Part 1: Leveraging Machine Learning and Hyperspectral Imaging for Predicting Adulteration in Ground Arabica Coffee with Robusta: A Focus on Spectral Preprocessing

AUTHOR
Derick Malavi

Introduction

Spectral preprocessing plays a critical role in hyperspectral imaging by enhancing data quality and interpretability for machine learning models. Raw spectral data frequently exhibit noise, baseline shifts, and scattering effects, which obscure differences between pure and adulterated samples, complicating accurate classification.

Techniques such as scatter corrections, smoothing, and derivatives directly address these interferences, improving the signal-to-noise ratio and preserving essential chemical and physical information. These enhancements enable more reliable predictions in detecting adulteration, such as Robusta in Arabica coffee.

This study applied Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and their combination with Savitzky-Golay derivatives (first or second order) using a 13-point window (6 left-side, 6 right-side, and 1 central point) to improve model performance.

```
# Load packages from the library
warning = FALSE
  suppressWarnings(suppressMessages({
  library(ggplot2)
  library(dplyr)
  library(readxl)
  library(readr)
  library(janitor)
  library(mdatools)
  library(rmarkdown)
  library(knitr)
  library(quarto)
  library(pander)
  }))
```

```
# Set working directory

setwd("C:/Users/abc/OneDrive - UGent/Documenten/Derick Malavi_PhD Docs_UGent/Manuscript 4_Coff
```

```
# Load or import our coffee raw file--->unprocessed

coffee_data_raw <- read.csv('coffee_data_pycleaned.csv')

dim(coffee_data_raw) # Check the number of rows and columns
```

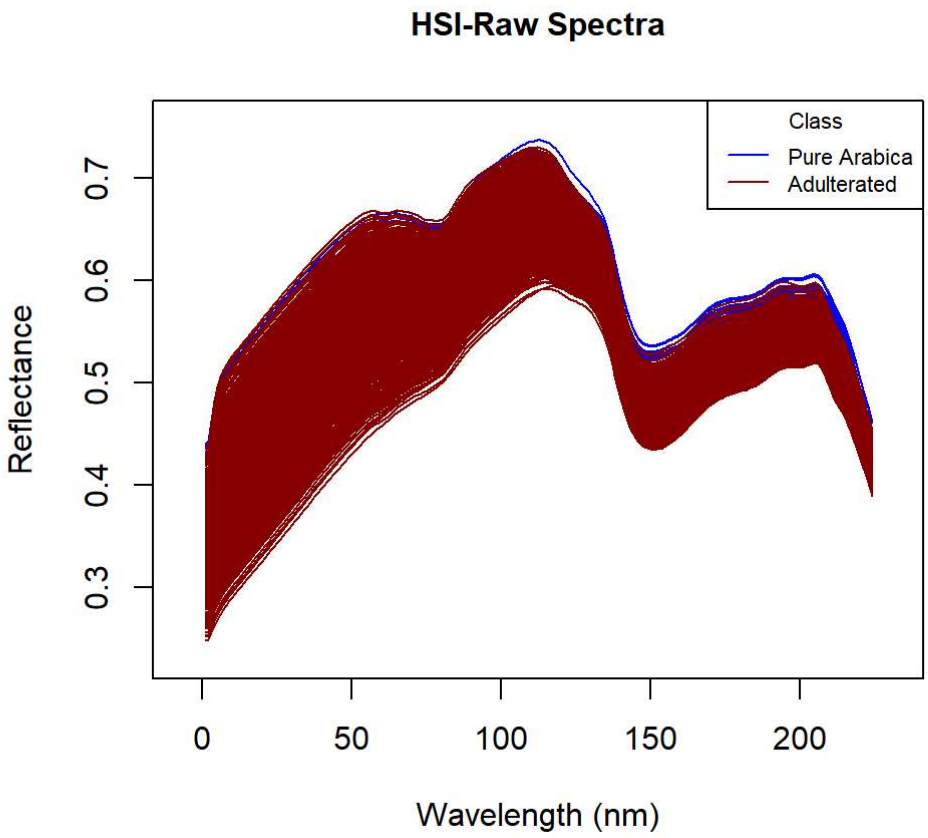
```
[1] 1469  229
```

```
kable(coffee_data_raw[1:5,c(1:12)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	0.347876	0.349376	0.363904	0
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	0.332850	0.334093	0.346032	0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.347639	0.349199	0.363783	0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.338413	0.339599	0.350815	0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.343888	0.345065	0.357601	0

```
detail_columns <- coffee_data_raw[,c(1:5)]

spectral_columns <- as.matrix(coffee_data_raw[,c(6:229)]) # Extract the spectral columns (nume
```

```
# Plot the raw spectra

# Define a color mapping for the binary classes
detail_columns$binary_class <-as.factor(detail_columns$binary_class)
class_colors <- ifelse(detail_columns$binary_class == "pure_arabica", "blue", "darkred")

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(spectral_columns, type = "l", xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-Raw Spectra", cex.main = 1.0)
```



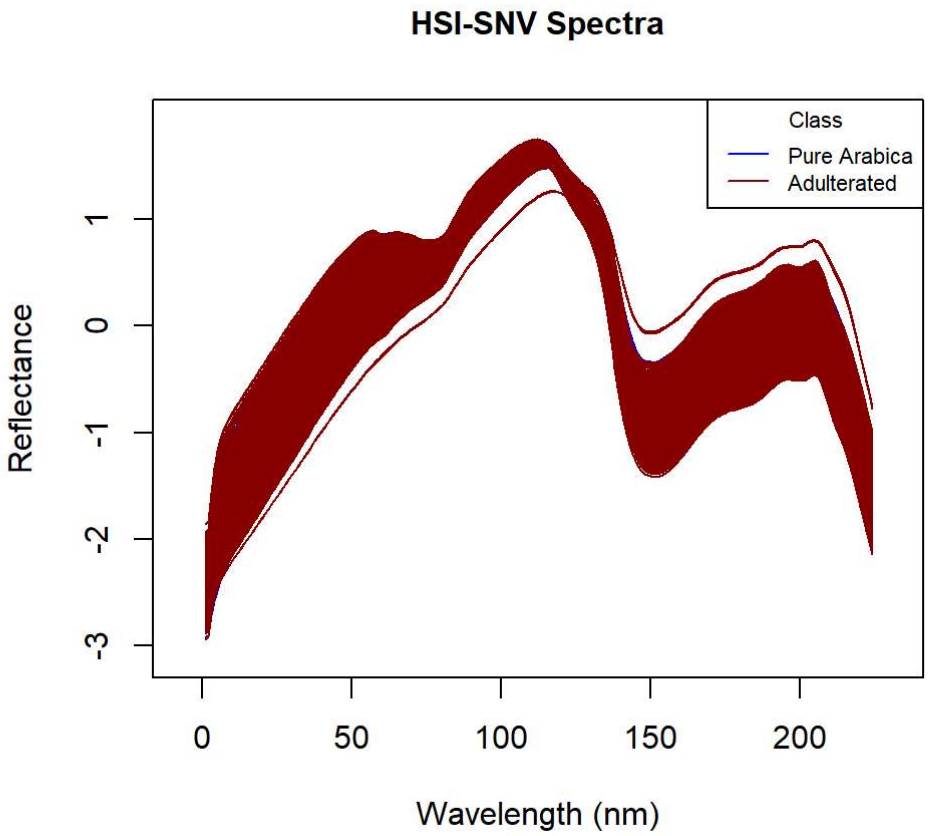

```
# Standard Normal Variate Treatment (SNV)
snv_data_1 <- prep.snv(spectral_columns)
snv_data <- cbind(detail_columns,snv_data_1)
kable(snv_data[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	-2.778104	-2.758083	-2.564173	-2
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	-2.803984	-2.788124	-2.635790	-2
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-2.800631	-2.780635	-2.593698	-2
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-2.746681	-2.731051	-2.583236	-2
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-2.770338	-2.755124	-2.593081	-2

```
#write.csv(snv_data,file = 'coffee_ground_snv.csv',row.names = FALSE) # save SNV data
```

```
# Plot Spectra preprocessed by SNV

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(snv_data_1, type = "l", xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)
```

```
# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-SNV Spectra", cex.main = 1.0)
```

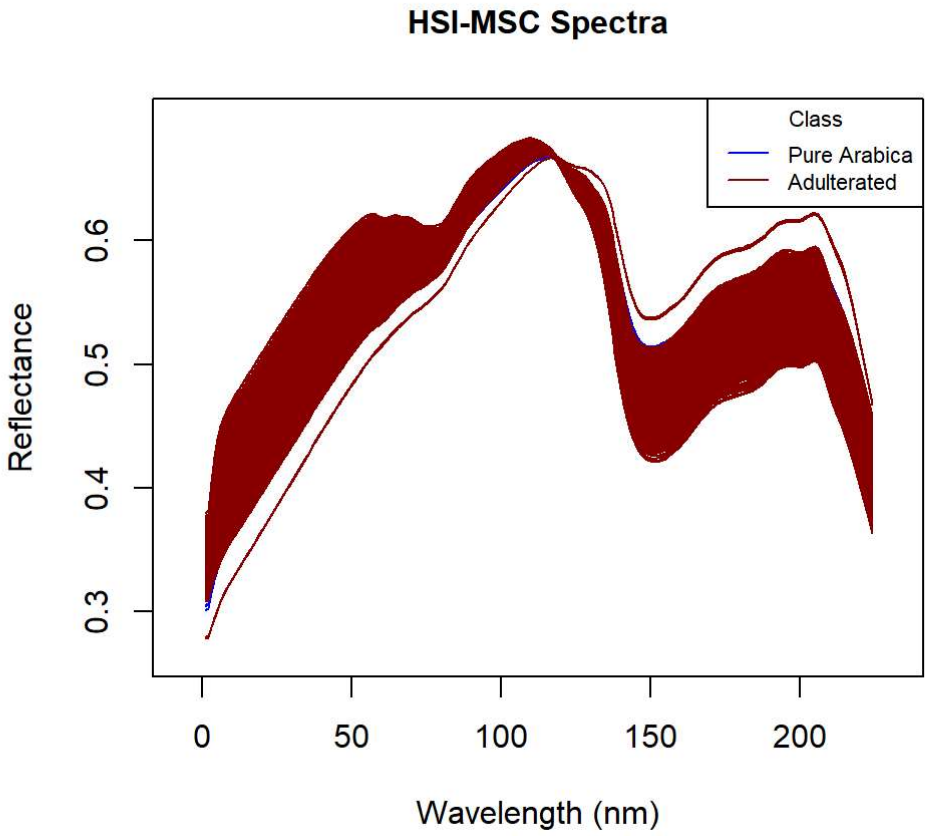



```
# Multiplicative Scatter Correction (MSC)

msc_data_1 <- prep.msc(spectral_columns)
msc_data <- cbind(detail_columns,msc_data_1)
kable(msc_data[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	0.3284397	0.3299852	0.3449542	0.3
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	0.3192311	0.3204962	0.3326475	0.3
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.3252248	0.3267790	0.3413082	0.3
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.3296052	0.3308189	0.3422974	0.3
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.3271205	0.3283055	0.3409268	0.3

```
#write.csv(msc_data,file = 'coffee_ground_msc.csv',row.names = FALSE) # save MSC data
```

```
# Plot Spectra preprocessed by MSC

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(msc_data_1, type = "l", xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-MSC Spectra", cex.main = 1.0)
```



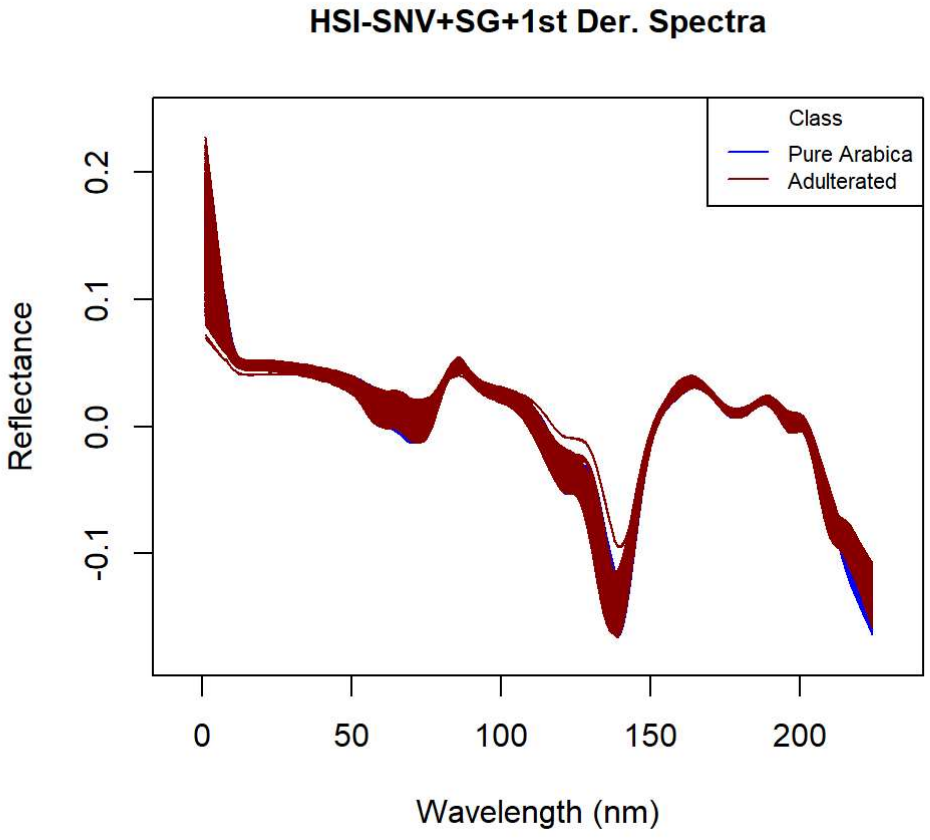

```
# Use of SNV and Savitzky-Golay First Derivative, Polynomial Order 2 and window of 13 points

snv_sg_1d_data_1 <- prep.savgol(snv_data_1,width = 13,porder = 2,dorder = 1)
snv_sg_1d <- cbind(detail_columns,snv_sg_1d_data_1)
kable(snv_sg_1d[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	0.1519042	0.1410505	0.1301968	0.1
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	0.1272989	0.1190992	0.1108994	0.1

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.1462943	0.1360424	0.1257906	0.1
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.1338453	0.1251104	0.1163756	0.1
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.1360254	0.1269853	0.1179452	0.1

```
#write.csv(snv_sg_1d,file = 'coffee_ground_snv_sg_1d.csv',row.names = FALSE) # save SNV_SG_1D_
```

```
# Plot Spectra preprocessed by SNV+Savitzky-Golay+1st Derivative (SNV+SG+1D)

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(snv_sg_1d_data_1, type = "l",xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-SNV+SG+1st Der. Spectra", cex.main = 1.0)
```

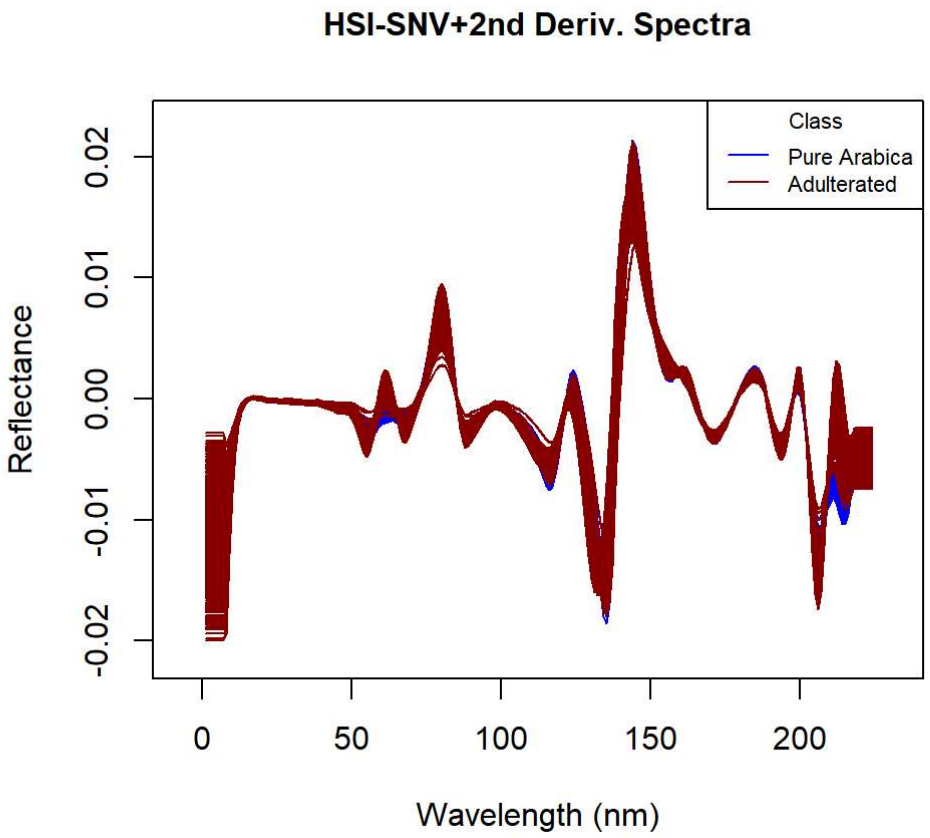


```
# Use of SNV and Savitzky-Golay Second Derivative, Polynomial Order 2 and window of 13 points

snv_sg_2d_data_1 <- prep.savgol(snv_data_1,width = 13,porder = 2,dorder = 2)
snv_sg_2d <- cbind(detail_columns,snv_sg_2d_data_1)
kable(snv_sg_2d[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	-0.0108537	-0.0108537	-0.0108537	-0.0
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	-0.0081997	-0.0081997	-0.0081997	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0102518	-0.0102518	-0.0102518	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0087349	-0.0087349	-0.0087349	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0090401	-0.0090401	-0.0090401	-0.0

```
#write.csv(snv_sg_2d,file = 'coffee_ground_snv_sg_2d.csv',row.names = FALSE) # save SNV_SG_2D_
```

```
# Plot Spectra preprocessed by SNV+Savitzky-Golay+2nd Derivative (SNV+SG+2D)

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(snv_sg_2d_data_1, type = "l", xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-SNV+2nd Deriv. Spectra", cex.main = 1.0)
```



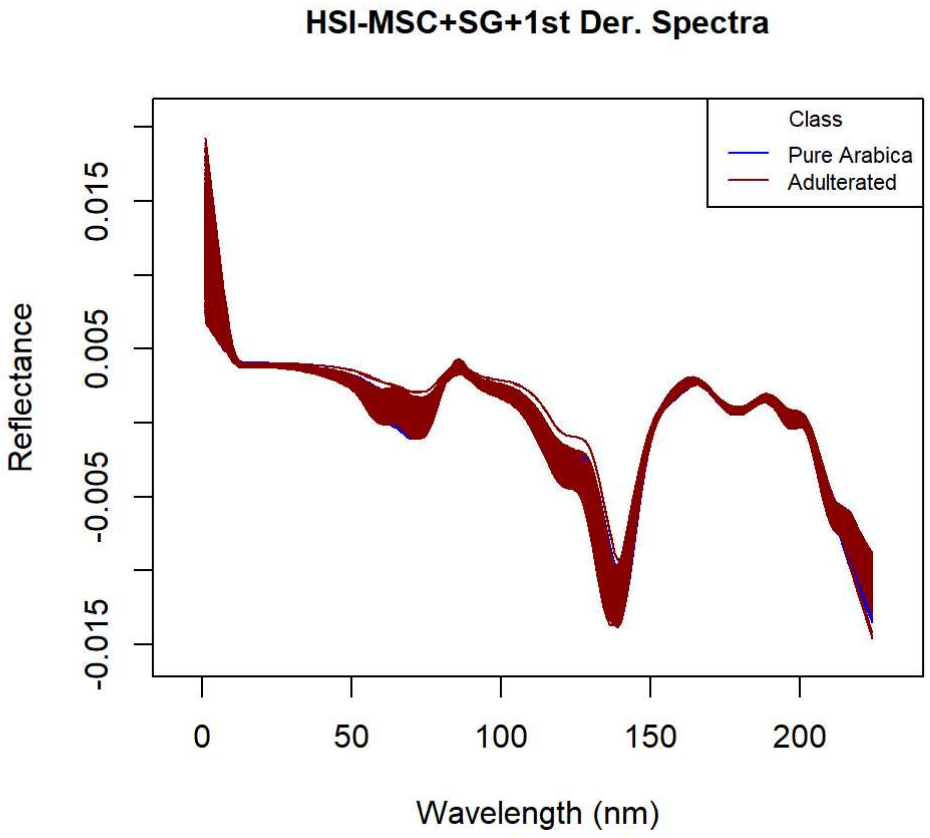

```
# Use of MSC and Savitzky-Golay First Derivative, Polynomial Order 2 and window of 13 points

msc_sg_1d_data_1 <- prep.savgol(msc_data_1,width = 13,porder = 2,dorder = 1)
msc_sg_1d <- cbind(detail_columns,msc_sg_1d_data_1)
kable(msc_sg_1d[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	0.0117264	0.0108885	0.0100506	0.0
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	0.0101543	0.0095002	0.0088462	0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.0113704	0.0105736	0.0097768	0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.0103937	0.0097154	0.0090371	0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	0.0105948	0.0098907	0.0091866	0.0

```
#write.csv(msc_sg_1d,file = 'coffee_ground_msc_sg_1d.csv',row.names = FALSE) # save SNV_SG_1D_
```

```
# Plot Spectra preprocessed by MSC+Savitzky-Golay+1st Derivative (MSC+SG+1D)

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(msc_sg_1d_data_1, type = "l", xlab = 'Wavelength (nm)',
```

```
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-MSC+SG+1st Der. Spectra", cex.main = 1.0)
```

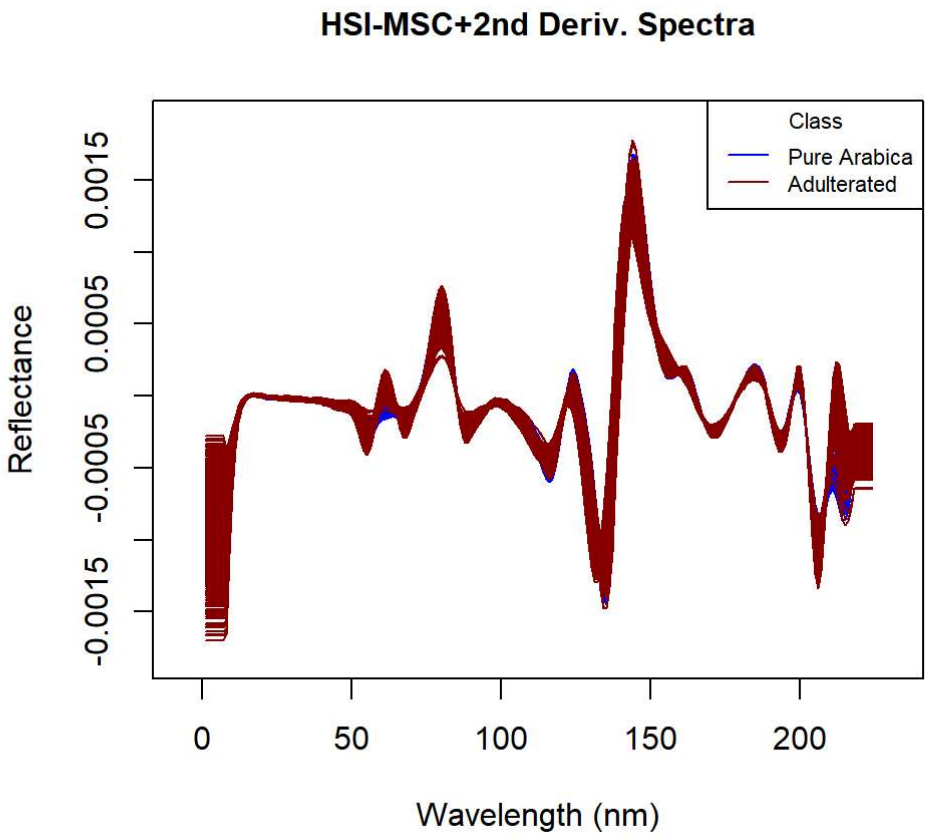



```
# Use of MSC and Savitzky-Golay Second Derivative, Polynomial Order 2 and window of 13 points

msc_sg_2d_data_1 <- prep.savgol(msc_data_1,width = 13,porder = 2,dorder = 2)
msc_sg_2d <- cbind(detail_columns,msc_sg_2d_data_1)
kable(msc_sg_2d[1:5, c(1:10)])
```

sample_id	binary_class	three_class	adult_percent	cal_val	X935.609985	X939.059998	X942.52002	X945
Pure_Arabica_10	pure_arabica	pure_arabica	0	1	-0.0008379	-0.0008379	-0.0008379	-0.0
Pure_Arabica_3	pure_arabica	pure_arabica	0	1	-0.0006541	-0.0006541	-0.0006541	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0007968	-0.0007968	-0.0007968	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0006783	-0.0006783	-0.0006783	-0.0
Pure_Arabica_2	pure_arabica	pure_arabica	0	1	-0.0007041	-0.0007041	-0.0007041	-0.0

```
#write.csv(msc_sg_2d,file = 'coffee_ground_msc_sg_2d.csv',row.names = FALSE) # save MSC_SG_2D_
```

```
# Plot Spectra preprocessed by MSC+Savitzky-Golay+2nd Derivative (MSC+SG+2D)

# Set the aspect ratio
par(pin = c(4, 3))  # Width 4 inches, height 3 inches

mdaplot(msc_sg_2d_data_1, type = "l", xlab = 'Wavelength (nm)',
        ylab = "Reflectance",col = class_colors,show.grid = FALSE)

# Add a legend to the plot
legend("topright", legend = c("Pure Arabica", "Adulterated"),
       col = c("blue", "darkred"), lty = 1, title = "Class", cex = 0.7)

# Add the title
title(main = "HSI-MSC+2nd Deriv. Spectra", cex.main = 1.0)
```



Next Steps

Reduction dimension by **PCA** and unsupervised learning/**clustering** by k-means or **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) will be explored.