Contents

1 Phát hiện Outliers nâng cao **1**

1.1 Chuẩn bị dữ liệu . . . 1
1.2 Hotelling T^2 Statistic . . . 2
1.3 Q Residuals (SPE - Squared Prediction Error) . . . 3
1.4 Combined T^2 vs Q Plot . . . 5
1.5 Phân tích Outliers theo Location . . . 7
1.6 Influence Plot . . . 8
1.7 Danh sách Outliers chi tiết . . . 10
1.8 So sánh phổ NIR của Outliers . . . 10
1.9 Quyết định xử lý Outliers . . . 12
1.10 Kết luận . . . 13

1 Phát hiện Outliers nâng cao

Trong phần này, chúng ta sử dụng các phương pháp thống kê nâng cao để phát hiện outliers trong dữ liệu NIR, đặc biệt là:

1. **Hotelling T^2 statistic**: Đo khoảng cách của mẫu đến trung tâm của mô hình PCA
2. **Q residuals (SPE)**: Đo phần dữ liệu không được giải thích bởi mô hình PCA
3. **Leverage và Influence**: Đánh giá ảnh hưởng của từng mẫu đến mô hình

1.1 Chuẩn bị dữ liệu

```
library(tidyverse)
library(FactoMineR)
library(factoextra)
library(knitr)
library(kableExtra)
library(gridExtra)

# Đọc dữ liệu
coffee_data <- read.csv("coffee_nirs.csv", sep = ";", row.names = 1, stringsAsFactors = FALSE)

# Chuyển đổi sang numeric
convert_to_numeric <- function(x) {
  if(is.character(x)) {
    x <- gsub("\\.", "", x)
    x <- gsub(",", ".", x)
    return(as.numeric(x))
  }
  return(x)
}
```

```r
for(col in names(coffee_data)) {
  if(col != "Localisation") {
    coffee_data[[col]] <- convert_to_numeric(coffee_data[[col]])
  }
}

# Định nghĩa các nhóm biến
chemical_vars <- c("CGA", "Cafeine", "Fat", "Trigonelline", "DM")
nir_vars <- grep("^S[0-9]+$", names(coffee_data), value = TRUE)

# Chuẩn bị dữ liệu NIR
data_nir <- coffee_data[, nir_vars]
data_nir_complete <- data_nir[complete.cases(data_nir), ]
location_info <- coffee_data$Localisation[complete.cases(data_nir)]

# Thực hiện PCA (nếu chưa có)
pca_nir <- PCA(data_nir_complete, scale.unit = TRUE, graph = FALSE)
```

1.2 Hotelling T^2 Statistic

Hotelling T^2 đo khoảng cách Mahalanobis của mỗi mẫu đến trung tâm của không gian PCA.

```r
# Số PC sử dụng (chọn theo cumulative variance > 95% hoặc elbow)
n_pcs <- min(10, ncol(pca_nir$ind$coord))
scores <- pca_nir$ind$coord[, 1:n_pcs]

# Tính Hotelling T2
eigenvalues <- pca_nir$eig[1:n_pcs, 1]
t2 <- rowSums(t(t(scores^2) / eigenvalues))

# Ngưỡng T2 (chi-square distribution với alpha = 0.05)
n <- nrow(scores)
p <- n_pcs
t2_limit <- qchisq(0.95, df = p)

# Xác định outliers
t2_outliers <- which(t2 > t2_limit)

cat("Ngưỡng T² (95%):", round(t2_limit, 2), "\n")
```

```
## Ngưỡng T² (95%): 11.07
```

```r
cat("Số mẫu vượt ngưỡng T²:", length(t2_outliers), "\n")
```

```
## Số mẫu vượt ngưỡng T²: 15
```

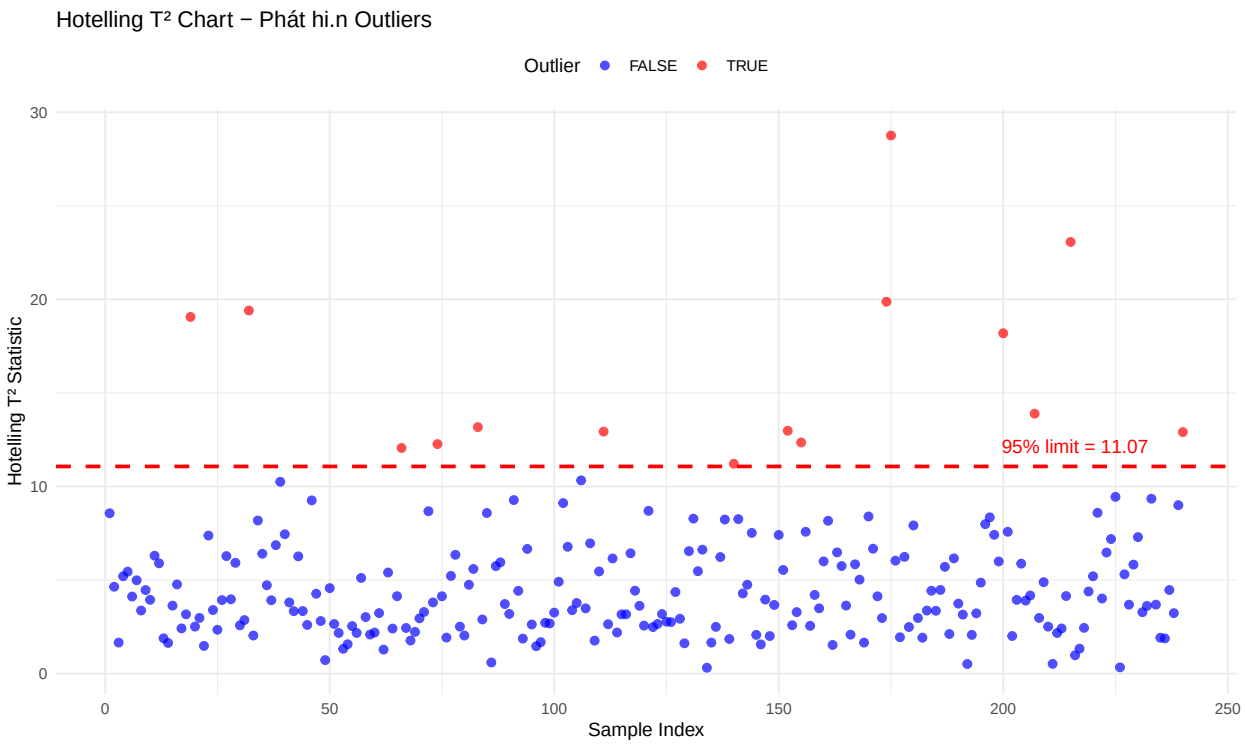
```r
cat("Tỷ lệ outliers:", round(length(t2_outliers) / n * 100, 2), "%\n")
```

```
## Tỷ lệ outliers: 6.25 %
```

1.2.1 Biểu đồ Hotelling T²

```r
# Tạo data frame
t2_df <- data.frame(
  Sample = 1:length(t2),
  T2 = t2,
  Location = factor(location_info),
  Outlier = t2 > t2_limit
)

# T2 chart
ggplot(t2_df, aes(x = Sample, y = T2, color = Outlier)) +
  geom_point(size = 2, alpha = 0.7) +
  geom_hline(yintercept = t2_limit, linetype = "dashed", color = "red", linewidth = 1) +
  annotate("text", x = max(t2_df$Sample) * 0.9, y = t2_limit * 1.1,
           label = paste0("95% limit = ", round(t2_limit, 2)), color = "red") +
  scale_color_manual(values = c("FALSE" = "blue", "TRUE" = "red")) +
  labs(
    title = "Hotelling T² Chart - Phát hiện Outliers",
    x = "Sample Index",
    y = "Hotelling T² Statistic"
  ) +
  theme_minimal() +
  theme(legend.position = "top")
```



1.3 Q Residuals (SPE - Squared Prediction Error)

Q residuals đo phần biến động không được giải thích bởi các PC đã chọn.

```r
# Tính Q residuals (SPE)
# Reconstruction của dữ liệu từ PCA
data_reconstructed <- scores %*% t(pca_nir$svd$V[, 1:n_pcs])
data_centered <- scale(data_nir_complete, center = TRUE, scale = TRUE)

# Q statistic = sum of squared residuals
q_stat <- rowSums((data_centered - data_reconstructed)^2)

# Ngưỡng Q (approximate chi-square)
# Sử dụng Jackson-Mudholkar approximation
theta1 <- sum(pca_nir$eig[(n_pcs+1):nrow(pca_nir$eig), 1])
theta2 <- sum(pca_nir$eig[(n_pcs+1):nrow(pca_nir$eig), 1]^2)
theta3 <- sum(pca_nir$eig[(n_pcs+1):nrow(pca_nir$eig), 1]^3)

h0 <- 1 - (2 * theta1 * theta3) / (3 * theta2^2)
ca <- qnorm(0.95)
q_limit <- theta1 * (1 + ca * sqrt(2 * theta2 * h0^2) / theta1 +
                     theta2 * h0 * (h0 - 1) / theta1^2)^(1/h0)

# Xác định outliers
q_outliers <- which(q_stat > q_limit)

cat("Ngưỡng Q (95%):", round(q_limit, 2), "\n")
```

```
## Ngưỡng Q (95%): 219.98
```

```r
cat("Số mẫu vượt ngưỡng Q:", length(q_outliers), "\n")
```

```
## Số mẫu vượt ngưỡng Q: 32
```

```r
cat("Tỷ lệ outliers:", round(length(q_outliers) / n * 100, 2), "%\n")
```

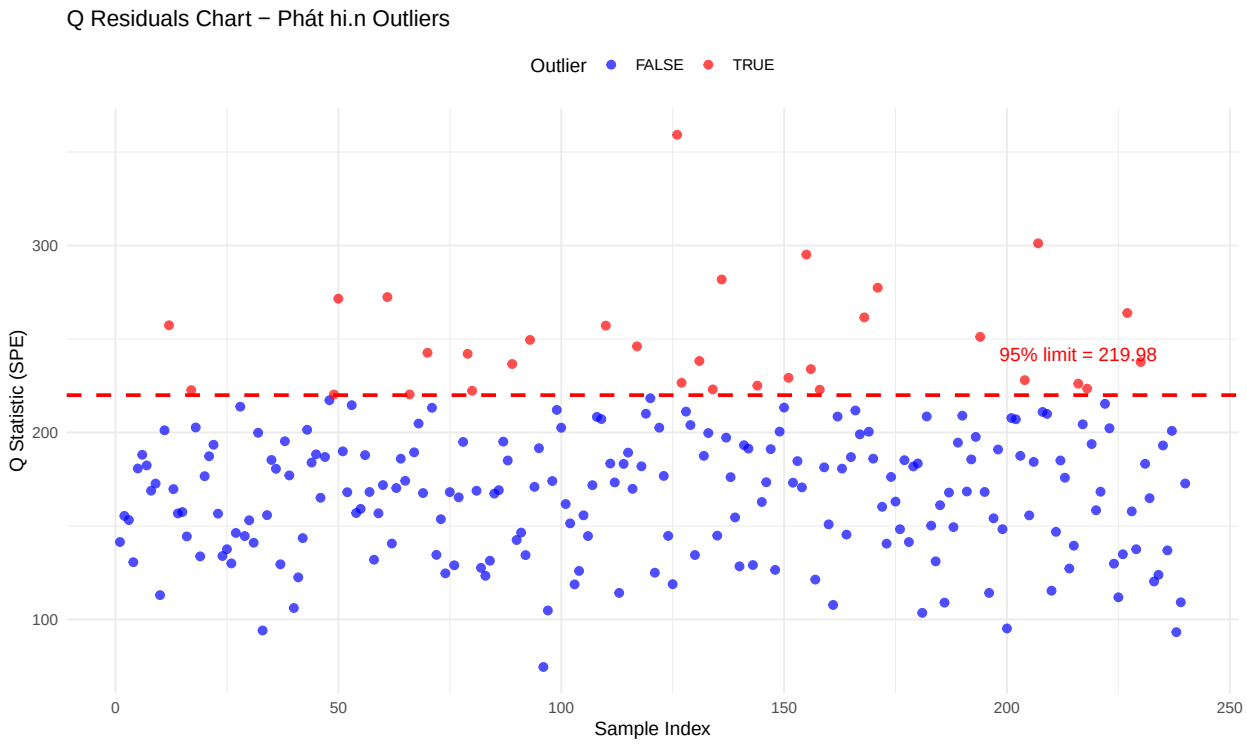
```
## Tỷ lệ outliers: 13.33 %
```

1.3.1 Biểu đồ Q Residuals

```r
# Tạo data frame
q_df <- data.frame(
  Sample = 1:length(q_stat),
  Q = q_stat,
  Location = factor(location_info),
  Outlier = q_stat > q_limit
)

# Q chart
ggplot(q_df, aes(x = Sample, y = Q, color = Outlier)) +
  geom_point(size = 2, alpha = 0.7) +
  geom_hline(yintercept = q_limit, linetype = "dashed", color = "red", linewidth = 1) +
  annotate("text", x = max(q_df$Sample) * 0.9, y = q_limit * 1.1,
           label = paste0("95% limit = ", round(q_limit, 2)), color = "red") +
```

```
  scale_color_manual(values = c("FALSE" = "blue", "TRUE" = "red")) +
  labs(
    title = "Q Residuals Chart - Phát hiện Outliers",
    x = "Sample Index",
    y = "Q Statistic (SPE)"
  ) +
  theme_minimal() +
  theme(legend.position = "top")
```



1.4 Combined T² vs Q Plot

Biểu đồ kết hợp T² và Q giúp phân loại outliers:

- **High T², Low Q**: Mẫu cực trị nhưng vẫn theo mô hình
- **Low T², High Q**: Mẫu gần trung tâm nhưng không theo mô hình
- **High T², High Q**: Mẫu outlier mạnh

```
# Kết hợp T2 và Q
combined_df <- data.frame(
  Sample = 1:n,
  T2 = t2,
  Q = q_stat,
  Location = factor(location_info),
  T2_outlier = t2 > t2_limit,
  Q_outlier = q_stat > q_limit
)
```

```r
# Phân loại outliers
combined_df$Outlier_Type <- "Normal"
combined_df$Outlier_Type[combined_df$T2_outlier & !combined_df$Q_outlier] <- "High T² only"
combined_df$Outlier_Type[!combined_df$T2_outlier & combined_df$Q_outlier] <- "High Q only"
combined_df$Outlier_Type[combined_df$T2_outlier & combined_df$Q_outlier] <- "Both High"

# T2 vs Q scatter plot
ggplot(combined_df, aes(x = T2, y = Q, color = Outlier_Type)) +
  geom_point(size = 2.5, alpha = 0.7) +
  geom_vline(xintercept = t2_limit, linetype = "dashed", color = "red") +
  geom_hline(yintercept = q_limit, linetype = "dashed", color = "red") +
  scale_color_manual(values = c(
    "Normal" = "blue",
    "High T² only" = "orange",
    "High Q only" = "purple",
    "Both High" = "red"
  )) +
  labs(
    title = "T² vs Q Plot - Phân loại Outliers",
    x = "Hotelling T²",
    y = "Q Residuals (SPE)",
    color = "Outlier Type"
  ) +
  theme_minimal() +
  theme(legend.position = "right")
```


```r
# Thống kê outliers
outlier_summary <- combined_df %>%
  count(Outlier_Type) %>%
```

Table 1: Phân loại Outliers theo T^2 và Q

Outlier_Type	n	Percentage
Both High	3	1.25
High Q only	29	12.08
High T^2 only	12	5.00
Normal	196	81.67

Table 2: Thống kê Outliers theo Location

Location	Total_Samples	N_Outliers	Pct_Outliers	Mean_T2	Mean_Q
7	19	7	36.84	9.21	170.08
3	26	8	30.77	4.39	185.24
1	50	9	18.00	5.09	172.87
6	84	14	16.67	4.25	186.46
2	26	4	15.38	3.51	179.93
5	22	2	9.09	6.49	152.98
4	13	0	0.00	5.03	148.20

```
  mutate(Percentage = round(n / nrow(combined_df) * 100, 2))

outlier_summary %>%
  kable(caption = "Phân loại Outliers theo T² và Q") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

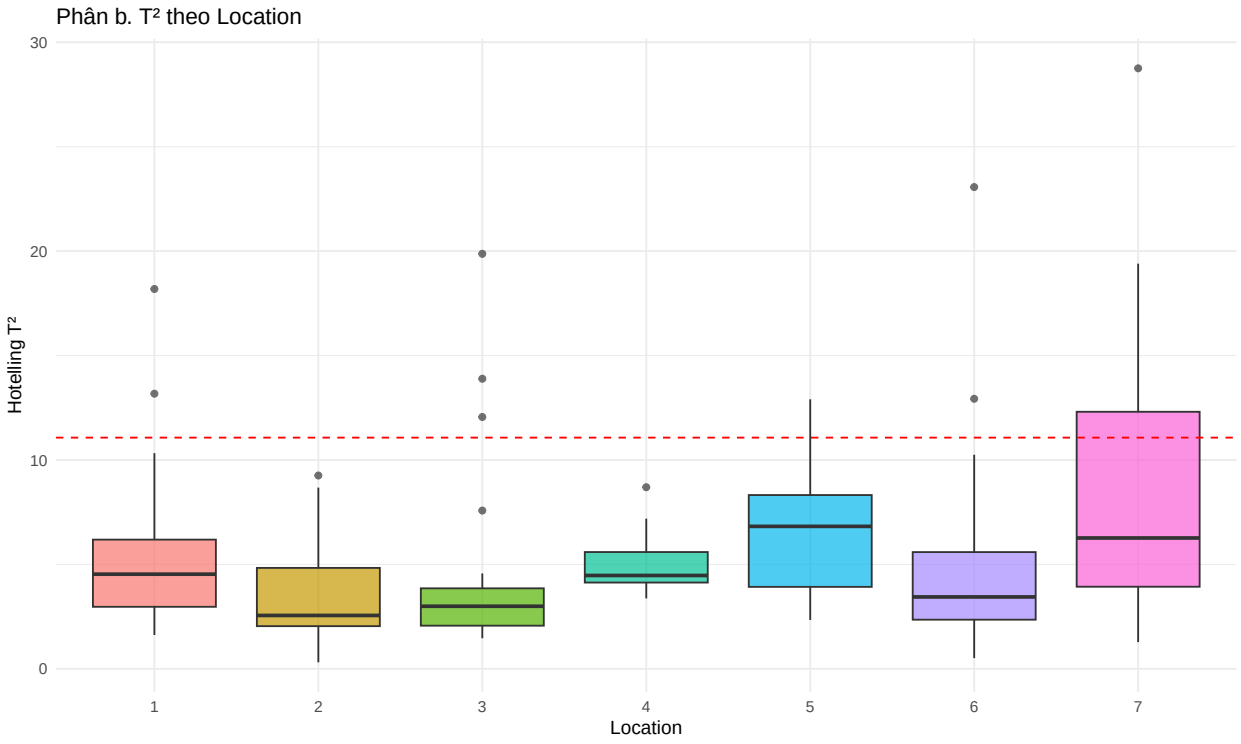
1.5 Phân tích Outliers theo Location

```
# Đếm outliers theo Location
outlier_location <- combined_df %>%
  mutate(Is_Outlier = T2_outlier | Q_outlier) %>%
  group_by(Location) %>%
  summarise(
    Total_Samples = n(),
    N_Outliers = sum(Is_Outlier),
    Pct_Outliers = round(N_Outliers / Total_Samples * 100, 2),
    Mean_T2 = round(mean(T2), 2),
    Mean_Q = round(mean(Q), 2)
  ) %>%
  arrange(desc(Pct_Outliers))

outlier_location %>%
  kable(caption = "Thống kê Outliers theo Location") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

```
# Boxplot T2 theo Location
ggplot(combined_df, aes(x = Location, y = T2, fill = Location)) +
  geom_boxplot(alpha = 0.7) +
  geom_hline(yintercept = t2_limit, linetype = "dashed", color = "red") +
```

```
  labs(
    title = "Phân bố T² theo Location",
    x = "Location",
    y = "Hotelling T²"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



1.6 Influence Plot

Influence plot kết hợp leverage (khoảng cách từ trung tâm) và residuals.

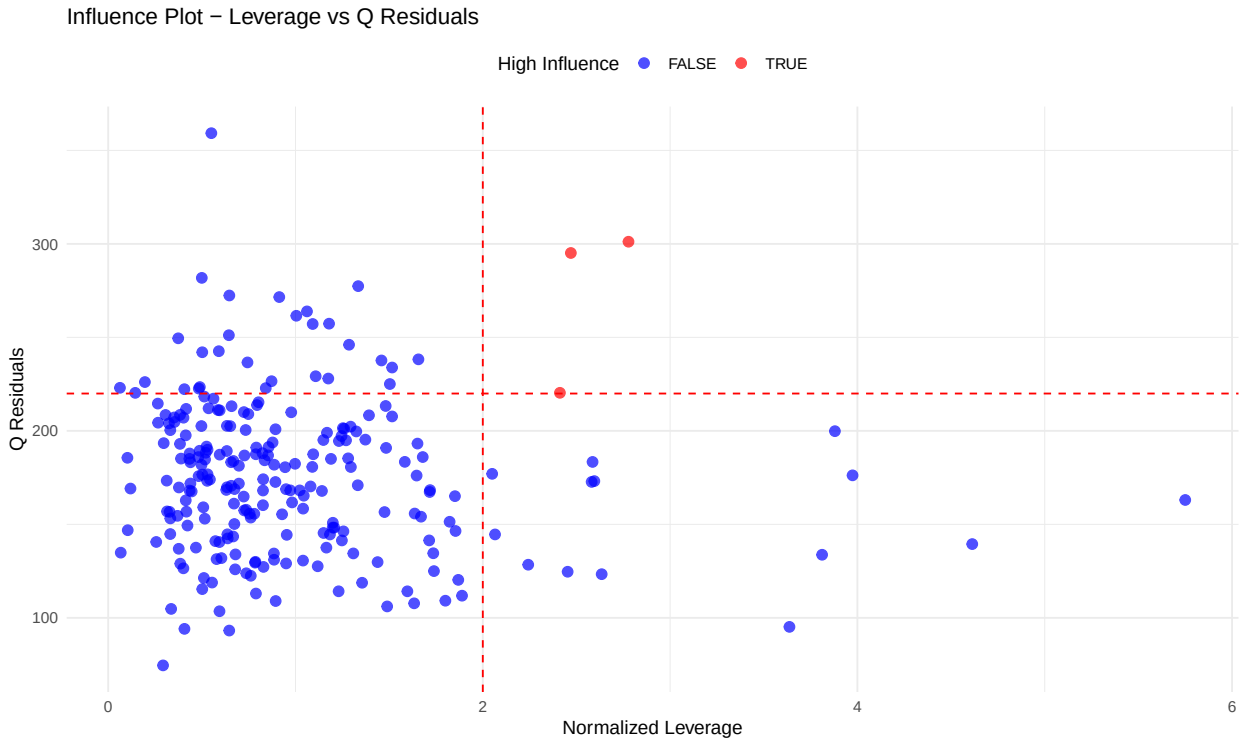
```
# Leverage: diagonal của hat matrix trong không gian PC
leverage <- diag(scores %*% solve(t(scores) %*% scores) %*% t(scores))

# Chuẩn hóa leverage
leverage_norm <- leverage / mean(leverage)

# Influence = leverage × residuals
influence_df <- data.frame(
  Sample = 1:n,
  Leverage = leverage_norm,
  Q = q_stat,
  Location = factor(location_info),
  High_Influence = leverage_norm > 2 & q_stat > q_limit
)

# Influence plot
```

```
ggplot(influence_df, aes(x = Leverage, y = Q, color = High_Influence)) +
  geom_point(size = 2.5, alpha = 0.7) +
  geom_vline(xintercept = 2, linetype = "dashed", color = "red") +
  geom_hline(yintercept = q_limit, linetype = "dashed", color = "red") +
  scale_color_manual(values = c("FALSE" = "blue", "TRUE" = "red")) +
  labs(
    title = "Influence Plot - Leverage vs Q Residuals",
    x = "Normalized Leverage",
    y = "Q Residuals",
    color = "High Influence"
  ) +
  theme_minimal() +
  theme(legend.position = "top")
```




```
high_influence <- which(influence_df$High_Influence)
cat("Số mẫu có influence cao:", length(high_influence), "\n")
```

```
## Số mẫu có influence cao: 3
```

```
if(length(high_influence) > 0) {
  cat("Các mẫu có influence cao:", paste(head(high_influence, 10), collapse = ", "), "\n")
}
```

```
## Các mẫu có influence cao: 66, 155, 207
```

Table 3: Top 20 Outliers (sắp xếp theo T^2 + Q)

	Sample	Location	T2	Q	Outlier_Type
V127	126	3	2.753825	359.2182	High Q only
V208	207	3	13.889167	301.1639	Both High
V156	155	7	12.350460	295.1462	Both High
V137	136	6	2.498920	281.8304	High Q only
V172	171	5	6.674063	277.4150	High Q only
V51	50	3	4.567273	271.5660	High Q only
V62	61	6	3.234206	272.4102	High Q only
V228	227	6	5.306836	263.9010	High Q only
V169	168	1	5.022219	261.5611	High Q only
V13	12	6	5.896187	257.3362	High Q only
V111	110	1	5.458945	257.1632	High Q only
V195	194	3	3.221745	251.1642	High Q only
V118	117	1	6.429499	246.0678	High Q only
V94	93	2	1.869872	249.5053	High Q only
V132	131	6	8.281691	238.2543	High Q only
V71	70	1	2.952866	242.6182	High Q only
V231	230	1	7.294970	237.6560	High Q only
V80	79	2	2.511432	242.0534	High Q only
V157	156	3	7.578982	233.8829	High Q only
V90	89	6	3.718666	236.6165	High Q only

1.7 Danh sách Outliers chi tiết

```
# Liệt kê tất cả outliers
all_outliers <- combined_df %>%
  filter(T2_outlier | Q_outlier) %>%
  arrange(desc(T2 + Q)) %>%
  select(Sample, Location, T2, Q, Outlier_Type)

if(nrow(all_outliers) > 0) {
  cat("Tìm thấy", nrow(all_outliers), "outliers:\n\n")
  head(all_outliers, 20) %>%
    kable(caption = "Top 20 Outliers (sắp xếp theo T² + Q)") %>%
    kable_styling(bootstrap_options = c("striped", "hover"))
} else {
  cat("Không tìm thấy outliers nào.\n")
}
```

```
## Tìm thấy 44 outliers:
```

1.8 So sánh phổ NIR của Outliers

```
# Lấy top 5 outliers mạnh nhất
if(nrow(all_outliers) > 0) {
  top_outliers <- head(all_outliers$Sample, 5)
```

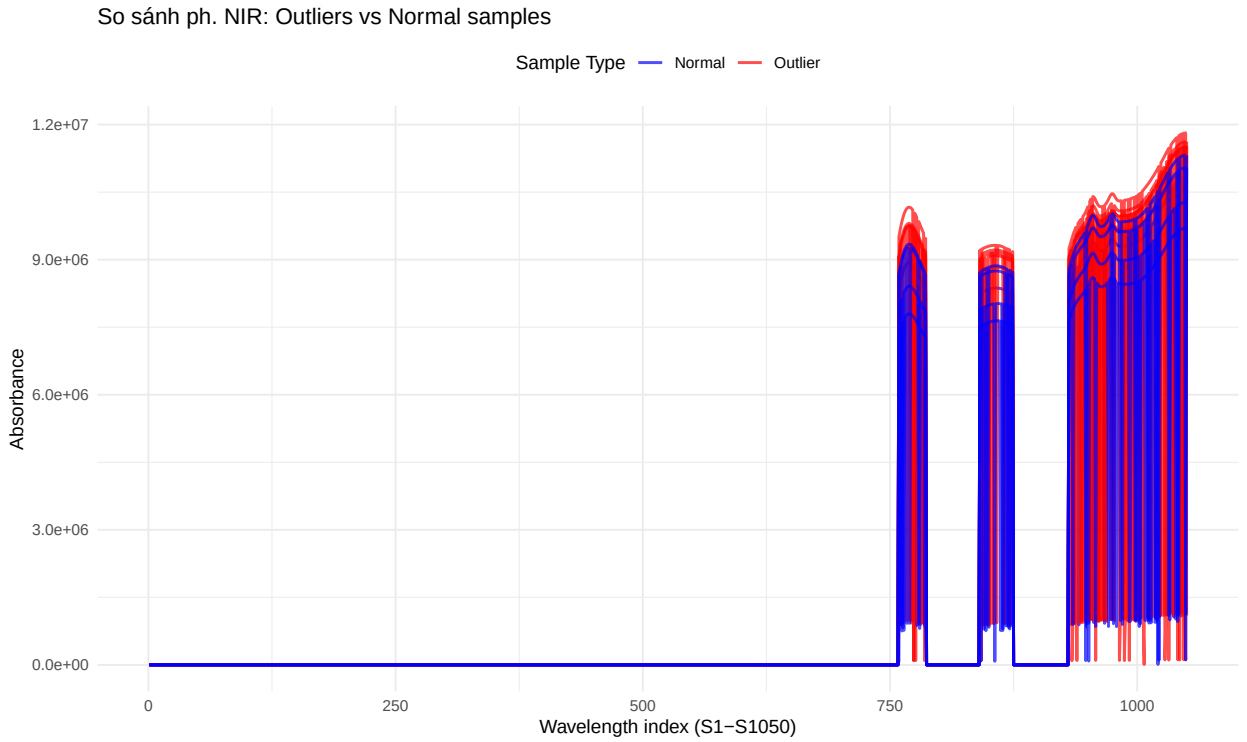
```r
  # Lấy mẫu bình thường để so sánh
  normal_samples <- combined_df %>%
    filter(!T2_outlier & !Q_outlier) %>%
    sample_n(min(5, sum(!combined_df$T2_outlier & !combined_df$Q_outlier))) %>%
    pull(Sample)

  # Chuẩn bị data
  selected_spectra <- data_nir_complete[c(top_outliers, normal_samples), ]

  # Thêm ID và Type TRƯỚC khi pivot
  sample_metadata <- data.frame(
    Sample_ID = paste0("Sample_", 1:nrow(selected_spectra)),
    Type = c(rep("Outlier", length(top_outliers)),
             rep("Normal", length(normal_samples)))
  )

  # Chuyển sang long format
  spectra_comparison <- selected_spectra %>%
    mutate(Sample_ID = sample_metadata$Sample_ID,
           Type = sample_metadata$Type) %>%
    pivot_longer(cols = matches("^S[0-9]+$"),
                 names_to = "Wavelength",
                 values_to = "Absorbance") %>%
    mutate(Wavelength = as.numeric(str_remove(Wavelength, "S")))

  # Plot
  ggplot(spectra_comparison, aes(x = Wavelength, y = Absorbance,
                                 group = Sample_ID, color = Type)) +
    geom_line(alpha = 0.7, linewidth = 0.8) +
    scale_color_manual(values = c("Outlier" = "red", "Normal" = "blue")) +
    labs(
      title = "So sánh phổ NIR: Outliers vs Normal samples",
      x = "Wavelength index (S1-S1050)",
      y = "Absorbance",
      color = "Sample Type"
    ) +
    theme_minimal() +
    theme(legend.position = "top")
}
```



1.9 Quyết định xử lý Outliers

```
cat("### Khuyến nghị xử lý Outliers:\n\n")
```

```
## ### Khuyến nghị xử lý Outliers:
```

```
total_outliers <- nrow(all_outliers)
outlier_pct <- total_outliers / n * 100

cat("1. **Tổng số outliers phát hiện:**", total_outliers,
    "(", round(outlier_pct, 2), "%)\n\n")
```

```
## 1. **Tổng số outliers phát hiện:** 44 ( 18.33 %)
```

```
if(outlier_pct < 1) {
  cat("2. **Mức độ:** Rất thấp - ít ảnh hưởng đến mô hình\n")
  cat("3. **Khuyến nghị:** Có thể giữ lại, chỉ cần lưu ý khi xây dựng mô hình\n\n")
} else if(outlier_pct < 5) {
  cat("2. **Mức độ:** Trung bình - cần xem xét kỹ\n")
  cat("3. **Khuyến nghị:**\n")
  cat("   - Kiểm tra lại dữ liệu gốc của các outliers\n")
  cat("   - Xem xét loại bỏ các outliers có cả T² và Q cao\n")
  cat("   - So sánh mô hình với và không có outliers\n\n")
} else {
  cat("2. **Mức độ:** Cao - cần xử lý nghiêm túc\n")
  cat("3. **Khuyến nghị:**\n")
```

```
  cat("   - Kiểm tra kỹ quy trình đo NIR và chuẩn bị mẫu\n")
  cat("   - Xem xét xây dựng mô hình robust (PLS-robust)\n")
  cat("   - Có thể cần thu thập thêm dữ liệu\n\n")
}
```

```
## 2. **Mức độ:** Cao - cần xử lý nghiêm túc
## 3. **Khuyến nghị:**
##    - Kiểm tra kỹ quy trình đo NIR và chuẩn bị mẫu
##    - Xem xét xây dựng mô hình robust (PLS-robust)
##    - Có thể cần thu thập thêm dữ liệu
```

```
cat("4. **Phân loại outliers:**\n")
```

```
## 4. **Phân loại outliers:**
```

```
if(nrow(all_outliers) > 0) {
  cat("   - High T² only:", sum(all_outliers$Outlier_Type == "High T² only"),
      "- Mẫu cực trị nhưng đúng pattern\n")
  cat("   - High Q only:", sum(all_outliers$Outlier_Type == "High Q only"),
      "- Mẫu không theo mô hình (có thể lỗi đo)\n")
  cat("   - Both High:", sum(all_outliers$Outlier_Type == "Both High"),
      "- Outliers mạnh (ưu tiên xem xét loại bỏ)\n\n")
}
```

```
##    - High T² only: 12 - Mẫu cực trị nhưng đúng pattern
##    - High Q only: 29 - Mẫu không theo mô hình (có thể lỗi đo)
##    - Both High: 3 - Outliers mạnh (ưu tiên xem xét loại bỏ)
```

```
cat("5. **Bước tiếp theo:**\n")
```

```
## 5. **Bước tiếp theo:**
```

```
cat("   - Xem xét phân tích thêm các mẫu outliers có influence cao\n")
```

```
##    - Xem xét phân tích thêm các mẫu outliers có influence cao
```

```
cat("   - Kiểm tra xem outliers có tập trung ở Location nào không\n")
```

```
##    - Kiểm tra xem outliers có tập trung ở Location nào không
```

```
cat("   - Thử xây dựng mô hình PLS với và không có outliers để so sánh hiệu suất\n")
```

```
##    - Thử xây dựng mô hình PLS với và không có outliers để so sánh hiệu suất
```

1.10 Kết luận

```r
cat("### Tóm tắt phát hiện Outliers:\n\n")
```

```
## ### Tóm tắt phát hiện Outliers:
```

```r
cat("- **Phương pháp sử dụng:**\n")
```

```
## - **Phương pháp sử dụng:**
```

```r
cat("  + Hotelling T² statistic (khoảng cách trong không gian PC)\n")
```

```
##   + Hotelling T² statistic (khoảng cách trong không gian PC)
```

```r
cat("  + Q residuals/SPE (phần không giải thích được)\n")
```

```
##   + Q residuals/SPE (phần không giải thích được)
```

```r
cat("  + Influence analysis (leverage × residuals)\n\n")
```

```
##   + Influence analysis (leverage × residuals)
```

```r
cat("- **Kết quả:**\n")
```

```
## - **Kết quả:**
```

```r
cat("  + Tổng số mẫu:", n, "\n")
```

```
##   + Tổng số mẫu: 240
```

```r
cat("  + Số PC sử dụng:", n_pcs, "\n")
```

```
##   + Số PC sử dụng: 5
```

```r
cat("  + Outliers theo T²:", length(t2_outliers), "\n")
```

```
##   + Outliers theo T²: 15
```

```r
cat("  + Outliers theo Q:", length(q_outliers), "\n")
```

```
##   + Outliers theo Q: 32
```

```r
cat("  + Tổng outliers (T² hoặc Q):", total_outliers, "\n\n")
```

```
##   + Tổng outliers (T² hoặc Q): 44
```

```r
cat("- **Ý nghĩa:**\n")
```

```
## - **Ý nghĩa:**
```

```r
cat("  + Outliers có thể do lỗi đo, contamination, hoặc mẫu thực sự khác biệt\n")
```

```
##   + Outliers có thể do lỗi đo, contamination, hoặc mẫu thực sự khác biệt
```

```r
cat("  + Cần kiểm tra kỹ trước khi quyết định loại bỏ\n")
```

```
##   + Cần kiểm tra kỹ trước khi quyết định loại bỏ
```

```r
cat("  + Xem xét ảnh hưởng của outliers đến mô hình dự đoán\n")
```

```
##   + Xem xét ảnh hưởng của outliers đến mô hình dự đoán
```