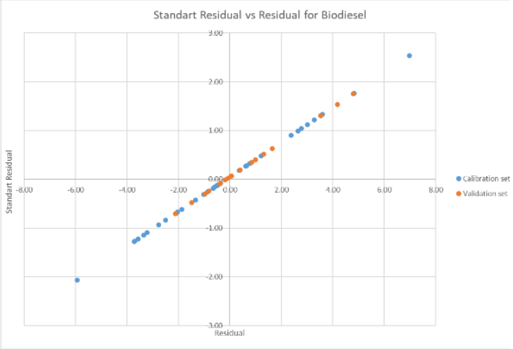




3. *Residual vs Standard Residual Plots:* Demonstrated strong prediction accuracy for PLS.







Simple Linear Regression Comparison

- **Sunflower Oil**: Higher error rates compared to PLS regression.
- **Biodiesel**: Significant limitations in capturing multivariate interactions.
- **Petroleum Diesel**: Errors were substantially higher, emphasizing the need for advanced methods like PLS regression.

Discussion

The results highlight the effectiveness of PLS regression in predicting the concentrations of sunflower oil, biodiesel, and petroleum diesel in ternary mixtures. PLS regression successfully captured latent multivariate relationships, providing low prediction errors across all components. In contrast, simple linear regression failed to account for the complex interactions between spectral variables and component concentrations, resulting in higher errors.

For petroleum diesel, PLS regression achieved SECV and SEP values of 2.86 and 2.13, respectively. These values, while slightly higher than those for biodiesel and sunflower oil, still indicate robust predictive performance. The challenges associated with diesel's spectral complexity highlight the importance of using advanced methods like PLS regression.

Future research could explore nonlinear calibration techniques, such as support vector machines or neural networks, to further enhance model accuracy. Incorporating additional variables, such as temperature and sample preparation conditions, may also improve prediction performance.

Conclusion

This study highlights the advantages of PLS regression in analysing FTIR spectra for ternary mixtures of biodiesel, sunflower oil, and petroleum diesel. Compared to basic linear regression models, PLS demonstrated superior precision and reliability, confirming its value as an analytical tool in fuel quality control. While the models were highly effective for sunflower oil and biodiesel, further optimization is required to address the challenges posed by petroleum diesel's complex spectral data. These findings reinforce the importance of adopting advanced chemometric methods in analytical chemistry to enhance the accuracy and efficiency of fuel analysis.