Taller de Consultoria - TP5

Gonzalo Barrera Borla

20/12/2019

Setup

```
library(tidyverse) # manipulación de datos en general, graficos
library(leaps) # Estimadores forward, backward, exhaustive & stepwise
library(broom) # limpieza y estructuracion de resultados de regresiones
```

Ejercicio 1

> El "Bovine viral diarrhea virus" (BVDV) afecta el sistema gastrointestinal del ganado causando diarrea grave. Es especialmente dañino para los animales preñados, por su capacidad para infectar el feto. Se desea estimar la concentración del virus en una solución. El BVDV tiene la propiedad de que cuando se lo cultiva en un disco de Petri, las partículas virales forman placas, regiones circulares en el medio de cultivo. Estas placas son fácilmente visibles al microscopio o directamente al ojo cuando se usa un tinte. Cada placa está asociada con una sola partícula viral. Contando el número de placas -las llamadas unidades formadoras de placas (UFP) por ml. de volumen, se puede estimar la concentración viral total.
>
> Una dificultad para estimar la concentración total es que con una concentración suficientemente grande, todo el disco de Petri se convierte en una única gran placa, y es imposible contar las UFP individuales. Para obtener placas contables se debe diluir la solución. Esto se hace en forma seriada. En el primer paso una parte de la solución se mezcla con dos partes de solución estéril. Esto continúa de la misma manera, de modo que en la dilución d, una parte de la solución d-1 se mezcla con dos partes de solución estéril. En cada paso la contaminación es 1/3 de la del paso anterior. En cada paso, 4 discos de Petri con una capacidad de 3 ml. cada uno ("réplicas") se preparan con el material de esa dilución. Con las primeras diluciones, los discos de Petri producen innumerables placas por la sobreabundancia de partículas virales. Para poder analizar los resultados, hacen falta al menos 2 diluciones que produzcan un número de placas contable (pero no nulo).
>
> El proceso es: contaminar, diluir, cultivar, y contar. Como la variabilidad puede afectar cada paso, el proceso se repite varias veces (llamadas muestras). Las preguntas son:
>
> 1. Dados los resultados de UFP de las diluciones seriadas, ¿cómo estimar la concentración viral en la solución original (sin diluir) y la precisión del estimador?

Antes que nada, un poco de notación. $[n, m] = \{n, n+1, \ldots, m\}$ denotará el conjunto de los naturales positivos de n hasta m inclusive, y omitiremos n cuando sea igual a 1 ($[m] = [1, m]$). X_{mdr} será la variable aleatoria que representa la cantidad de UFP en el cultivo de la muestra $m \in [9]$, en la dilución $[d] \in [3, 9]$ y la réplica $r \in [4]$. Llamaremos ρ a la concentración viral total (medida en UFP/ml.), y ρ_m a la concentración viral en la muestra m sin diluir. Como las UFP son relativamente escasas, la concentración viral total no se puede asumir uniforme, y por ende $\rho_m \neq \rho$. En su lugar, diremos que la concentración de las muestras son VAIID con $E(\rho_m) = \rho$. Con estimaciones de $\hat{\rho}_m \; \forall \, m \in [9]$, podemos estimar tanto la media como el desvio de ρ.

De la descripción del proceso, parece razonable suponer que las X_{mdr} se distribuyen independientemente como Poisson(λ_{mdr}), donde el parámetro λ_{mdr} depende de la concentración inicial de la muestra, el número de diluciones realizadas, el volumen de cada cultivo (que en este caso, es siempre 3ml.), y *no* depende de la réplica. Como en los datos la primer dilución contable figura como $d = 3$, pero el enunciado menciona que ese

dato se obtiene luego de *dos* diluciones, asumimos que $d = 1$ se corresponde a la solución original sin diluir, y la fórmula de λ_{mdr} resulta:

$$\lambda_{mdr} = \rho_m \times \left(\tfrac{1}{3}\right)^{d-1} \times v_{mdr}$$

Como asumimos que las X_{mdr} son independientes entre sí, su función de densidad conjunta es igual al producto de las densidades individuales, y la estimación de los ρ_m es bastante directa. Llamamos $y = (x_{111}, x_{112}, \ldots, x_{994})$ al vector de todos los conteos, $\gamma = (\rho_1, \ldots, \rho_9)$ al vector de todas las concentraciones muestrales, y planteamos la función de verosmilitud y su logaritmo natural:

$$\begin{aligned} L(\gamma|y) &= \prod_{(m,d,r)\in[9]\times[3,9]\times[4]} \frac{\lambda_{mdr}^{x_{mdr}} e^{-\lambda_{mdr}}}{x_{mdr}!} \\ \ln L(\gamma|y) &= \sum_{(m,d,r)\in[9]\times[3,9]\times[4]} x_{mdr} \ln(\lambda_{mdr}) - \lambda_{mdr} - \ln(x_{mdr}!) \end{aligned}$$

Reemplazando λ_{mdr} por su expresión y derivando $\ln L$ con respecto de cierto ρ_i resulta:

$$\begin{aligned} \ln L(\gamma|y) &= \sum_{(m,d,r)\in[9]\times[3,9]\times[4]} x_{mdr} \left[\ln(\rho_m) + (d-1)\ln(1/3) + \ln(v_{mdr})\right] - \rho_m \left(\tfrac{1}{3}\right)^{d-1} v_{mdr} - \ln(x_{mdr}!) \\ \frac{d}{d\rho_i} \ln L(\gamma|y) &= \sum_{(d,r)\in\times[3,9]\times[4]} x_{idr}\rho_i^{-1} - \left(\tfrac{1}{3}\right)^{d-1} v_{mdr} \end{aligned}$$

Como $v_{mdr} = v = 3\text{ml}$ para todos los cultivos, la ecuación se puede simplificar un poco más. Llamando $n_r = 4$ al número de réplicas y $\bar{x}_{md} = n_r^{-1} \sum_i x_{mdi}$ al promedio de conteos entre las réplicas de una misma muestra y dilución, obtenemos:

$$\frac{d}{d\rho_i} \ln L(\gamma|y) = \frac{n_r}{\rho_i} \sum_{d\in\times[3,9]} \bar{x}_{id} - n_r \cdot v \sum_{d\in[3,9]} \left(\tfrac{1}{3}\right)^{d-1}$$

Finalmente, reemplazando ρ_i por su EMV e igualando a cero, resulta:

$$\hat{\rho}_i = \frac{\sum_d \bar{x}_{id}}{v \sum_d \left(\tfrac{1}{3}\right)^{d-1}}$$

Al computar los EMV para cada muestra, obtenemos:

Muestra	EMV(ρ_m)
1	217.60
2	132.06
3	184.58
4	239.11
5	178.58
6	140.56
7	126.56
8	134.56
9	131.06

Aunque no observamos directamente ρ, todos los ρ_m son realizaciones de una VA con esperanza igual a rho. Por ende, nuestra mejor estimación $\hat{\rho}$ de la concentración viral total será el promedio de las estimaciones en cada muestra, que es 164.96 partículas virales por mililitro. El desvío estándar también se puede calcular, y es 41.93.

2. (Optativo) ¿Qué causa la mayor parte de la variación en el número de UFP/ml.?. ¿Se la puede atribuir a diferencias entre muestras y/o diferencias dentro de la mismas?.

Dentro del experimento, lo que causa la mayor parte de la variación en el número de UFP/ml, es la cantidad de diluciones realizadas. Luego, si queremos comparar la variabilidad *dentro* o *entre* muestras, habrá que descontar el efecto de las diluciones.

La forma que conocemos de estudiar las diferencias "entre/dentro" de cirtos grupos, es a través del análisis de la varianza, para lo cual necesitamos que los distintos grupos bajo estudio sean homocedásticos. Al estar tratando con VA de distribución Poisson, la transformación estabilizadora de la varianza que corresponde aplicar a los conteos, es la raíz cuadrada.

Consideraremos tanto a la muestra como también a la dilución como variables *categóricas*, ajustaremos un modelo multiplicativo ($\sqrt{ufp} \sim muestra \times dilución$), y compararemos las sumas de cuadrados de la tradicional tabla de ANOVA. Los gráficos de diagnóstico no muestran anomalías severas, así que presentamos directamente la tabla de suma de cuadrados para efectos principales, interacciones y "residuos":

Coefs	Suma de Cuadrados	P-valor
muestra	20.68	6.07e-12
dilucion	1565.48	7.01e-142
muestra:dilucion	37.75	1.49e-08
Residuals	47.36	NA

Desde un punto de vista explicativo, tanto el efecto principal de la muestra como las interacciones son relevantes, pero resulta evidente de los números que el factor más relevante, por lejos, es la cantidad de diluciones. Descontado este efecto, la segunda mayor suma de cuadrados es la de los "residuos", que nos está dando una medida de la variación *dentro* de cada grupo (muestra, dilución). En otras palabras, después de las diluciones, la mayor variabilidad parece provenir de diferencias entre réplicas de una misma muestra y dilución, en tercer lugar vienen los efectos de las interacciones (que interpretamos como "variabilidad proveniente de *cómo* se hizo cada dilución, para cada muestra") y en último lugar el efecto principal de las muestras, que da una idea de las diferencias *entre* ellas.

Ejercicio 2

Los siguientes datos corresponden a un trabajo para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico. Como el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo. Con este objetivo se tomó una muestra de 180 vasijas, a las que se realizó una espectrometria de rayos X sobre 1920 frecuencias, y también un análisis de laboratorio para determinar el contenido de 13 compuestos químicos, a saber:

Na_2O MgO Al_2O_3 SiO_2 P_2O_5 SO_3 Cl K_2O CaO MnO Fe_2O_3 BaO PbO

Cada fila del archivo Vessel_X es el espectro de una vasija, limitado a las frecuencias 100 a 400, pues las demás tienen valores casi nulos. Cada fila del archivo Vessel_Y tiene los contenidos de los 13 compuestos en esa vasija. Se trata de predecir el compuesto 1 (óxido de sodio) usando sólo las columnas 10, 20,... etc. de X.

1. Para familiarizarse con los datos, grafique en función de la frecuencia las medias y varianzas de X, y también algunos espectros.
2. Luego tome una muestra al azar de 120 vasijas. Con ella calcule los estimadores de mínimos cuadrados, Forward y Backward; y cualquier otro que se le ocurra. Para cada estimador estime el error cuadrático medio de predicción (ECM).
3. Luego aplíquelos a las otras 60 vasijas para estimar el error de predicción (este será insesgado). Compare los estimadores, y también compare las estimaciones del ECM con el inicial.

Observando detenidamente los datos y consultando con allegados capacitados en la materia, podemos concluir que

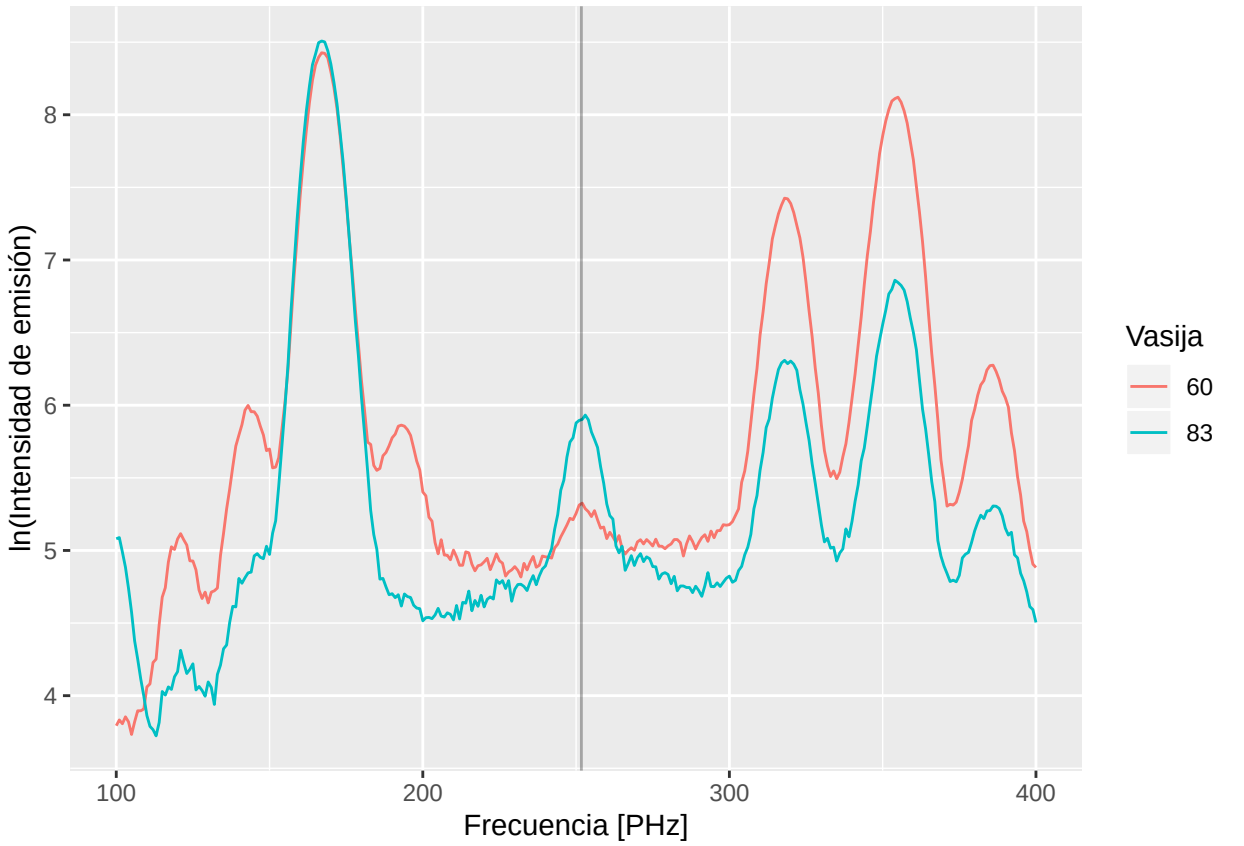
- la técnica de espectroscopía utilizada es la "fluorescencia de rayos X",
- las frecuencias del espectro están medidas en PHz (peta-hertz, $1 \times 10^{15}\mathrm{s}^{-1}$), y
- las "cantidades" de cada compuesto, son en realidad el porcentaje masa/masa (% m/m) de los óxidos en cada muestra ($100 \times \frac{\text{masa compuesto}}{\text{masa muestra}}$).

Al exponer un material a rayos X de longitudes de onda cortas o a rayos gamma, pueden ionizarse los átomos que constituyen el material. La ionización consiste en eyección de uno o más electrones desde el átomo. Tanto los rayos X como los gamma pueden ser suficientemente energéticos para desprender electrones fuertemente ligados en los orbitales internos del átomo. Tal remoción electrónica deja en condición inestable a la estructura electrónica del átomo, y los electrones de orbitales más elevados «caen» hacia el orbital más bajo, ocupando los "huecos" de los electrones internos desprendidos. En esta caída, o transición, se genera energía mediante emisión de un fotón.

Existe una cantidad finita de variantes de esta transición (entre pares de capas electrónicas) y a las transiciones principales se les han asignado nombres: $K_\alpha, K_\beta, L_\alpha, \ldots$. Cada una de estas transiciones produce un fotón fluorescente dotado de una energía característica que es igual a la diferencia de energía entre los orbitales inicial y final. La longitud de onda de esta radiación fluorescente se puede calcular a partir del postulado de Planck $\lambda = h \cdot c/E$, o si se prefiere, las longitudes de onda λ características se pueden expresar como frecuencias, $f = c/\lambda$.

Por ejemplo, la longitud de onda correspondiente a la linea espectral K_α para el sodio (Na) es de 1.191 nanómetros [1], que nos da una frecuencia equivalente de $\frac{299{,}792{,}458\ \mathrm{m/s}}{1.191 \times 10^{-9}\mathrm{m}} = 251.7 \times 10^{15}\mathrm{s}^{-1} = 251.7\,\mathrm{PHz}$.

A continuación, graficaremos los espectros (en escala logarítmica) para las vasijas 60 y 83, que son las de menor (0.986% m/m) y mayor (17.586% m/m) concentración de óxido de sodio, respectivamente. En gris, la línea K_α del sodio, ≈ 251.7PHz:



¡Eureka! Efectivamente, la intensidad de emisión alrededor de los 252 PHz aumenta significativamente para la vasija de mayor concentración de sodio.

Normalmente, un químico tiene una muestra pura del elemento cuya concentración se desea conocer, y utiliza su espectro como patrón de calibración para estimar la concentración en muestras de origen desconocido. Para aumentar la precisión, lo que se mide no es directamente la intensidad de la señal en la linea de emisión teórica, sino el *área debajo del espectro*, a la que se le resta una *línea de base* de "ruido".

Trabajar de esta manera seguramente sirva para mejorar las estimaciones "mecánicas"que podemos proponer conociendo sólo de estadística, pero no es sencillo en las circunstancias actuales. Por un lado, no es fácil encontrar tablas con líneas de emisión para todos los elementos de interés, ni es obvio cómo transladar las cantidades relativas (los porcentajes masa/masa) a una escala absoluta (la intensidad de emisión) sin conocer la masa total de cada muestra. Además, el espectro está reducido a un región muy pequeña de frecuencias, y la precisión del instrumento parece más bien baja, ya que los picos alrededor de la linea teórica son bastante "gruesos" en general.

En su lugar, podemos plantear una línea de investigación menos ambiciosa: distinguir empíricamente las frecuencias donde se ven "picos" para cada elemento, buscar los estimadores forward y backward de entre dicho conjunto de frecuencias, y compararlos con los obtenidos de entre las frecuencias "múltiplo de 10" sugeridas por el enunciado. En el Apéndice incluimos gráficos similares al anterior para cada elemento, y de su análisis concluimos que existen 11 frecuencias "pico" en el espectro analizado: $\{100, 121, 145, 167, 195, 227, 252, 318, 351, 355, 386\}$ (el primer pico está por debajo de los 100PHz, pero esa es la primer frecuencia disponible; el pico de 351 está oculto debajo del de 355 en general, pero se ve claro para K_2O y MgO).

Aprovechando que el paquete `leaps` nos permite calcular no sólo los estimadores "forward" y "backward", sino también los "exhaustivos" (para 50 o menos covariables), y el "stepwise", buscaremos los mejores modelos por método, conjunto de frecuencias y cantidad de covariables, para las siguientes combinaciones:

Método	Frecuencias
mod10	exhaustive
picos	exhaustive
todas	forward
todas	backward
todas	seqrep

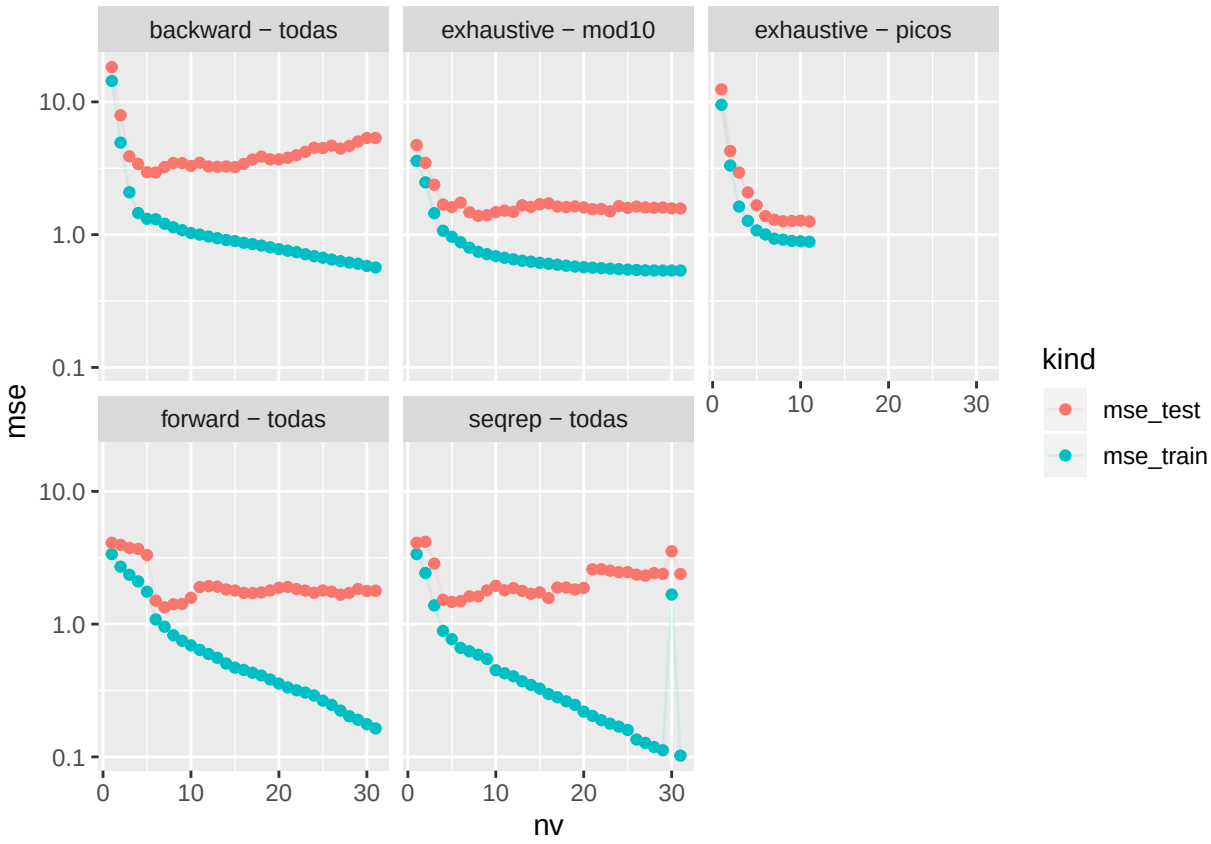
Como los conjuntos "Módulo 10" ($p = 31$) y "Picos" ($p = 11$) son pequeños, buscaremos directamente los mejores estimadores con el método exhaustivo. Para "Todas" ($p = 301$), buscaremos estimadores según cada uno de los métdos disponibles, hasta un total de 31 predictores.

La selección del "mejor modelo" constará de dos pasos. Dividiremos los datos disponibles en un conjunto de entrenamiento (*train*, n=120) y uno de prueba (*test*, n=60). El conjunto de entrenamiento se usará para determinar los modelos candidatos (uno por grupo de frecuencias, método y cantidad de variables), y de entre ellos elegiremos al que minimice el ECM en el conjunto de prueba. Contamos con un solo conjunto de prueba y por ende una única estimación puntual del ECM, así que no podemos estimar su varianza, y los criterios del estilo R1DE quedan descartados.

De los 135 modelos ajustados, a continuación presentamos los resultados para los 5 mejores:

Grupo	# Vars.	ECM Test	ECM Train	Formula
seqrep - todas	8	0.964	0.618	Na2O ~ 102 + 107 + 135 + 175 + 232 + 241 + 270 + 342
forward - todas	9	1.036	0.677	Na2O ~ 100 + 107 + 135 + 139 + 173 + 232 + 241 + 270 + 341
forward - todas	10	1.039	0.621	Na2O ~ 100 + 107 + 135 + 139 + 173 + 232 + 241 + 270 + 341 + 344
forward - todas	18	1.071	0.396	Na2O ~ 100 + 105 + 107 + 135 + 139 + 173 + 201 + 206 + 232 + 241 + 270 + 286 + 288 + 305 + 341 + 344 + 371 + 377
exhaustive - picos	7	1.105	1.068	Na2O ~ 100 + 145 + 167 + 227 + 252 + 318 + 351

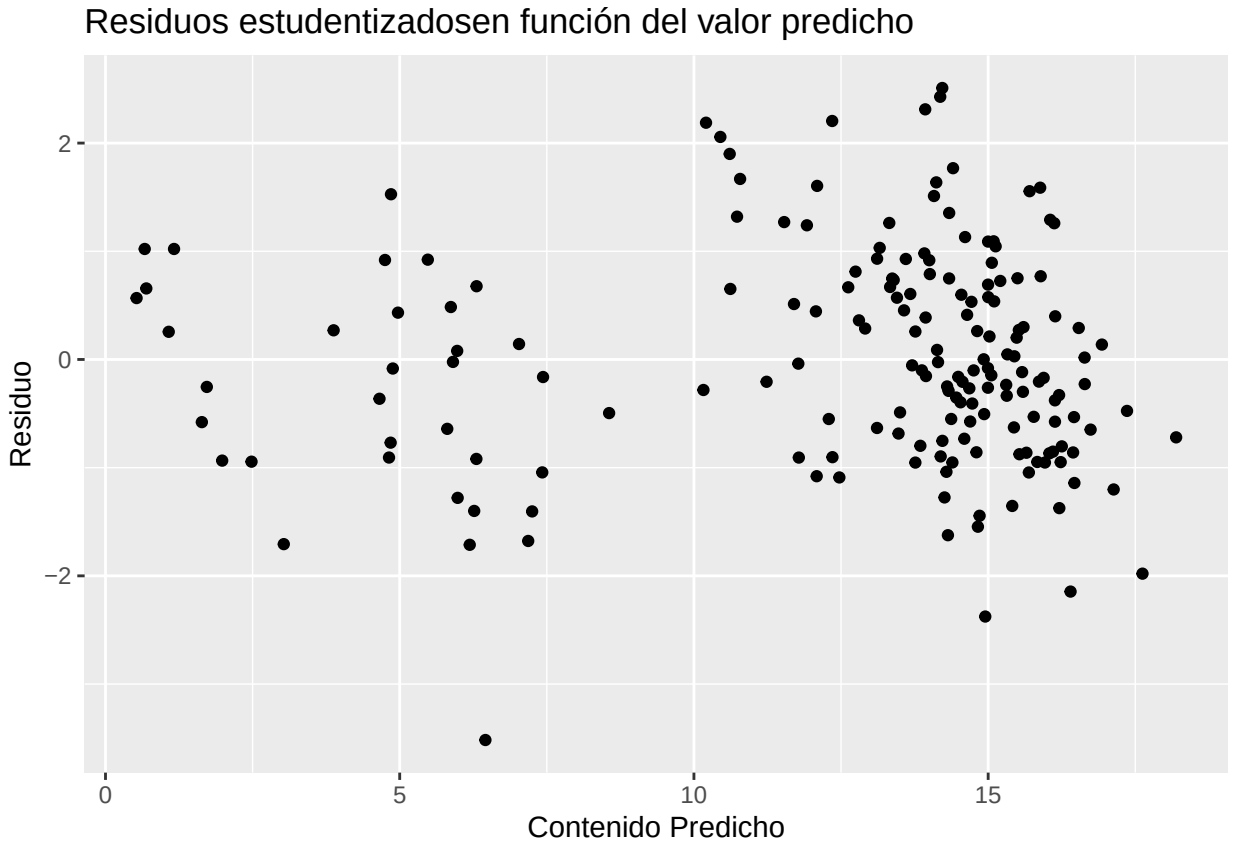
Gráficamente, es interesante observar la evolución del ECM de entrenamiento y prueba para cada conjunto de frecuencia y método empleado. Aunque el ECM de entrenamiento siempre disminuye a medida que se agregan covariables, su equivalente de prueba tiende a empeorar a partir de cierto punto en que comienza el sobreajuste. Otro punto interesante, es que aunque el "mejor" modelo proviene de otro conjunto de frecuencias, la performance del grupo de "picos" es muy buena aún con pocas covariables, y no se observa un sobreajuste a medida que agregamos predictores: evidentemente, se trate de frecuencias informativas.

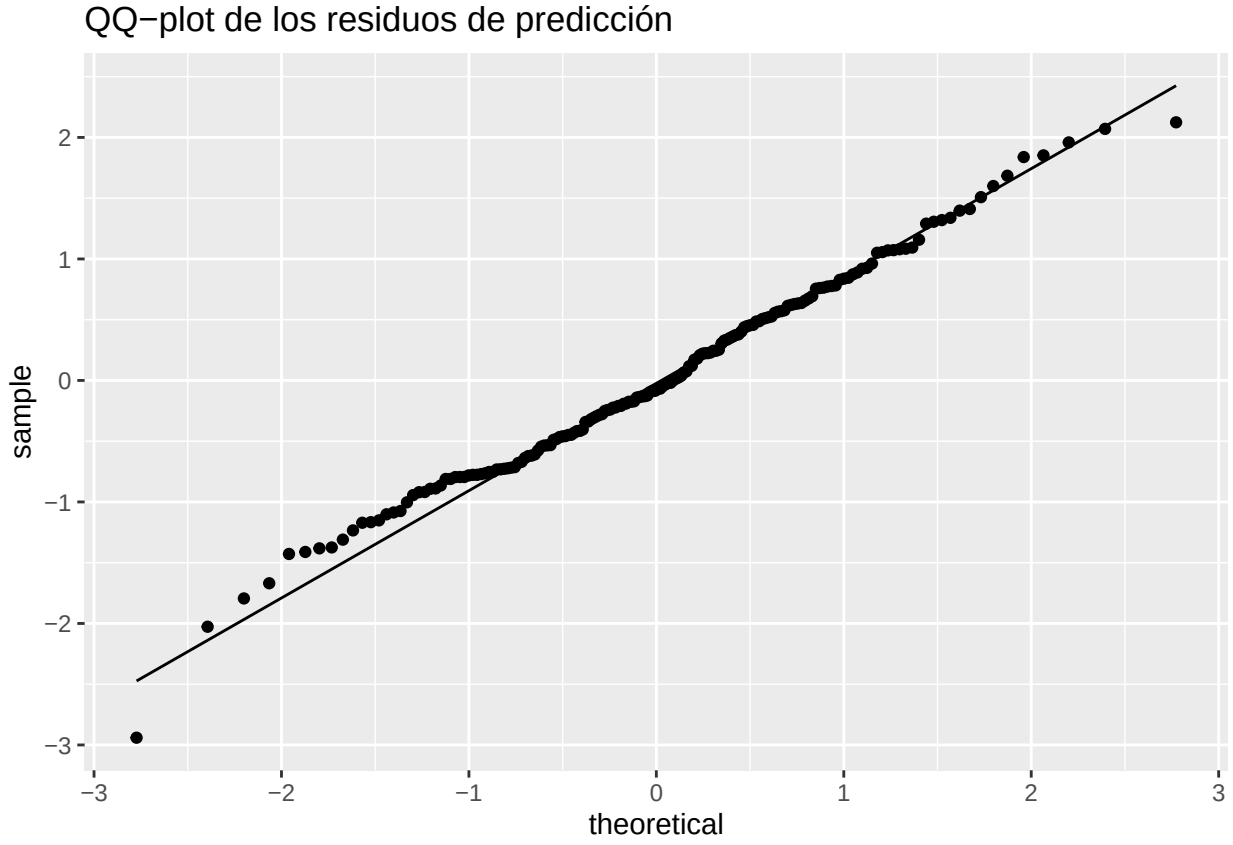


A continuación, presentamos los principales resultados del modelo elegido, junto con los gráficos de diagnóstico típicos. Vale recordar que los p-valores enunciados son estimativos, y debido a la selección de modelos no constituyen verdaderas probabilidades.

R^2 aj.	F obs.	P-valor
0.96	545	2.16e-117

Coef.	Estimado	P-valor
(Intercept)	23.2	2.03e-52
102	0.027	5.94e-06
107	0.0686	7.06e-06
135	-0.0656	2.81e-26
175	-0.00721	1.35e-36
232	-0.0269	7.88e-11
241	0.0506	1.22e-06
270	0.0297	0.00328
342	-0.0212	1.63e-30





Mas allá de un potencial outlier en el gráfico de residuos versus predichos, el modelo parece ajustar más que

adecuadamente. Antes de concluir, una palabra de cautela: en casos como este, en que hay más predictores que variables, las técnicas de validación cruzada no garantizan elegir el "mejor" modelo, ya que lo que conseguimos es más bien el modelo mejor ajustado al conjunto de prueba. Para ilustrarlo, a continuación presentamos los "mejores modelos" obtenidos en 10 repeticiones idénticas del mismo proceso, salvo por la elección aleatoria de los conjuntos de prueba y entrenamiento. El lector inquieto puede comprobar personalmente que los gráficos de diagnóstico son sumamente buenos para todos ellos:

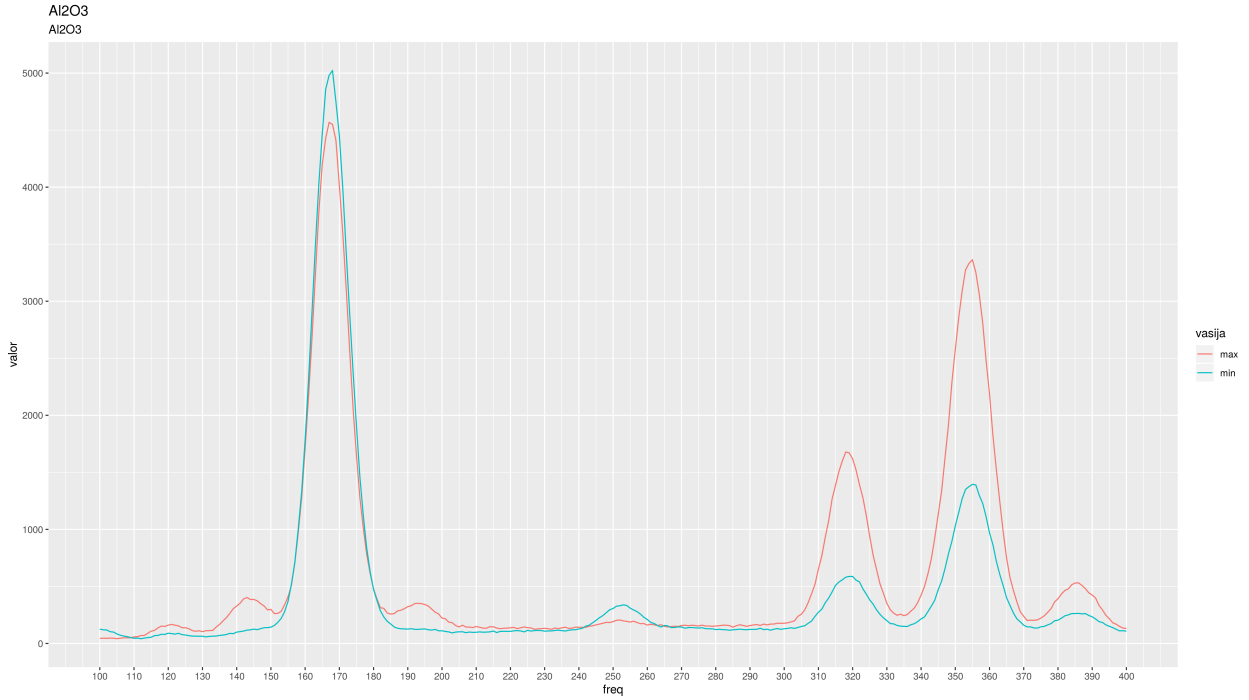
Grupo	# Vars.	ECM Test	ECM Train	Formula
exhaustive - picos	11	1.255	0.885	Na2O ~ 100 + 121 + 145 + 167 + 195 + 227 + 252 + 318 + 351 + 355 + 386
exhaustive - mod10	27	0.841	0.701	Na2O ~ 100 + 110 + 130 + 140 + 170 + 180 + 190 + 200 + 210 + 220 + 230 + ...
forward - todas	31	1.136	0.171	Na2O ~ 100 + 105 + 107 + 111 + 112 + 115 + 135 + 158 + 164 + 167 + 175 + ...
exhaustive - mod10	30	0.660	0.753	Na2O ~ 100 + 110 + 120 + 130 + 140 + 150 + 160 + 170 + 180 + 190 + 200 + ...
forward - todas	7	0.909	0.885	Na2O ~ 100 + 135 + 158 + 169 + 232 + 244 + 342
seqrep - todas	4	0.881	1.164	Na2O ~ 105 + 135 + 174 + 342
exhaustive - picos	6	1.070	1.133	Na2O ~ 100 + 145 + 167 + 227 + 318 + 351
seqrep - todas	11	0.934	0.473	Na2O ~ 105 + 107 + 146 + 175 + 225 + 241 + 297 + 305 + 352 + 375 + 380
seqrep - todas	6	0.988	0.805	Na2O ~ 100 + 135 + 169 + 232 + 244 + 342
seqrep - todas	8	0.964	0.618	Na2O ~ 102 + 107 + 135 + 175 + 232 + 241 + 270 + 342

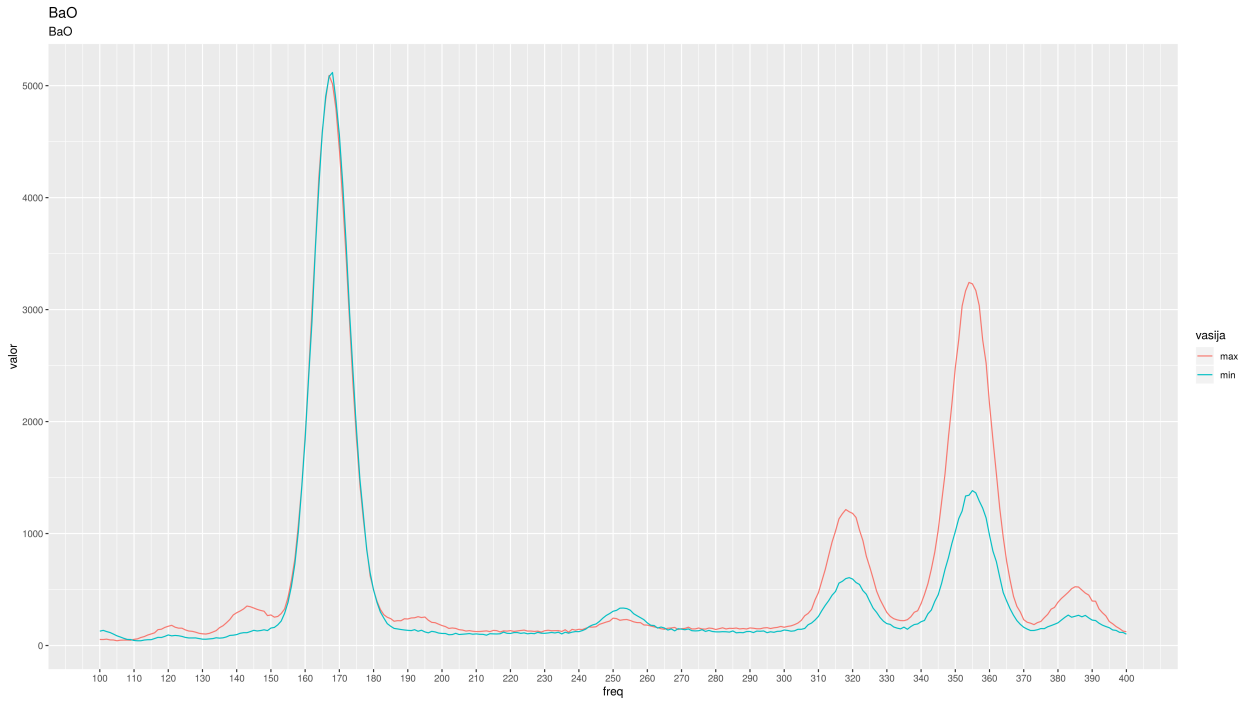
Honestamente, no resulta intuitivo cómo avanzar en situaciones como ésta, así que habiendo obtenido un ajuste más que razonable, concluimos aquí la exploiración de los datos.

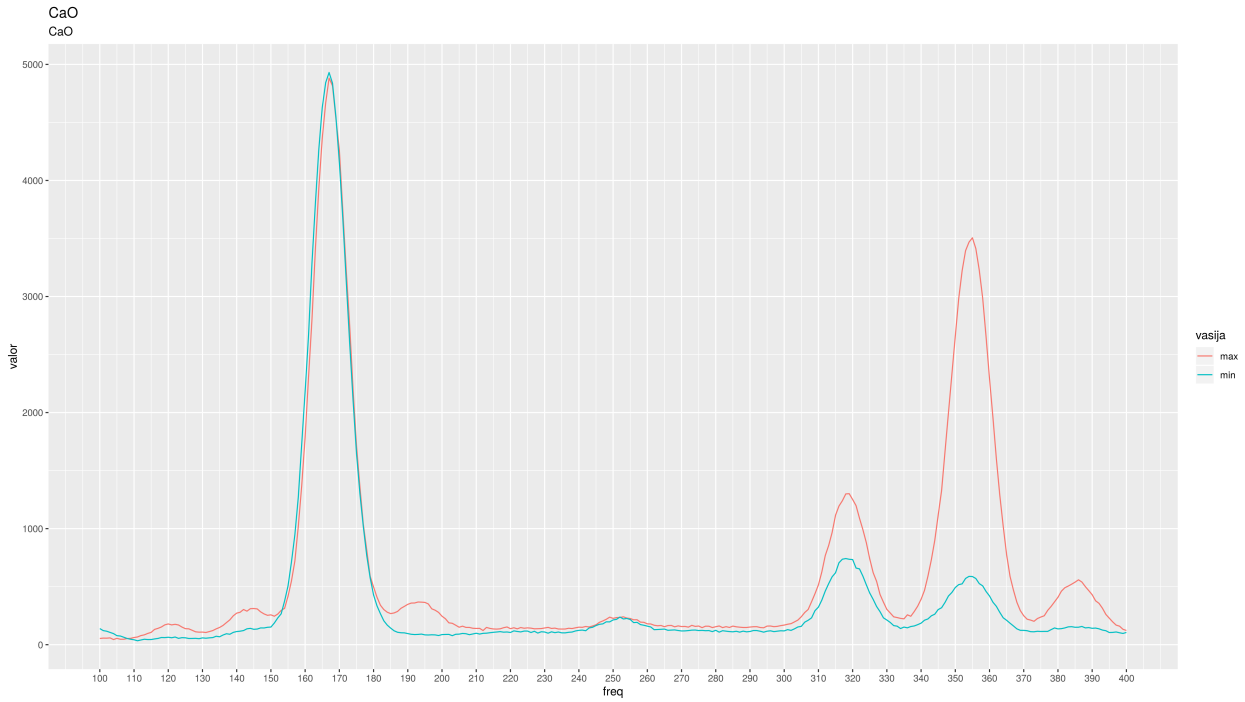
Referencias

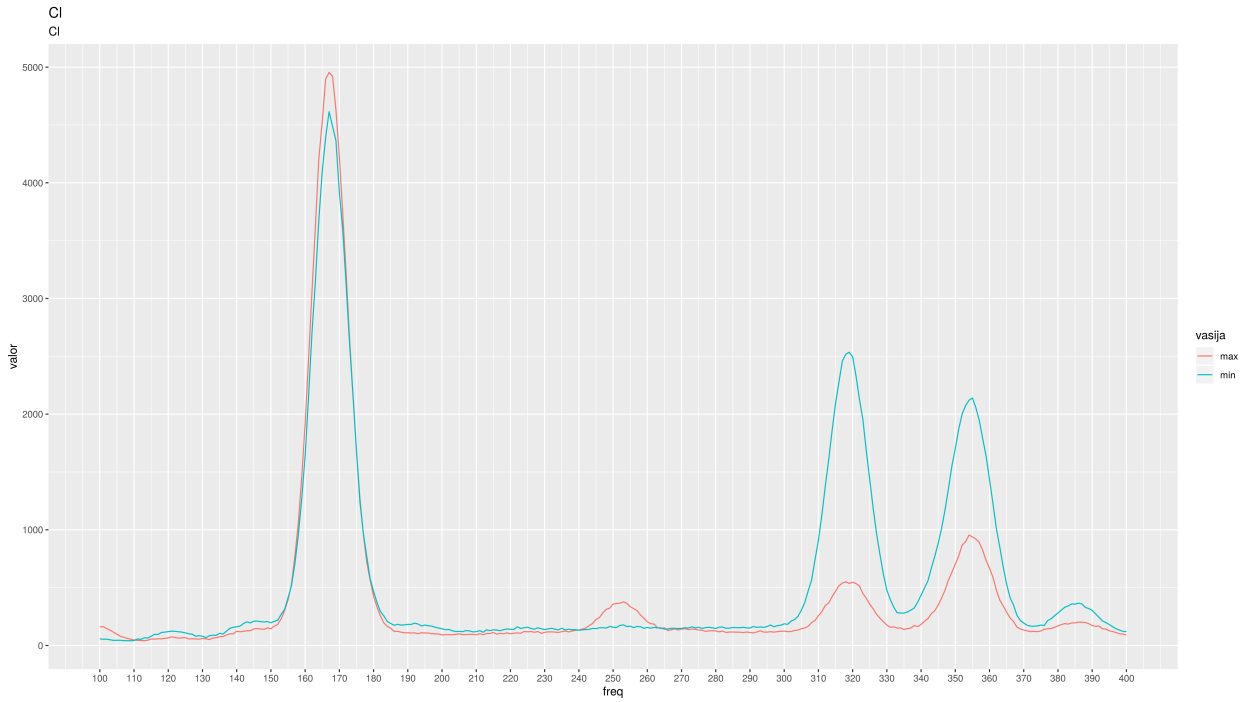
[1] Líneas de análisis para fluorescencia de rayos X

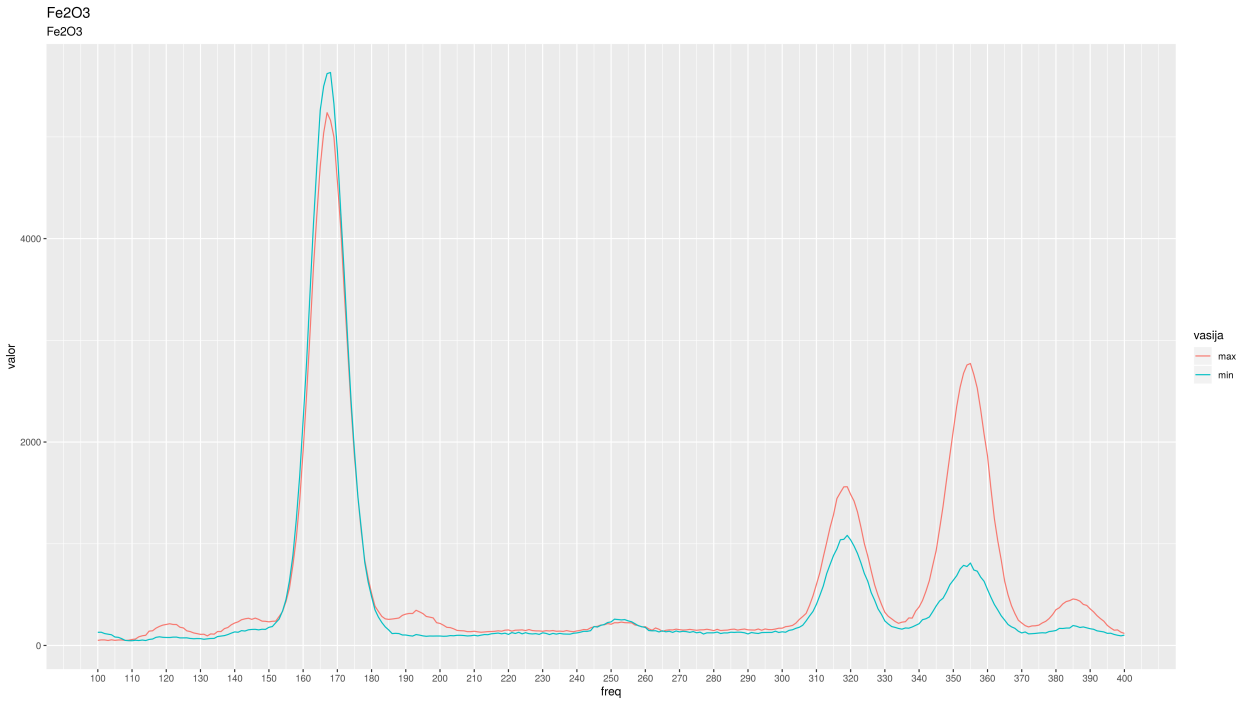
Apéndice: Comparación de espectros para vasijas con mínimas y máximas cantidades de cada elemento.

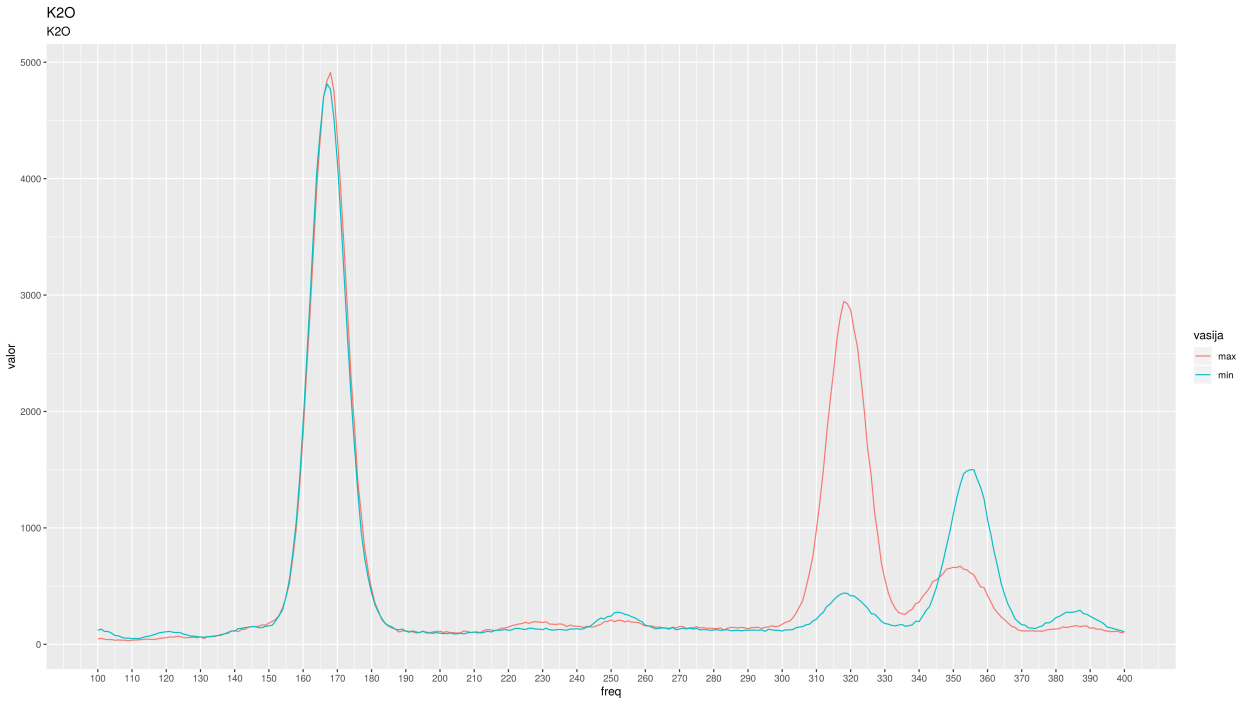


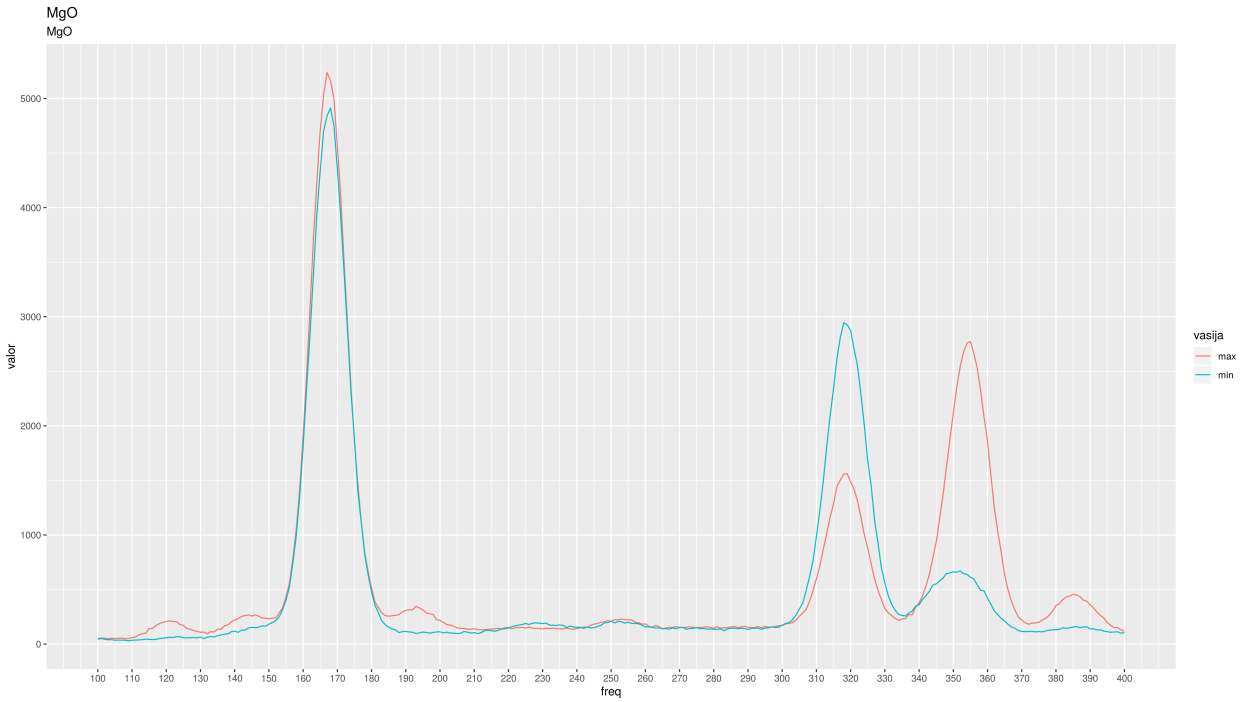


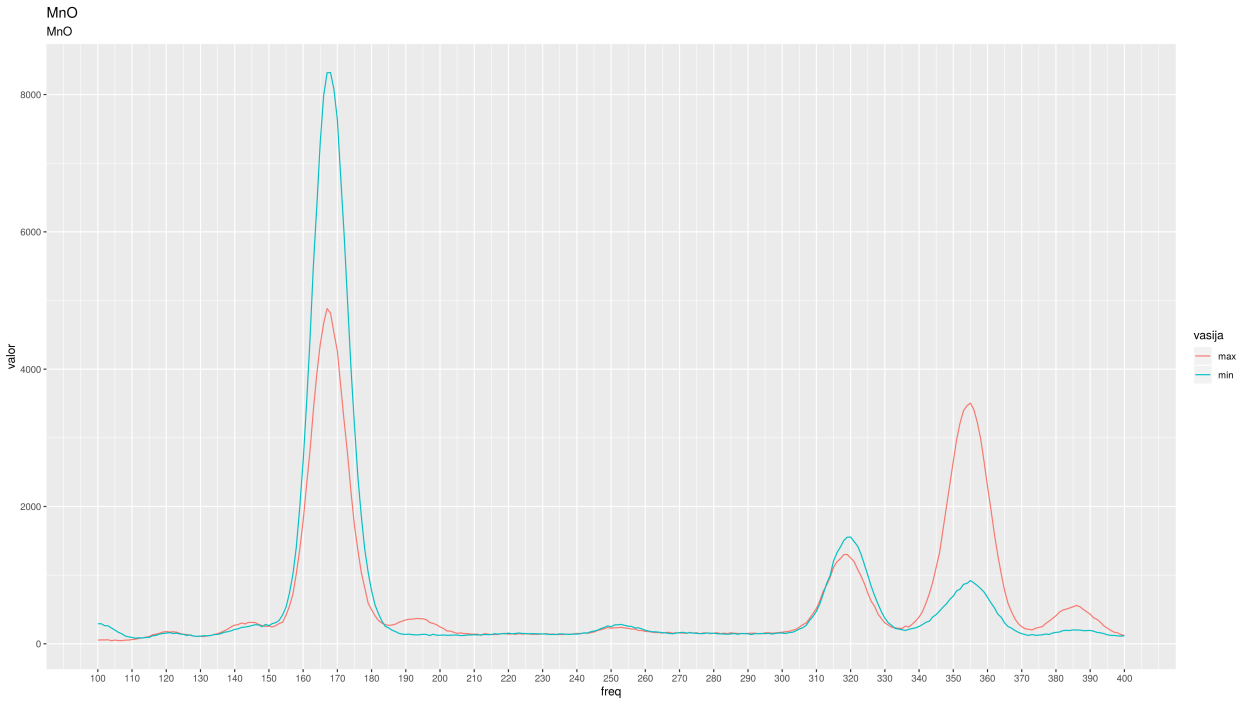

CaO
CaO
valor
freq
vasija
max
min

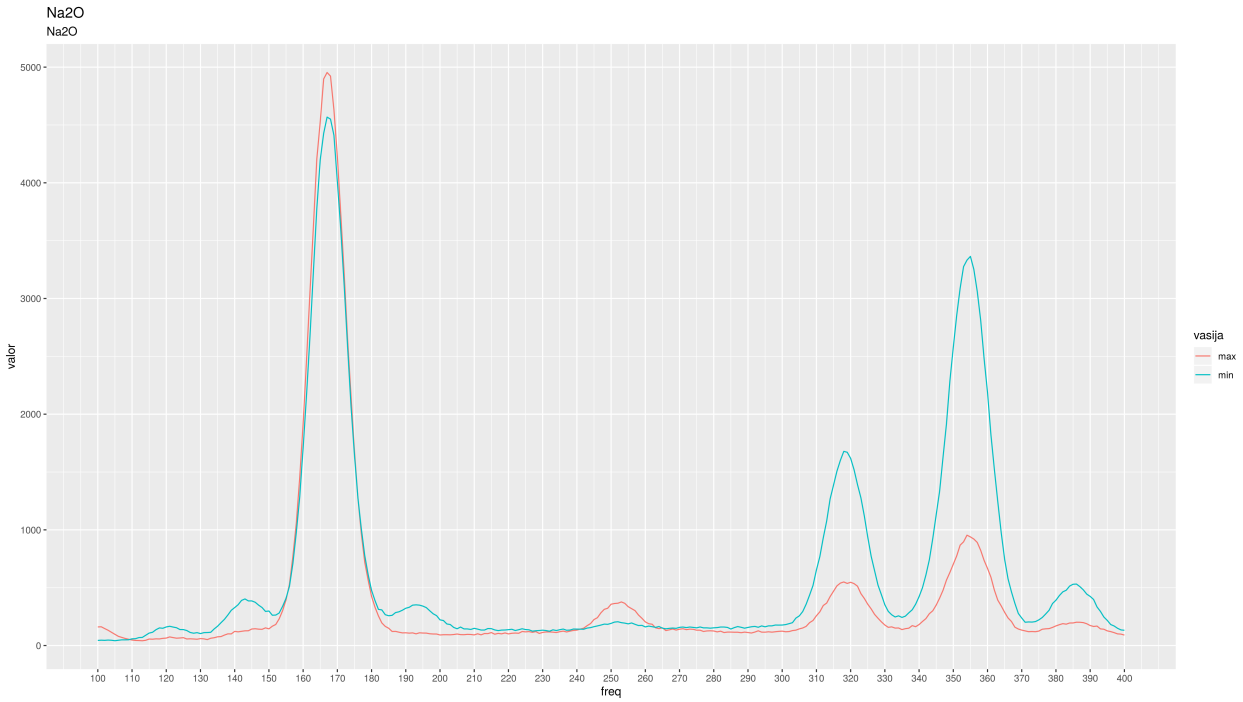

Cl
Cl
valor
freq
vasija
max
min

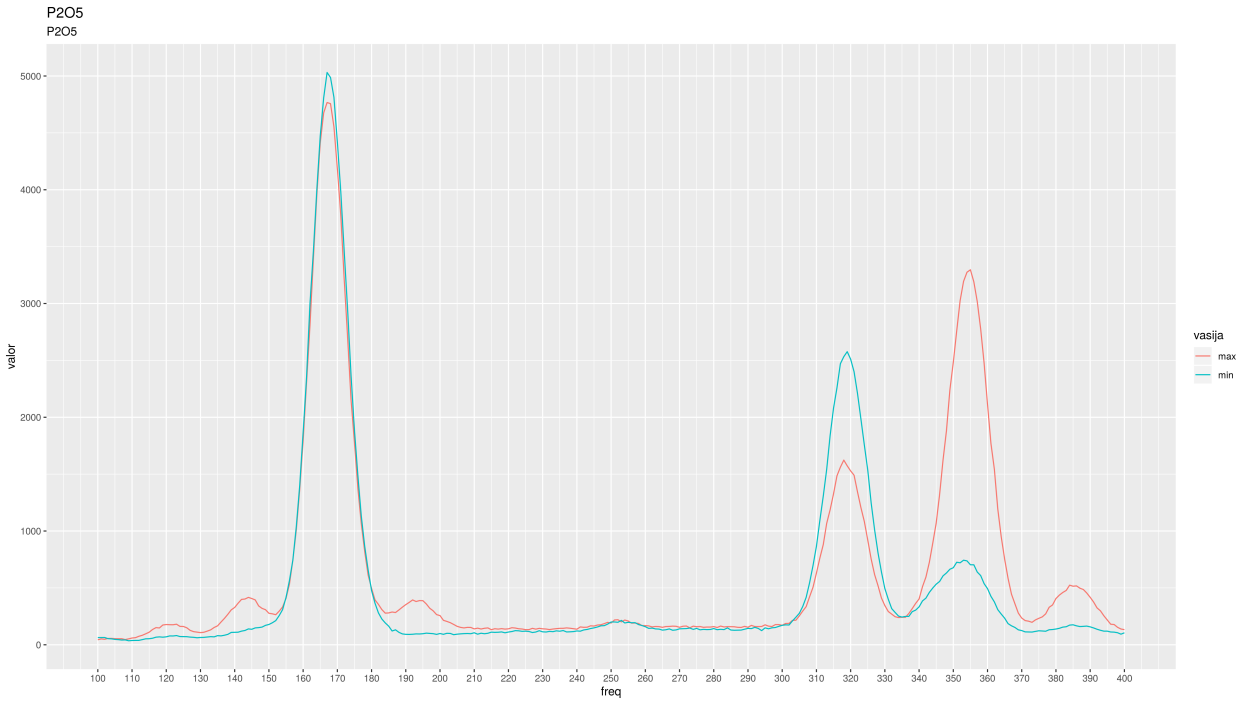

Fe2O3
Fe2O3
valor
freq
vasija
max
min

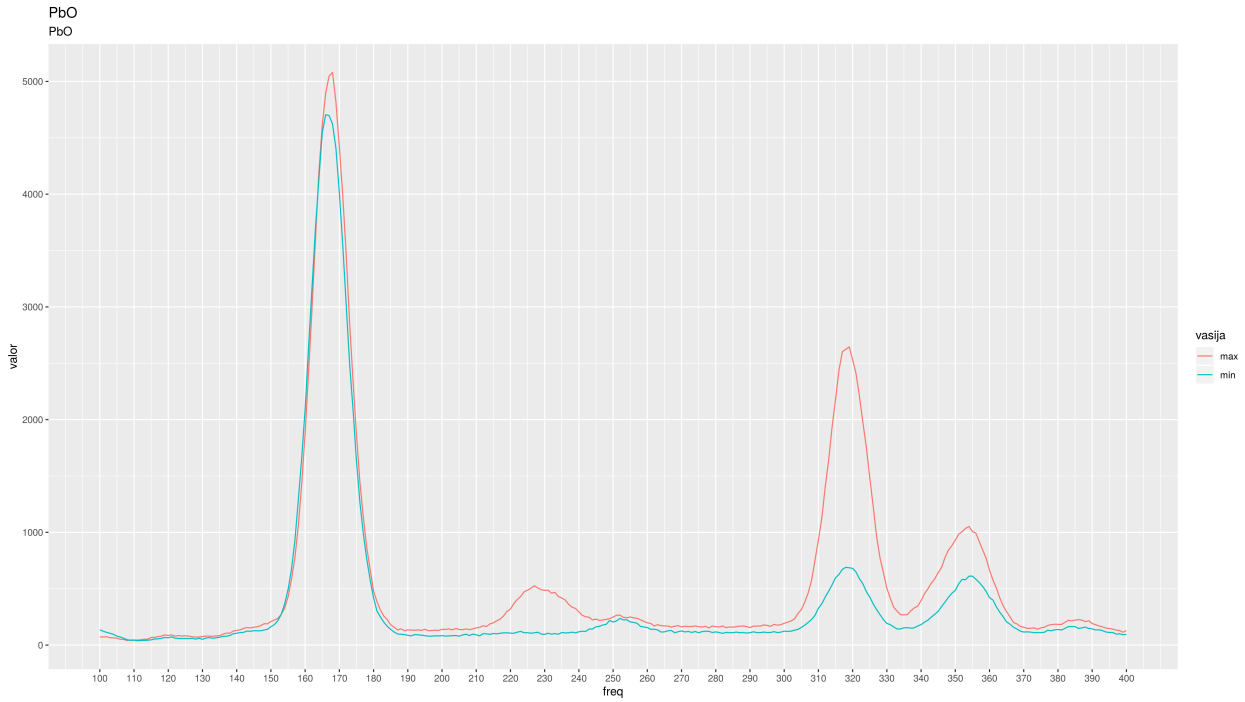

K2O
K2O
valor
freq
vasija
max
min

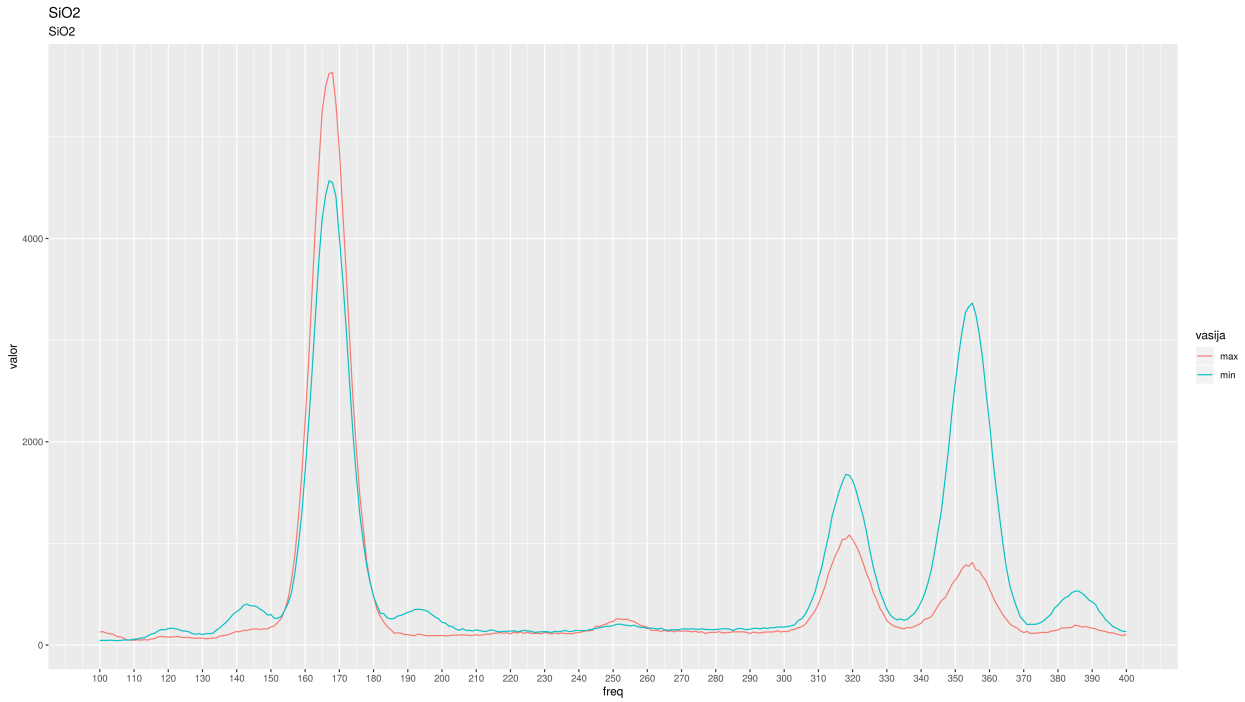

MgO
MgO
valor
freq
vasija
max
min

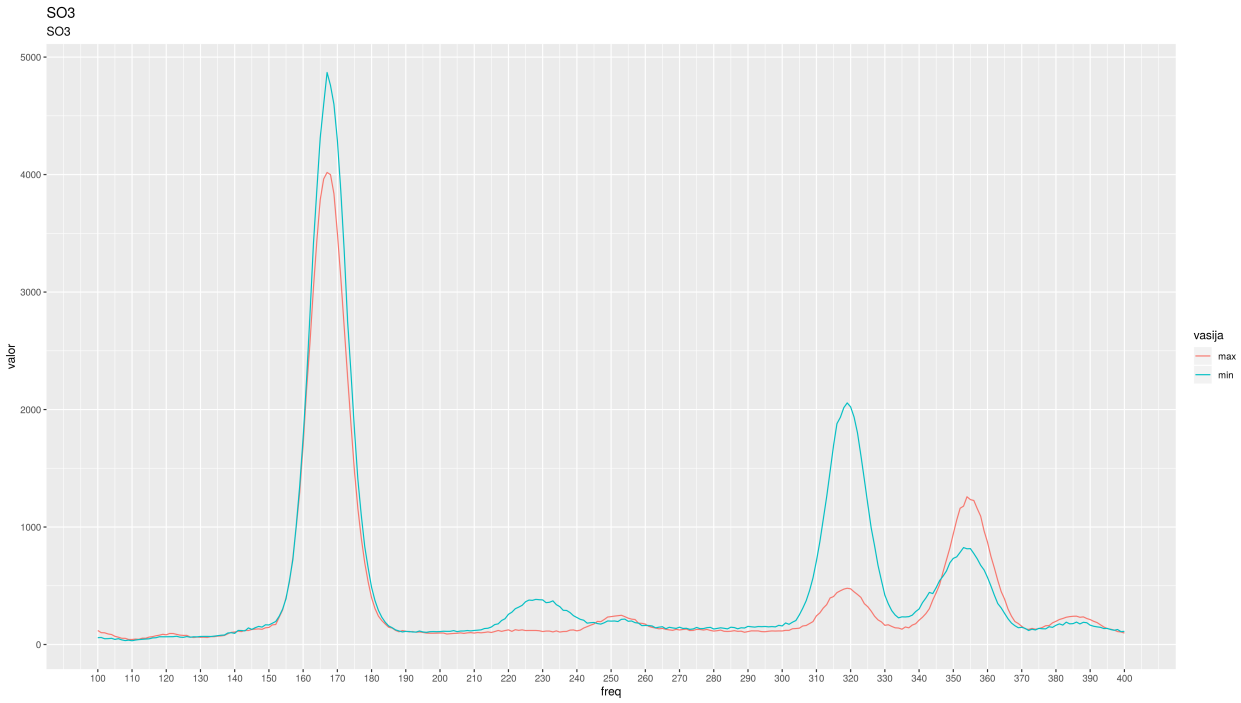

MnO
MnO
valor
freq
vasija
max
min


Na2O
Na2O
valor
freq
vasija
max
min


P2O5
P2O5
valor
freq
vasija
max
min


PbO
PbO
valor
freq
vasija
max
min


SiO2
SiO2
valor
freq
vasija
max
min


SO3
SO3
valor
freq
vasija
max
min
5000
4000
3000
2000
1000
0
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400