Statistical Data Analysis Project

Alexandre De Cuyper

University of A Coruña

February 22, 2024

Introduction

- Analyze the powder X-ray diffraction data,
- Objectives and hypotheses: Investigating if there are significant differences in peak positions and intensity among crystals,
- Importance of the analysis: Understanding the change in structure of the crystals when changing the composition.

Data Overview

- Overview of the dataset: Peaks from different crystals with compositions and varying 2 theta values.
- Key variables and their significance: $`2_theta`$, $Composition$, $Intensity$, $andCluster$.

Data Preprocessing

- Cleaning and handling missing values.
- Transformation of data for analysis.

```
> head(data)
  2_theta Ni75Co25 Ni50Co50 Ni25Co75   Co
1    8.00     20.0    18.00    24.00 19.0
2    8.05     19.5    21.98    23.01 20.5
3    8.10     19.0    26.00    22.00 22.0
4    8.15     19.0    27.49    21.50 25.0
5    8.20     19.0    29.00    21.00 28.0
6    8.25     17.0    24.53    24.98 24.5
```

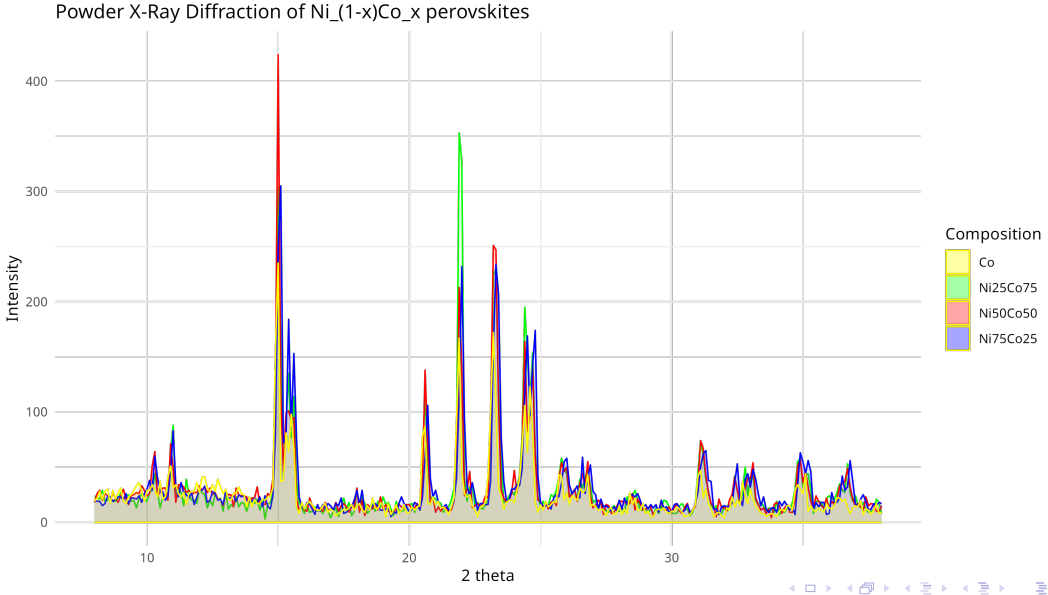
Exploratory Data Analysis

- Visualizations and summary statistics.
- Identification of patterns and trends.

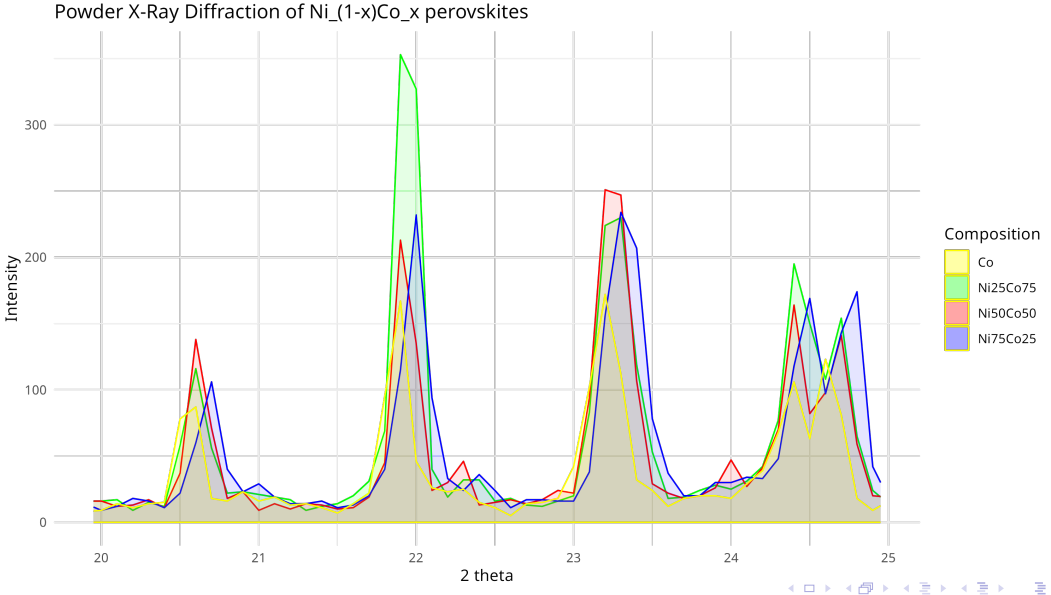
```
  # Exploratory Data Analysis (EDA)
  ggplot(data, aes(x = `2_theta`)) +
geom_ribbon(aes(ymin = 0, ymax = `Ni25Co75`, fill = "Ni25Co75"),
  alpha = 0.1, color = "green"
) +
geom_ribbon(aes(ymin = 0, ymax = `Ni50Co50`, fill = "Ni50Co50"),
  alpha = 0.1, color = "red"
) +
geom_ribbon(aes(ymin = 0, ymax = `Ni75Co25`, fill = "Ni75Co25"),
  alpha = 0.1, color = "blue"
) +
geom_ribbon(aes(ymin = 0, ymax = `Co`, fill = "Co"),
  alpha = 0.1, color = "yellow"
) +
labs(
  title = "Powder X-ray Diffraction of Ni_(1-x)Co_x perovskites",
  x = "2 theta",
  y = "Intensity",
  fill = "Composition"
) +
scale_fill_manual(values = c(
  "Ni75Co25" = "blue", "Ni50Co50" = "red",
  "Ni25Co75" = "green", "Co" = "yellow"
)) +
theme_minimal()

```

Visualization of the Data



Visualization of the Data



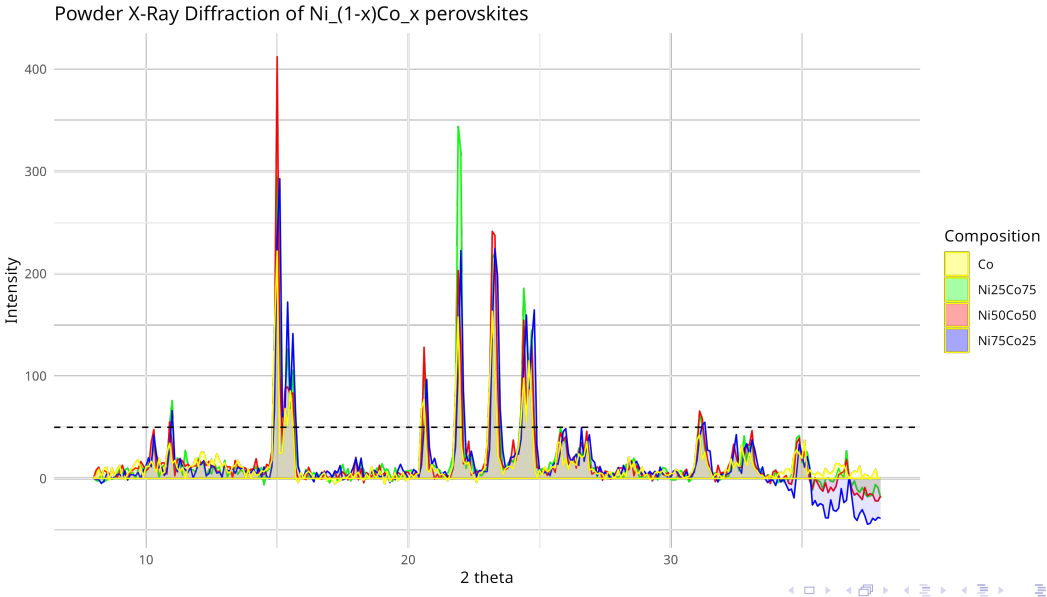
Exploratory Data Analysis

- Visualizations and summary statistics.
- Identification of patterns and trends.

```
    #Threshold value
      +
    geom_hline(
      yintercept = threshold,
      linetype = "dashed", color = "black"
    )

```

Visualization of the Data

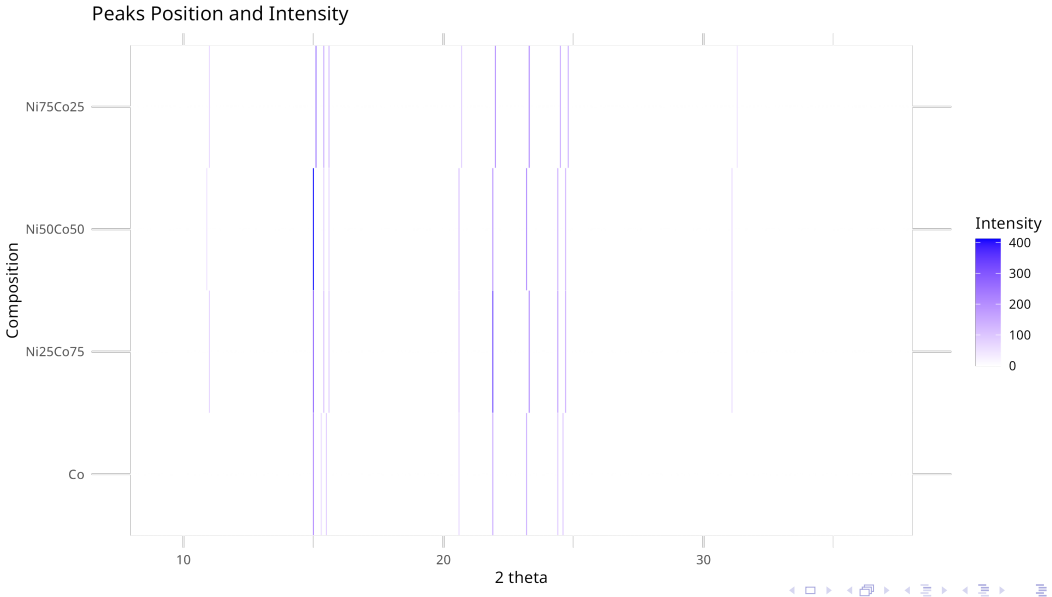


Peak Identification

- Overview of the importance of identifying peaks in X-ray diffraction data.
- Explanation of the methodology used for peak identification.

```r
find_peaks <- function(data) {
  if (length(data) < 2) {
    return(NULL) # No peak in lists with 0 or 1 element
  }
  peaks <- rep(0, length(data))
  # Create a vector to store the value and index of the peaks
  for (i in 2:(length(data) - 1)) {
    if (data[i] > data[i - 1] && data[i] > data[i + 1]) { # Looking for a peak
      peaks[i] <- data[i]
    }
  }
  # Checking if the first and last value is a peak or not
  if (data[1] > data[2]) {
    peaks[1] <- data[1]
  }
  if (tail(data, 1) > tail(data, 2)[1]) {
    peaks[length(peaks)] <- data[length(data)]
  }
  return(peaks)
  }
```

Peak Identification Results



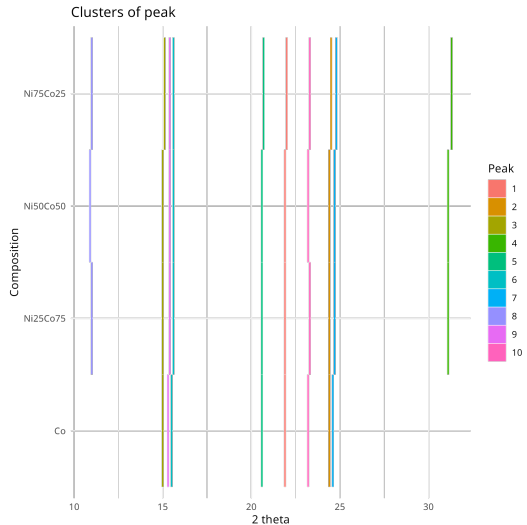
Clustering Process

- Explanation of the clustering algorithm used.
- Description of the data preparation steps.

```
1
2 # Specify the number of clusters (k)
3       k <- 10
4
5 # Perform k-means clustering based only on 2_theta
6       cluster_assignments <- kmeans(data_for_clustering,
7         centers = k,
8         nstart = 4
9       )$cluster
10
11
```

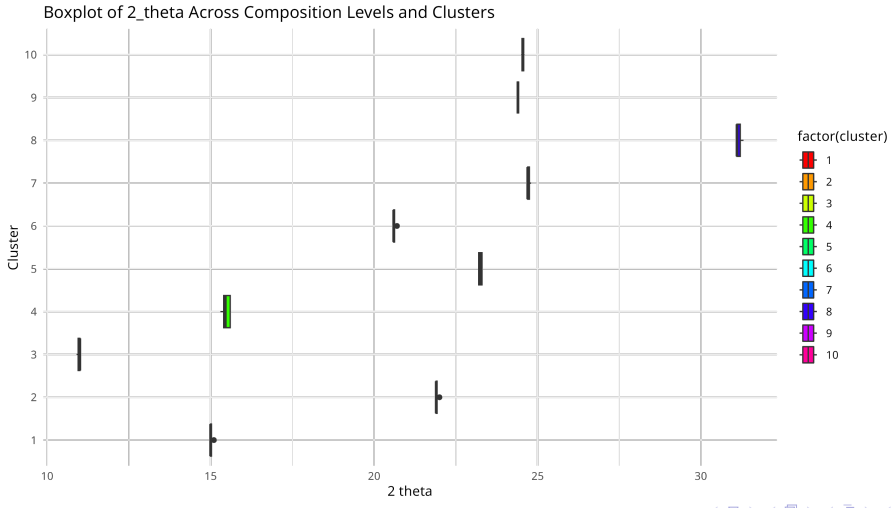
Clustering Results

- Overview of the clustering results.
- Interpretation of clusters and their characteristics.



Clustering Results

- Overview of the clustering results.
- Interpretation of clusters and their characteristics.



ANOVA Results

```r
# Perform ANOVA
anova_result_composition <- aov(`2_theta` ~ Composition *
  cluster, data = non_zero_data_long)

# Print ANOVA summary
summary(anova_result_composition)

```

ANOVA Results

Factor	Df	Sum Sq	Mean Sq	F Intensity	Pr(>F)
Composition	3	0.5	0.15	5.137	0.0286*
Cluster	9	1103.8	122.65	4175.310	1.03×10^{-13}***
Composition:Cluster	17	0.0	0.00	0.066	1.0000
Residuals	8	0.2	0.03		

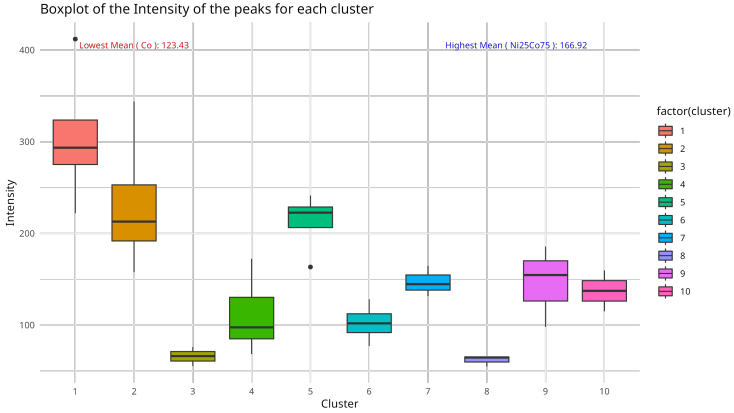
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of ANOVA Results

- The Composition factor shows a significant effect on the response variable (p-value $= 0.0286$).
- The Cluster factor has a highly significant impact on the response variable (p-value $= 1.03 \times 10^{-13}$).
- The interaction between Composition and Cluster is not significant (p-value $= 1.0000$).
- Residuals indicate the unexplained variance in the model.

Peak Intensity

- Powder X-ray Diffraction (PXRD) data often contains peaks that correspond to specific crystallographic planes.
- Peak intensity is a crucial parameter in PXRD analysis, reflecting the abundance or concentration of particular crystallographic phases.



ANOVA Results

- Analysis of Variance (ANOVA) was performed to assess the impact of 'Composition' and 'cluster' on the 'Intensity' variable.
- Statistical significance was evaluated based on p-values.

```
> anova_result <- aov(Intensity ~ Composition * cluster, data = non_zero_data_long)


> summary(anova_result)
                    Df Sum Sq Mean Sq F value   Pr(>F)
Composition          3   9496    3165   14.80 0.012440 *
cluster              9 220856   24540  114.78 0.000182 ***
Composition:cluster 21  31651    1507    7.05 0.035380 *
Residuals            4    855     214
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

- **Composition:** The p-value (0.012440) indicates a significant difference in means across 'Composition' levels.
- **Cluster:** A very low p-value (0.000182) suggests significant differences in means across clusters.
- **Interaction:** The interaction between 'Composition' and 'cluster' is significant (p-value = 0.035380).

Results and Discussion

- Interpretation of ANOVA results.
- Comparison of clusters: Assessing the significance of peak position variations.
- Implications of the findings: Understanding how composition affects peak positions.

Conclusion

Summary of key findings.

- Peak position,
- Intensity.

Limitations and areas for future research.

- **Ni** composition for further analysis,
- Improve the peak finding algorithm,
- Random variability in the clustering process.

Any Questions?