Statistical Data Analysis Project

Alexandre De Cuyper

University of A Coruña

February 22, 2024

Introduction

- Analyze the powder X-ray diffraction data from Complex Materials,
- 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 Processing

- Exporting the data from MATCH!
- Importing the data with R.

```
> 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
```

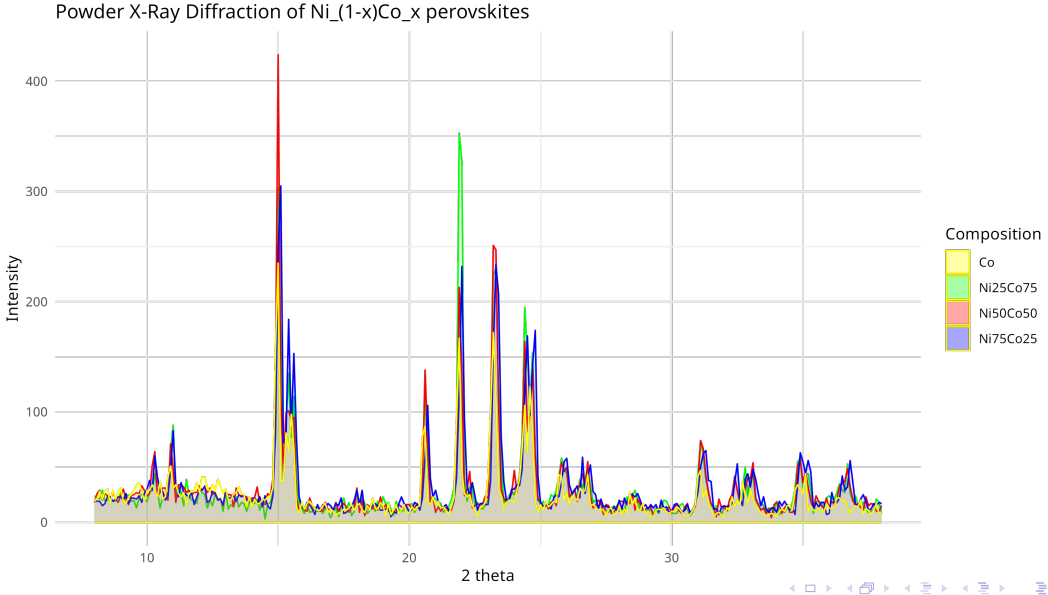
Exploratory Data Analysis

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

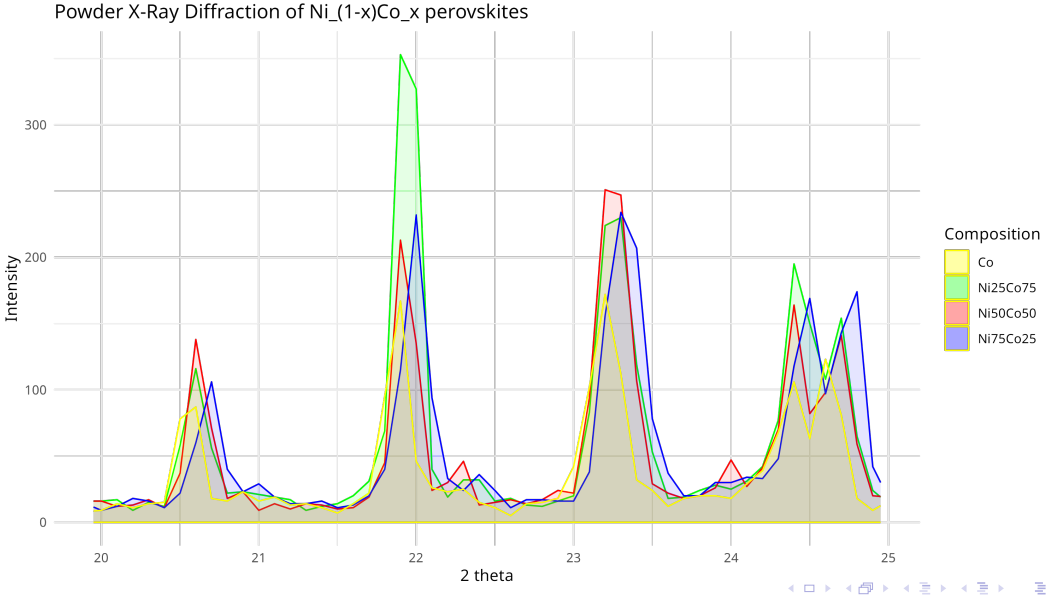
```
  # Exploratory Data Analysis (EDA)
  ggplot(data, aes(x = `2_theta`)) +
geom_ribbon(aes(ymin = 0, ymax = `Ni25Co75`, fill = "Ni25Co75"),
  alpha = 0.1, color = "green"
) +
geom_ribbon(aes(ymin = 0, ymax = `Ni50Co50`, fill = "Ni50Co50"),
  alpha = 0.1, color = "red"
) +
geom_ribbon(aes(ymin = 0, 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



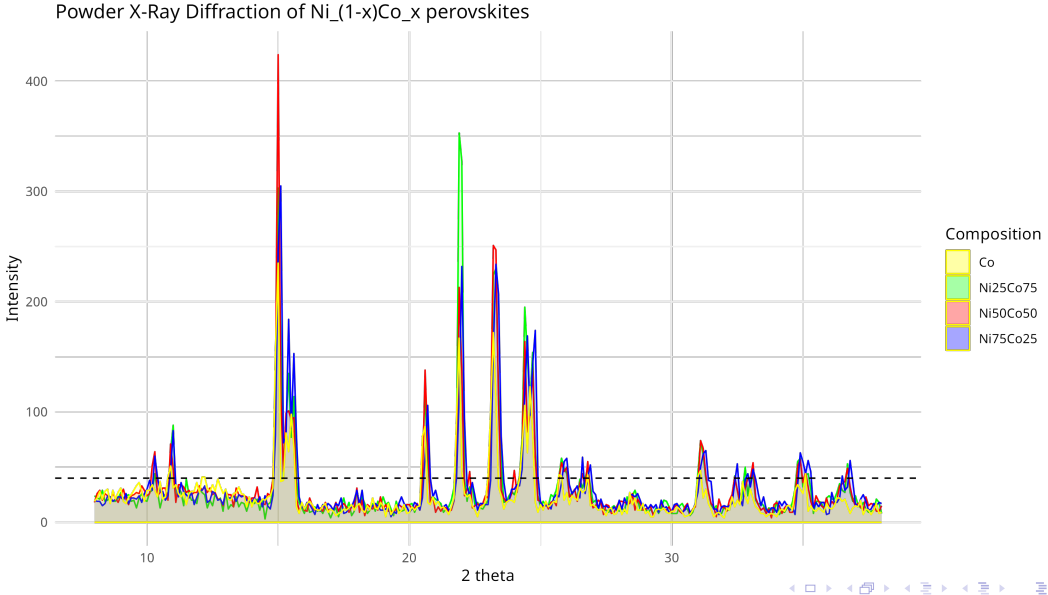
Exploratory Data Analysis

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

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

```

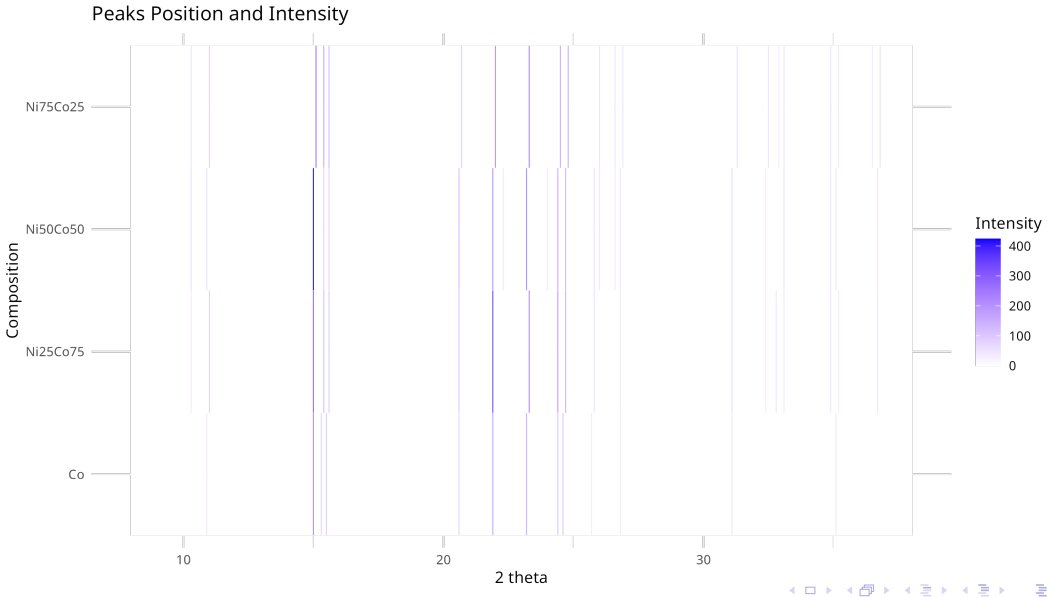
Visualization of the Data



Peak Identification

```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



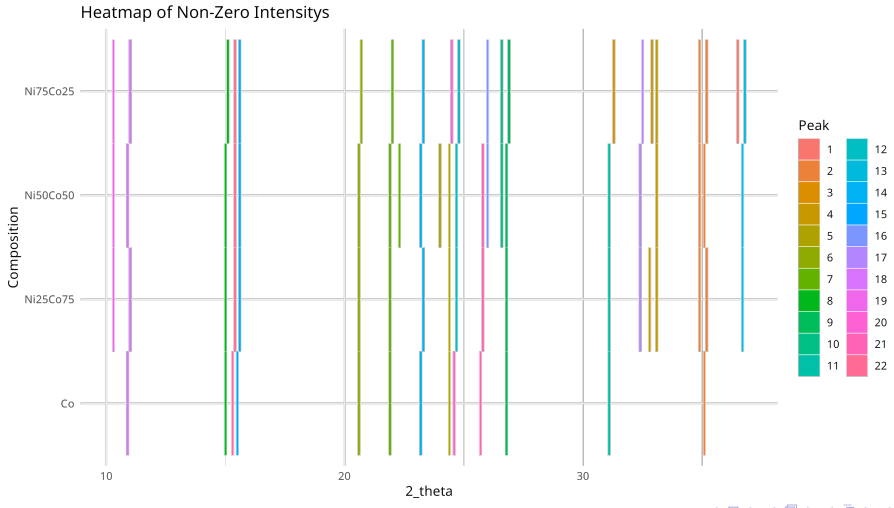
Clustering Process

- K-means clusters data by iteratively assigning points to the nearest cluster center and updating those centers until convergence, aiming to group similar points together.

```
1
2  # Specify the number of clusters (k)
3        k <- 22
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
```

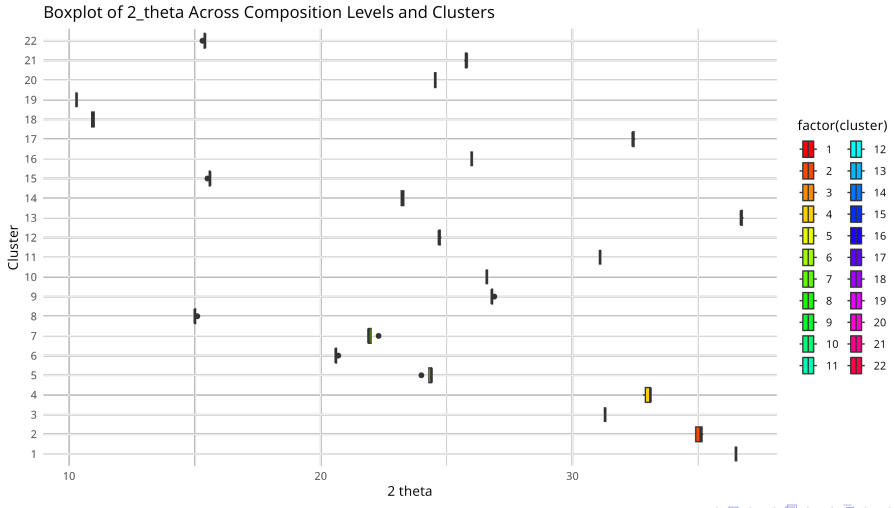
Clustering Results

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



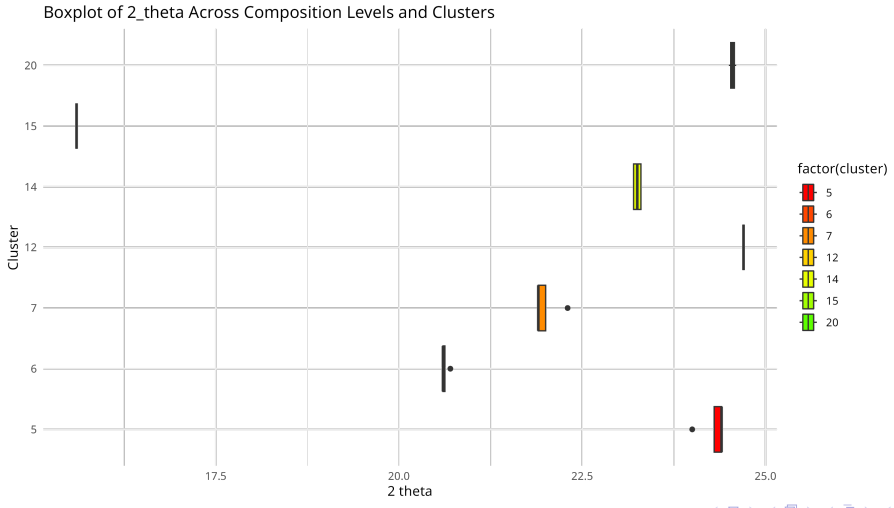
Clustering Results

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



Clustering Results

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



ANOVA Test

```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

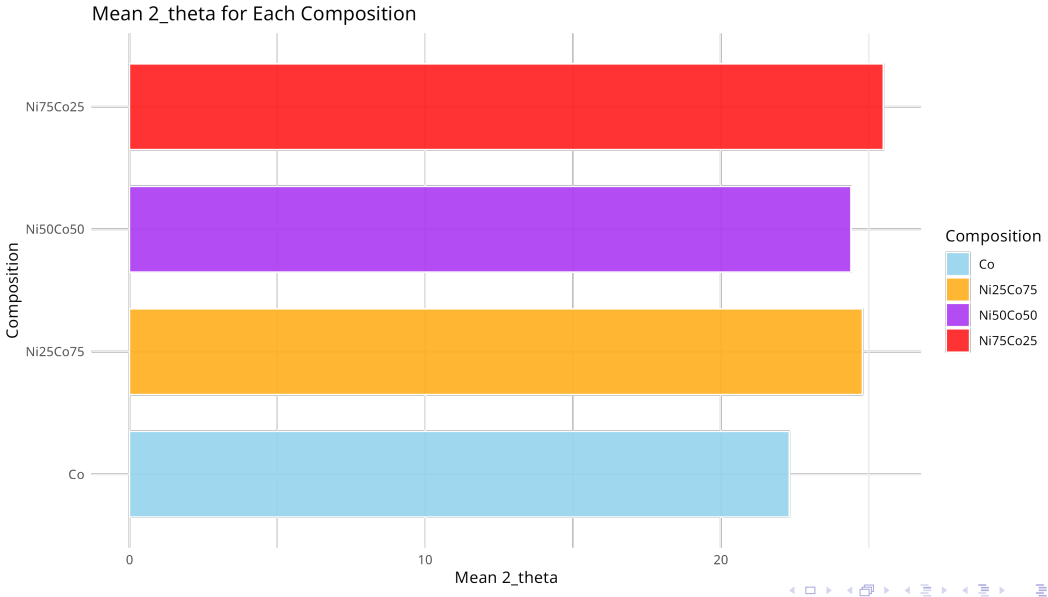
Source	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Composition	3	156.5	52.15	81.746	0.000114***
Cluster	15	2896.1	193.08	302.627	2.33e-06***
Composition:cluster	31	0.9	0.03	0.044	1.000000
Residuals	5	3.2	0.64		

Table: ANOVA Results for `2theta` with Composition and Cluster.

Analysis of ANOVA Results

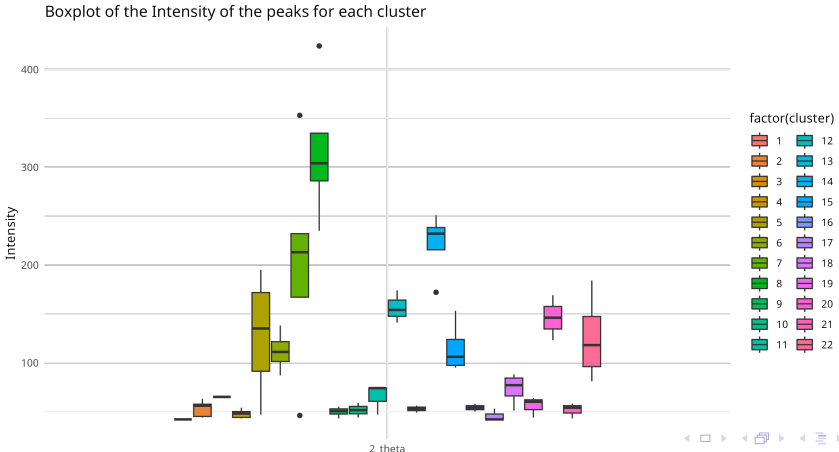
- The Composition factor shows a significant effect on the response variable (p-value = 0.000114).
- The Cluster factor has a highly significant impact on the response variable (p-value = 2.33e-06).
- The interaction between Composition and Cluster is not significant (p-value = 1.0000).
- Residuals indicate the unexplained variance in the model.

Results


Mean 2_theta for Each Composition
Ni75Co25
Ni50Co50
Ni25Co75
Co
Composition
0
10
20
Mean 2_theta
Composition
Co
Ni25Co75
Ni50Co50
Ni75Co25

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.



Peak Intensity Analysis

Composition	Mean Intensity	Variance Intensity
Co	99.8	3653.0
Ni25Co75	116.0	8636.0
Ni50Co50	104.0	8514.0
Ni75Co25	109.0	6062.0

Table: Mean and Variance Intensities for Each Composition.

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)
```

Factor	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	–	–

Table: ANOVA Summary

ANOVA Results

- **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).

Conclusion

Summary of key findings.

- Significant peak position and intensity variation depending on the composition,
- Implies a change in the cristalline structure of the material.

Limitations and areas for future research.

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

Any Questions?