INFORME DE PROYECTO DEL CURSO: “INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R”

Alumno: Federico Ricardo Checozzi

Profesores: Dr. Ignacio E. Schor

Lic. Martín García Solá

Lic. Nicolás Méndez

Datos utilizados en el proyecto

En la carpeta “data” del proyecto hay varios archivos .csv con espectros obtenidos del análisis de ciertos compuestos químicos, cuyos nombres consisten de algunas letras correspondientes a un compuesto específico más un número que corresponde a una muestra (ej.: PE1.csv).

Cada archivo posee una tabla cuya primera columna corresponde a longitudes de onda; las columnas restantes corresponden a un “barrido” espectral (cada celda tiene el nivel de intensidad observado a cierta longitud de onda). La primera línea es ignorada cuando se cargan los archivos. La tabla final junta la información de los archivos cargados más una columna para identificar su procedencia.

	A	B	C	D	E	F	G	H	I	J	K	L
1		ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m	ZnPENT_15m
2	Wavelength	Intensity	intensity	intensity	intensity	intensity	intensity	intensity	intensity	intensity	intensity	intensity
3	275,726	18	223	361	440	386	464	381	439	399	407	440
4	276,055	47,721	195,2	390,72	535,87	473,24	544,53	396,34	526,24	479,53	450,14	442,88
5	276,384	11,381	210,52	458,98	481,28	468,74	473,68	500,68	503,39	529,79	542,84	442,08
6	276,713	40,422	250,56	436,08	524,19	497,79	501,17	522,27	532,55	400,81	485,28	562,05
7	277,042	20,783	253,49	509,66	556,05	513,69	510,18	533,19	485,23	487,22	488,53	498,83

⇒

Archivo	Longitud de onda	Barrido 1	Barrido 2	…	Barrido 11
“PEi”	W1	I1-1	I1-2	…	I1-11
“PEi”	W2	I2-1	I2-2	…	I2-11
…	…	…	…	…	…
“PEi”	Wn	In-1	In-2	…	In-11
…	…	…	…	…	…

Objetivo del análisis a realizar

La meta final del procesamiento de estos espectros químicos es la clasificación de compuestos. Dado que eso está fuera del alcance del curso, lo que se busca realizar en este proyecto es cierto procesamiento complementario:

- Pre-procesar los espectros,
- explorar en busca de anomalías,
- observar la relación entre barridos de una misma muestra, entre muestras del mismo compuesto y entre diferentes compuestos,
- analizar cuán útil es realizar reducciones dimensionales sobre datos,
- tratar de encontrar longitudes de onda que permitan distinguir entre químicos,
- extraer algunas características (y compararlos con la información original).

Procedimiento y resultados

Ordenamiento de datos:

Los espectros obtenidos deben estar en dos formatos para realizar los análisis deseados: "Tidy" (cada fila corresponde a un barrido y la columna a variables, en este caso las longitudes de onda y variables categóricas) y "Gathered" (hay ciertos procesamientos en los que conviene que las longitudes de onda estén en una variable numérica).

Tidy:

	Archivo	Grupo	NBarrido	W1	…	Wn
PEi-Bj	"PEi"	"PE"	"Barridoj"	I1ij	…	Inij
…	…	…	…	…	…	…

Para lograrlo once columnas con información de intensidad fueron convertidas a tres columnas, una indicando el barrido y las otras dos las intensidad y longitud de onda. De ahí se convirtieron las longitudes de onda a variables como se ve en la tabla de arriba.

Gathered:

Archivo	Grupo	NBarrido	Longitud de onda	Intensidad
"PEi"	"PE"	Barridoj	W1ij	I1ij
…	…	…	…	…
"PEi"	"PE"	Barridoj	Wnij	Inij
…	…	…	…	…

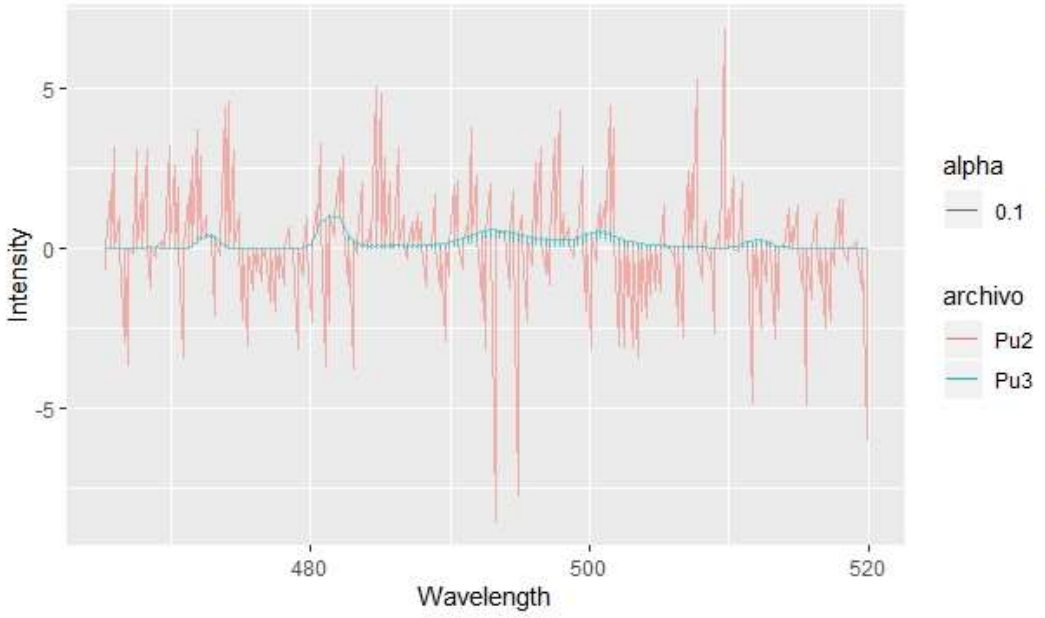
En este caso las longitudes de onda volvieron a ser una columna.

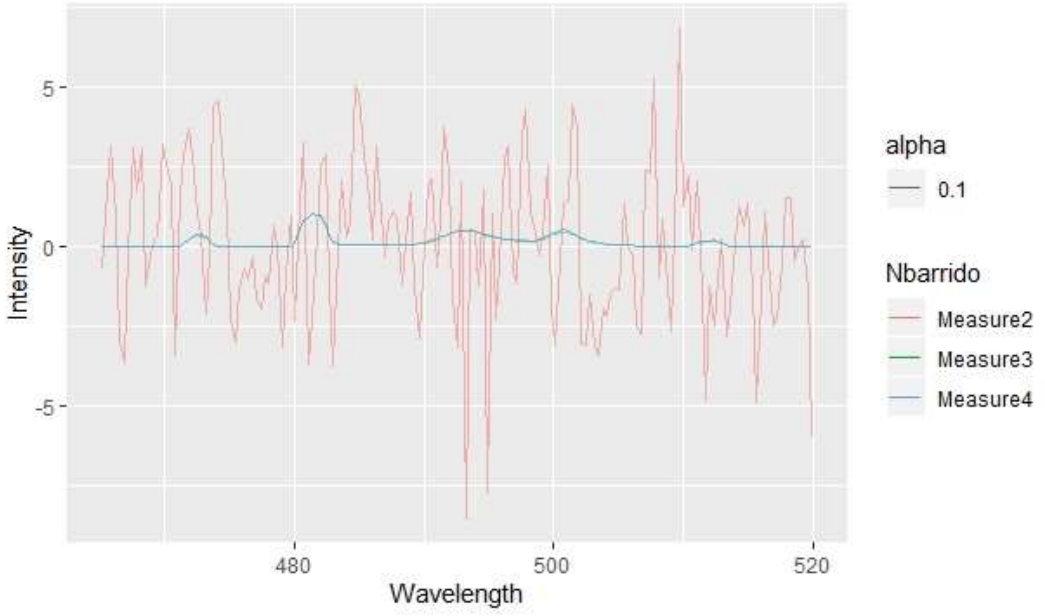
Pre-procesamiento:

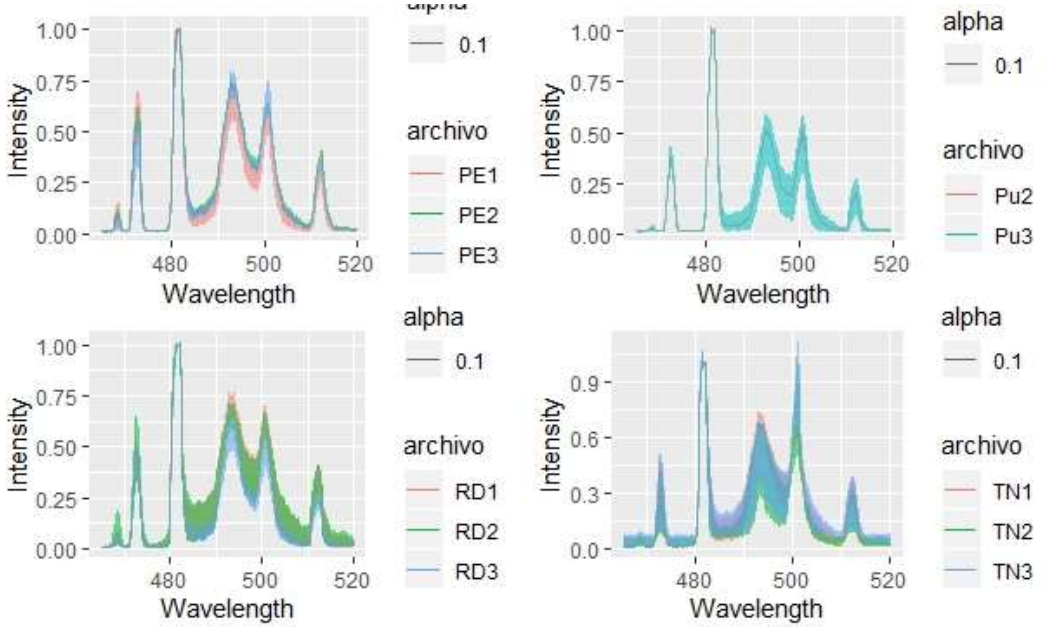
Las intensidades de los espectros fueron normalizadas utilizando los valores de intensidad registrados a una longitud de onda específica, y las longitudes de onda fueron limitadas a un cierto rango, en base a valores sugeridos por quienes realizaron las mediciones (todos los valores empleados fueron extraídos de "explosivos.py", dentro de la carpeta "R").

Exploración de los datos:

La forma más básica de visualizar la información contenida en los espectros es simplemente graficar la intensidad medida en función de la longitud de onda. Gráficos por agregación de todos los espectros de un grupo químico específico permiten visualizar cualquier desviación importante. De esta forma se encontró un grupo con una anomalía, y una inspección posterior a nivel barrido permitió eliminar un espectro de mala calidad.

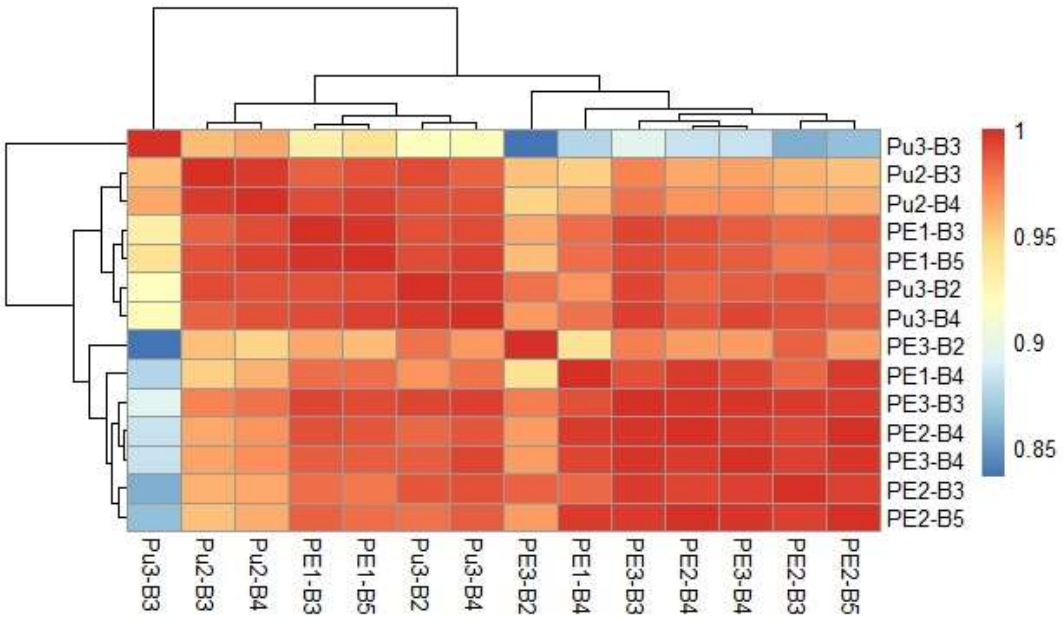

5
0
-5
Intensity
480
500
520
Wavelength
alpha
0.1
archivo
Pu2
Pu3

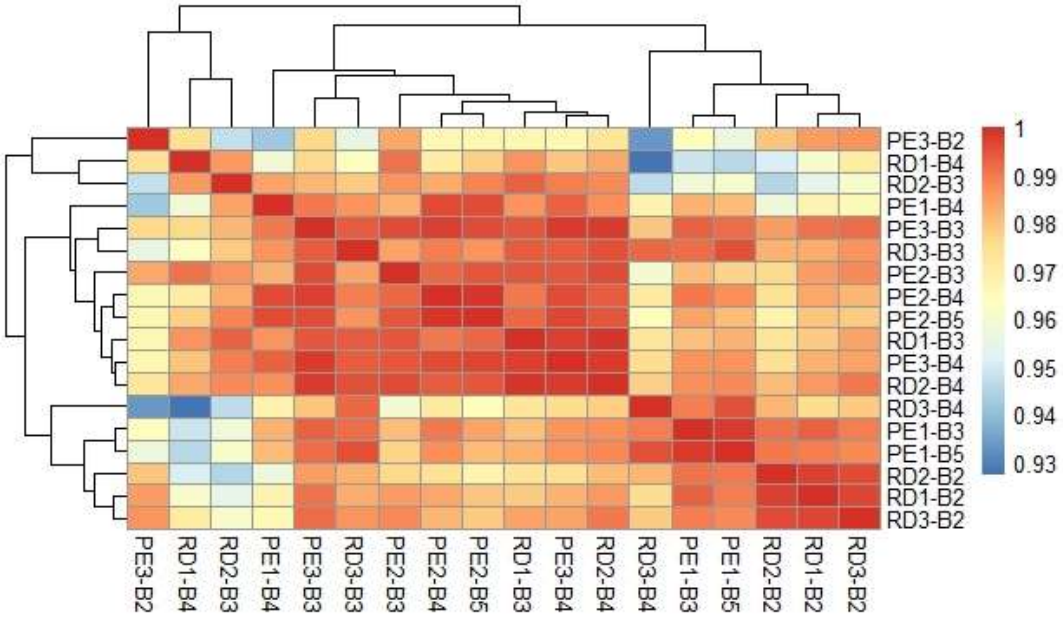

5
0
-5
Intensity
480
500
520
Wavelength
alpha
0.1
Nbarrido
Measure2
Measure3
Measure4

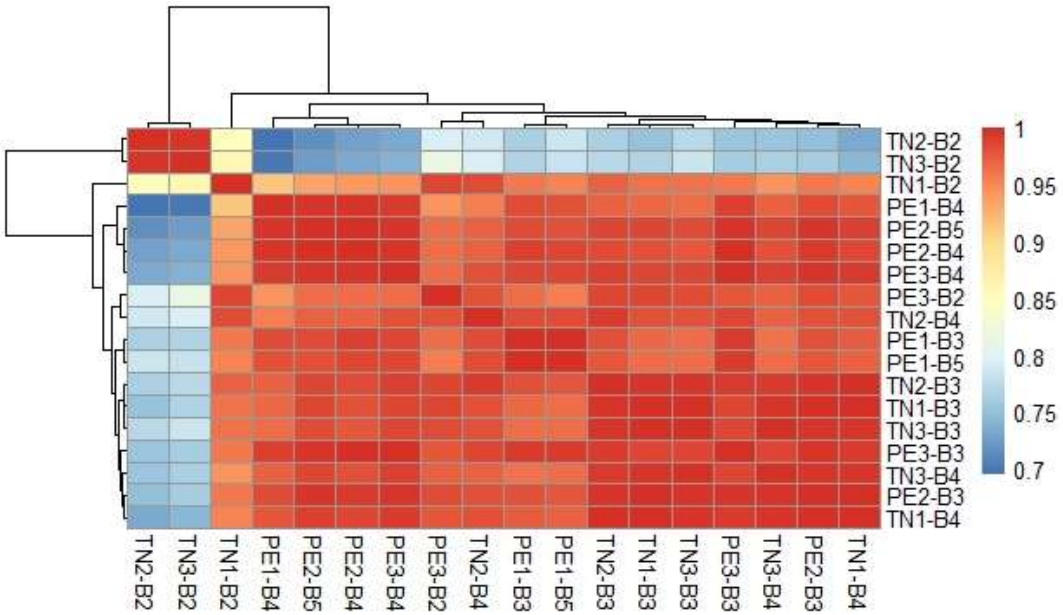


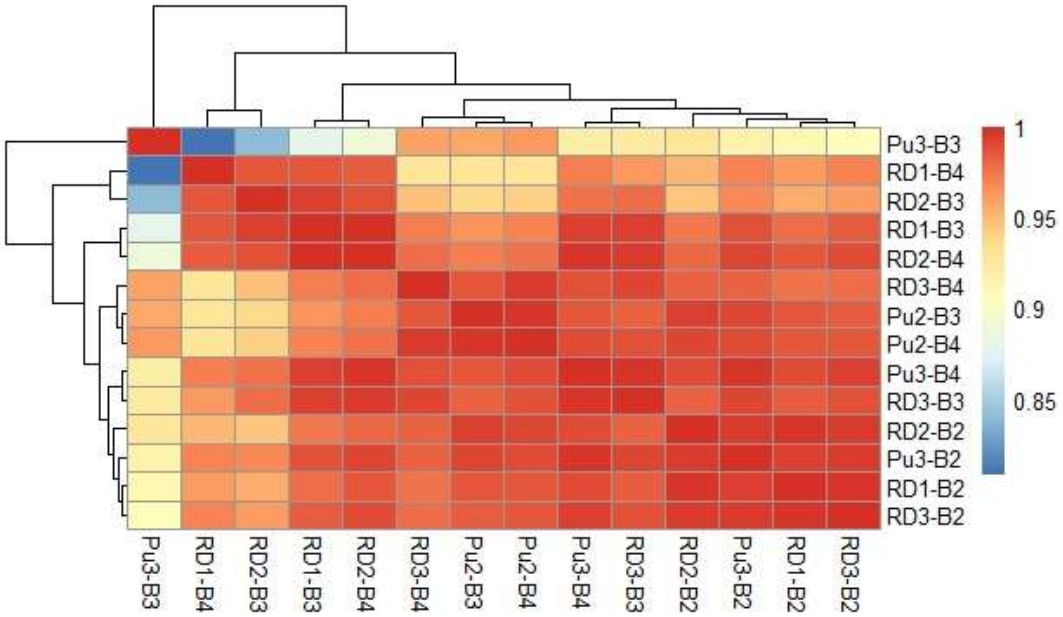
También se observó que a simple vista los espectros no eran tan diferentes entre grupos.

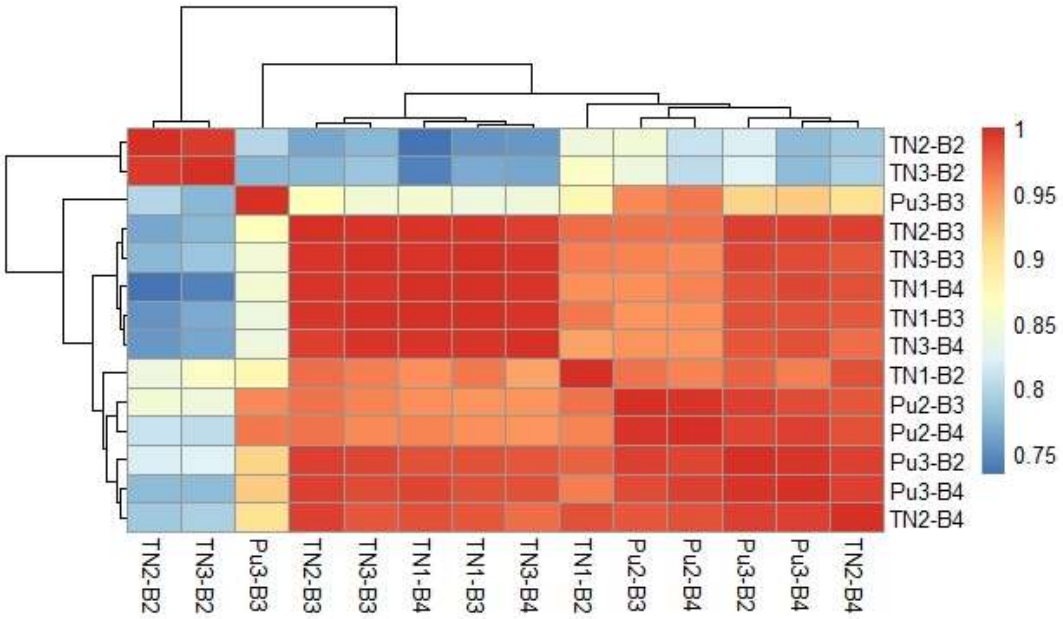
Posteriormente se graficaron varios mapas de calor de las correlaciones entre diferentes barridos de dos grupos de compuestos, de forma de visualizar las similitudes y diferencias entre compuestos del mismo tipo o de tipos diferentes.

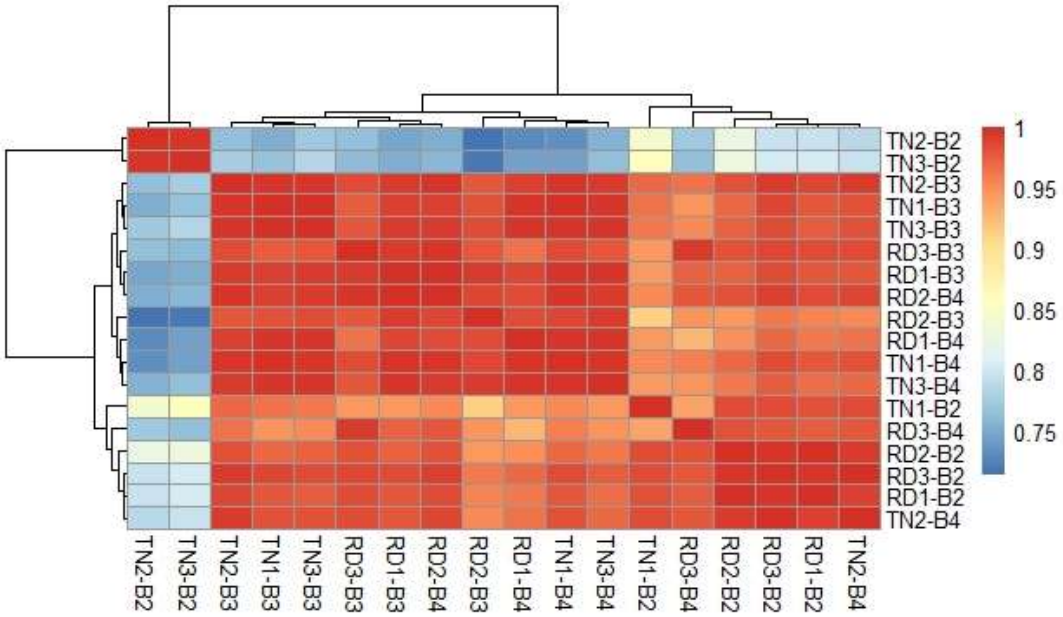



PE3-B2
RD1-B4
RD2-B3
PE1-B4
PE3-B3
RD3-B3
PE2-B3
PE2-B4
PE2-B5
RD1-B3
PE3-B4
RD2-B4
RD3-B4
PE1-B3
PE1-B5
RD2-B2
RD1-B2
RD3-B2
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93


TN2-B2
TN3-B2
TN1-B2
PE1-B4
PE2-B5
PE2-B4
PE3-B4
PE3-B2
TN2-B4
PE1-B3
PE1-B5
TN2-B3
TN1-B3
TN3-B3
PE3-B3
TN3-B4
PE2-B3
TN1-B4
1
0.95
0.9
0.85
0.8
0.75
0.7

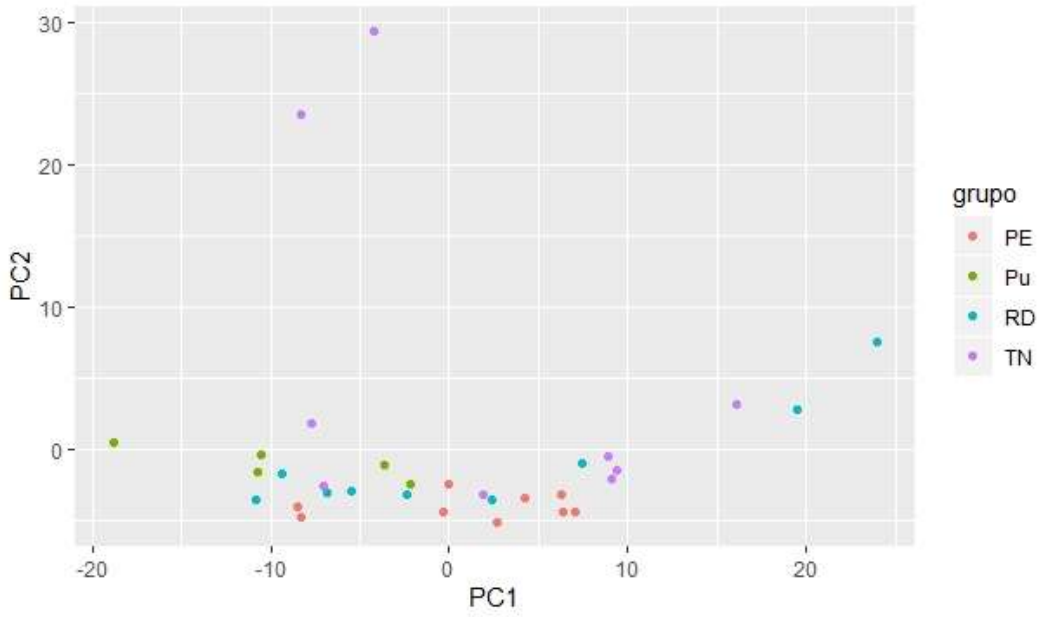

Pu3-B3
RD1-B4
RD2-B3
RD1-B3
RD2-B4
RD3-B4
Pu2-B3
Pu2-B4
Pu3-B4
RD3-B3
RD2-B2
Pu3-B2
RD1-B2
RD3-B2
1
0.95
0.9
0.85


TN2-B2
TN3-B2
Pu3-B3
TN2-B3
TN3-B3
TN1-B4
TN1-B3
TN3-B4
TN1-B2
Pu2-B3
Pu2-B4
Pu3-B2
Pu3-B4
TN2-B4
1
0.95
0.9
0.85
0.8
0.75



En general no se observó una separación demasiado notoria como se hubiera esperado, en donde haya más parecido entre barridos del mismo tipo y más diferencias entre barridos de diferentes tipos.

Finalmente se realizó una reducción dimensional de los datos mediante el análisis de componentes principales, de forma de centrarse en la varianza de los espectros para encontrar alguna diferencia entre grupos.



No pudo observarse ninguna agrupación clara, y un análisis posterior mediante ANOVA para PC1 y Kruskal-Wallis para PC2 (componentes principales de mayor peso confirmó) no encontró diferencias significativas entre grupos en el primer caso pero sí encontró algo para el segundo caso (específicamente es capaz de diferenciar entre el grupo PE y otros, aunque dado la falta

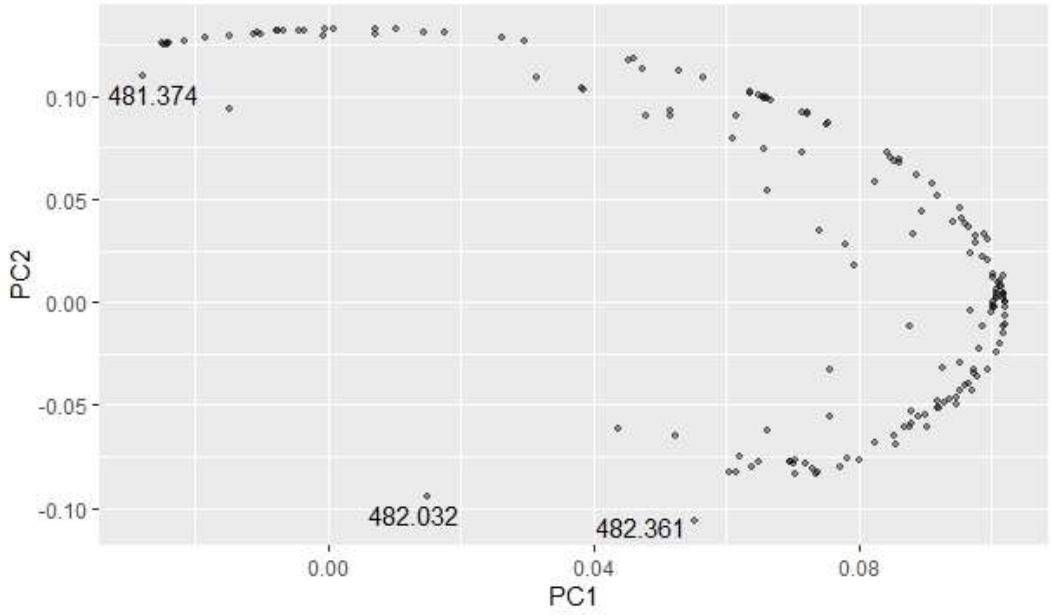
de un grupo de control y lo observado en los gráficos tendría que hacerse pruebas adicionales antes de confirmarlo, es un tanto dudoso como resultado).

```
> aov_res <- aov(PC1 ~ grupo, data = df_out)
> summary (aov_res)#PC1 no es suficientemente bueno para dividir entre grupos
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        3  505.3  168.42   1.962  0.143
Residuals   28 2403.0   85.82
> TukeyHSD(aov_res)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PC1 ~ grupo, data = df_out)

$grupo
              diff        lwr       upr     p adj
Pu-PE -10.23847179 -24.346630  3.869686 0.2189612
RD-PE   0.97416383 -10.949406 12.897734 0.9959910
TN-PE   0.95882928 -10.964740 12.882399 0.9961749
RD-Pu  11.21263562  -2.895522 25.320794 0.1564801
TN-Pu  11.19730107  -2.910857 25.305459 0.1573424
TN-RD  -0.01533455 -11.938904 11.908235 1.0000000
```

```
> kruskal.test(PC2 ~ grupo, data = df_out)#aunque PC2 parece ser interesante

        Kruskal-Wallis rank sum test

data:  PC2 by grupo
Kruskal-Wallis chi-squared = 16.649, df = 3, p-value = 0.0008345

> pairwise.wilcox.test(df_out$PC2, df_out$grupo,p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  df_out$PC2 and df_out$grupo

   PE     Pu     RD
Pu 0.0030 -      -
RD 0.0080 0.3572 -
TN 0.0017 0.6993 0.2039

P value adjustment method: BH
```

Búsqueda de longitudes de onda de interés:

Para este propósito, se trató de identificar las longitudes de onda que tuvieran más peso en la varianza total. Esto es, aquellas longitudes de onda cuyos coeficientes en la matriz de rotación correspondientes a las primeras dos componentes principales sean valores extremos. Para ello se graficó los coeficientes de PC1 vs. PC2 y se identificó aquellos que tuvieran su posición más allá de un área rectangular alrededor del cero en el gráfico.



Una vez encontradas estas frecuencias se examinó si eran útiles mediante tests de Kruskal-Wallis.

```
> kruskal.test(`481.374` ~ grupo, data = datos.tidy)#no es significativo

        Kruskal-Wallis rank sum test

data:  481.374 by grupo
Kruskal-Wallis chi-squared = 6.6601, df = 3, p-value = 0.08356

> kruskal.test(`482.032` ~ grupo, data = datos.tidy)#no es significativo

        Kruskal-Wallis rank sum test

data:  482.032 by grupo
Kruskal-Wallis chi-squared = 4.5793, df = 3, p-value = 0.2053

> kruskal.test(`482.361` ~ grupo, data = datos.tidy)#no es significativo

        Kruskal-Wallis rank sum test

data:  482.361 by grupo
Kruskal-Wallis chi-squared = 7.0167, df = 3, p-value = 0.07137
```

En este caso la separación entre grupos no parece ser adecuada en ninguna de las longitudes de onda seleccionadas.

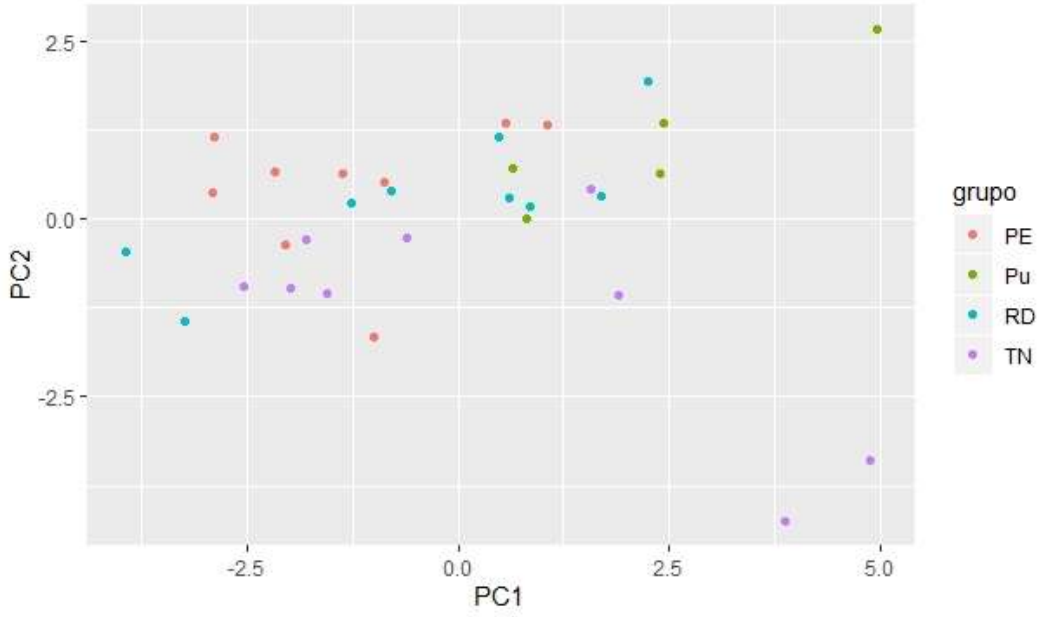
Extracción de características de los espectros:

Por último, se creó una tabla con información extraída de los espectros, con la esperanza de que este nuevo conjunto de datos permita separar los grupos de compuestos mejor. Para ese propósito se hizo una nueva tabla resumen en base a ocho “estadísticas” para cada barrido (cuatro picos a ciertas longitudes de onda decididas empíricamente, y cuatro integrales de intensidad en los picos observados, una especie de medida de la energía en ciertas longitudes de onda). En este proceso se empleó la tabla “Gathered” agrupando por barrido.

Features:

	Archivo	Grupo	NBarrido	P1	…	I1	…
PEi-Bj	“PEi”	“PE”	“Barridoj”	P1ij	…	Inij	…
…	…	…	…	…	…	…	…

Se realizó un PCA más MANOVA para hacer una estimación de qué tan adecuadamente las nuevas variables permiten discriminar entre distintos compuestos.



Mejora un poco el clustering respecto del gráfico de PCA anterior, pero no por mucho.

```
> manova_res <- manova(cbind(PC1,PC2) ~ grupo, data = featureout)
> summary(manova_res)
          Df  Pillai approx F num Df den Df    Pr(>F)
grupo      3 0.65109    4.505      6     56 0.0008632 ***
Residuals 28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.aov(manova_res)#dividido por el número de test sólo es interesante PC2
 Response PC1 :
            Df  Sum Sq Mean Sq F value  Pr(>F)
grupo        3  43.141 14.3803  3.2465 0.03671 *
Residuals   28 124.025  4.4294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response PC2 :
            Df Sum Sq Mean Sq F value   Pr(>F)
grupo        3 23.794  7.9315  6.0433 0.002627 **
Residuals   28 36.748  1.3124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dado que tenemos dos tests en el MANOVA hay que ajustar los valores p por dos, de manera de que sólo nos interesa analizar lo que ocurre en PC2.

```
> aov_res <- aov(PC2 ~ grupo, data = featureout)
> summary (aov_res)
            Df Sum Sq Mean Sq F value  Pr(>F)
grupo        3  23.79   7.931   6.043 0.00263 **
Residuals   28  36.75   1.312
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(aov_res)#las únicas diferencias significativas las veo para comparar TN con
 el resto
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PC2 ~ grupo, data = featureout)

$grupo
             diff       lwr        upr     p adj
Pu-PE  0.6329766 -1.111670  2.3776235 0.7559187
RD-PE -0.1545089 -1.629005  1.3199868 0.9916526
TN-PE -1.7549620 -3.229458 -0.2804662 0.0149987
RD-Pu -0.7874855 -2.532132  0.9571614 0.6120757
TN-Pu -2.3879386 -4.132585 -0.6432917 0.0044386
TN-RD -1.6004531 -3.074949 -0.1259573 0.0295394
```

Tras una inspección más cuidadosa de PC2 parece ser que podría separar el grupo TN del resto, pero (de forma similar a como se vio en un caso anterior) se necesitaría un mejor diseño experimental y observar una mejor separación entre grupos antes de creer este resultado.

Conclusión

No fue posible hacer una discriminación adecuada entre grupos con los métodos utilizados hasta ahora. Hay algunos resultados que podrían ser falsos positivos, aunque al menos la metodología empleada para este trabajo es aplicable a análisis futuros.

Dicho eso, algunas de las posibles alternativas para mejorar sobre lo realizado en este proyecto son:

-Mejorar los métodos de exploración de datos, por ejemplo buscar regiones con longitudes de onda de interés en vez de utilizar una región elegida empíricamente. Áreas con variaciones importantes así como también consideraciones sobre las naturalezas de los espectros contribuirían en ese sentido.

-Investigar si conviene eliminar algunos outliers.

-Buscar otras características a extraer que sean más representativas de los compuestos.

-En el lado experimental, convendría agregar un grupo de control para mejorar la calidad de los tests realizados.