



Universidad de Valladolid

Facultad de Ciencias

Grado en Estadística

**ANÁLISIS DE PARÁMETROS
AMBIENTALES Y CALIDAD DEL AIRE
EN HOGARES EUROPEOS**

Autor: David Moretón Pavón

Tutor: Jesús Alberto Tapia García

Curso: 2024 - 2025

Resumen

En nuestras vidas, la salud y el bienestar son esenciales y están estrechamente vinculados a nuestro hogar. Aquí, la calidad del aire es crucial, ya que pasamos la mayor parte de nuestro tiempo en interiores. Este TFG tiene como objetivo examinar la calidad del aire interior en casas de Austria, Polonia y España, empleando métodos estadísticos para analizar varios factores ambientales y sus relaciones.

Para esto, se realiza un estudio de algunos parámetros, tanto en las relaciones entre sí como su comportamiento a elementos como la calefacción, fumar en la casa y la estación del año. También se utilizan técnicas como el Análisis de Componentes Principales (ACP), redes bayesianas y pruebas no paramétricas para identificar patrones y relaciones entre grupos de variables. El objetivo es obtener una visión más clara de la calidad del aire interior al ofrecer un punto de vista estadístico con implicaciones importantes para la salud ambiental.

Palabras clave

Calidad del aire, análisis de componentes principales, redes bayesianas, patrones, relaciones.

Abstract

In our lives, health and well-being are essential and closely linked to our homes. Here, air quality is crucial, as we spend most of our time indoors. This project aims to examine indoor air quality in homes in Austria, Poland, and Spain, using statistical methods to analyse various environmental factors and their relationships.

To achieve this, a study is conducted on several parameters, both in terms of their interrelationships and their behaviour in relation to elements such as heating, smoking indoors, and the time of year. Techniques such as Principal Component Analysis (PCA), bayesian networks and non-parametric tests are also used to identify patterns and relationships between groups of variables. The goal is to gain a clearer understanding of indoor air quality by providing a statistical perspective with important implications for environmental health.

Keywords

Air quality, Principal Component Analysis, Bayesian networks, patterns, relationships.

Índice general

Resumen	I
Abstract	I
Índice de figuras	IV
Índice de tablas	VI
1. Introducción	1
1.1. Contexto del problema	1
1.2. Objetivos	1
1.3. Entorno de trabajo	2
1.4. Estructura	2
2. Metodología	3
2.1. Estadística descriptiva	3
2.2. Pruebas de normalidad	4
2.3. Análisis de Componentes Principales	4
2.4. Redes Bayesianas	5
2.5. Kruskal-Wallis y análisis post-hoc	6
3. Datos	7
3.1. Características de los datos	7
3.2. Importar y preprocesar los datos	9
3.2.1. Recodificación de variables	9
3.2.2. Identificación de hogares fumadores	9
3.2.3. Unificación y promedio horario	9
3.2.4. Clasificación de la calidad del aire	10

4. Análisis de los datos y resultados	11
4.1. Estadística descriptiva	11
4.2. Normalidad en los datos	22
4.2.1. CO ₂	22
4.2.2. Humedad	23
4.2.3. Partículas	24
4.2.4. Temperatura	25
4.2.5. TVOCs	26
4.2.6. Formaldehído	27
4.3. Análisis de Componentes Principales	28
4.3.1. Datos originales	28
4.3.2. Datos transformados	32
4.4. Redes Bayesianas	37
4.5. Pruebas de Kruskal-Wallis y de Dunn	40
4.5.1. Kruskal-Wallis	40
4.5.2. Dunn	41
5. Conclusiones	44
5.1 Trabajos futuros	44
6. Bibliografía	46
Anexo	48
A.1. Importar y preprocesar los datos	48
A.2. Estadística descriptiva	53
A.3. Pruebas de normalidad	58
A.4. Análisis de Componentes Principales	61
A.5 Redes Bayesianas	63
A.6. Pruebas de Kruskal-Wallis y de Dunn	65

Índice de figuras

3.1. Histograma sobre la cantidad de muestras a lo largo del tiempo	8
4.1. Distribuciones del CO ₂	13
4.2. Distribuciones de la humedad	14
4.3. Distribuciones de PM _{2.5}	15
4.4. Distribuciones de la temperatura	16
4.5. Distribuciones de los volátiles	17
4.6. Distribuciones del formaldehído	18
4.7. Distribución de las observaciones por las interacciones de las variables categóricas en Austria	19
4.8. Distribución de las observaciones por las interacciones de las variables categóricas en España	20
4.9. Distribución de las observaciones por las interacciones de las variables categóricas en Polonia	21
4.10. Log-verosimilitud perfilada de CO ₂ en Austria	22
4.11. Q-Q plot de CO ₂ en Austria	23
4.12. Q-Q plot de CO ₂ en España	23
4.13. Q-Q plot de CO ₂ en Polonia	23
4.14. Q-Q plot de humedad en Austria	23
4.15. Q-Q plot de humedad en España	23
4.16. Q-Q plot de humedad en Polonia	24
4.17. Log-verosimilitud perfilada de PM _{2.5} en España	24
4.18. Q-Q plot de PM _{2.5} en Austria	25
4.19. Q-Q plot de PM _{2.5} en España	25
4.20. Q-Q plot de PM _{2.5} en Polonia	25
4.21. Q-Q plot de temperatura en Austria	25
4.22. Q-Q plot de temperatura en España	25
4.23. Q-Q plot de temperatura en Polonia	26
4.24. Log-verosimilitud perfilada de TVOCs en España	26
4.25. Q-Q plot de TVOCs en Austria	27

4.26. Q-Q plot de TVOCs en España	27
4.27. Q-Q plot de TVOCs en Polonia	27
4.28. Q-Q plot de formaldehído en España	27
4.29. Scree plot de Austria	28
4.30. Biplot de Austria	29
4.31. Scree plot de España	30
4.32. Biplot de España	30
4.33. Scree plot de Polonia	31
4.34. Biplot de Polonia	32
4.35. Scree plot de Austria con datos transformados	33
4.36. Biplots de Austria con datos transformados	33
4.37. Scree plot de España con datos transformados	34
4.38. Biplots de España con datos transformados	35
4.39. Scree plot de Polonia con datos transformados	35
4.40. Biplots de Polonia con datos transformados	36
4.41. Redes bayesianas para Austria	37
4.42. Redes bayesianas para España	38
4.43. Redes bayesianas para Polonia	39
4.44. Redes bayesianas para el total	39

Índice de tablas

3.1. Resumen de los conjuntos de datos	8
3.2. Desviaciones típicas originales frente al promedio	9
3.3: Umbrales clasificadores del AQI	10
4.1. Matriz de correlaciones entre variables observadas en Austria	12
4.2. Matriz de correlaciones entre variables observadas en Polonia	12
4.3. Matriz de correlaciones entre variables observadas en España	12
4.4. Resumen del CO ₂	13
4.5. Resumen de la humedad	14
4.6. Resumen de PM _{2.5}	15
4.7. Resumen de la temperatura	16
4.8. Resumen de los volátiles	17
4.9. Resumen del formaldehído	18
4.10. Resumen normalidad de CO ₂	22
4.11. Resumen normalidad de humedad	23
4.12. Resumen normalidad de partículas	24
4.13. Resumen normalidad de temperatura	25
4.14. Resumen normalidad de volátiles	26
4.15. Resumen normalidad de formaldehído	27
4.16. Importancia de las componentes principales en Austria	28
4.17. Cargas de las variables originales en Austria	29
4.18. Importancia de las componentes principales en España	29
4.19. Cargas de las variables originales en España	30
4.20. Importancia de las componentes principales en Polonia	31
4.21. Cargas de las variables originales en Polonia	32
4.22. Importancia de las componentes principales en Austria con datos transformados	32
4.23. Cargas de las variables transformadas en Austria	33
4.24. Importancia de las componentes principales en España con datos transformados	34
4.25. Cargas de las variables transformadas en España	34
4.26. Importancia de las componentes principales en Polonia con datos transformados	35

4.27. Cargas de las variables transformadas en Polonia	36
4.28. Resumen de las pruebas de Kruskal-Wallis	41
4.29. Resumen del test de Dunn sobre season	41
4.30. Resumen del test de Dunn sobre HOM_smoking	42
4.31. Resumen del test de Dunn sobre AQI_co2	43
4.32. Resumen del test de Dunn sobre AQI_pm25	43

Capítulo 1

Introducción

1.1. Contexto del problema

Este estudio está emparejado con el proyecto K-HEALTHinAIR [1]. Es una iniciativa europea coordinada por el centro tecnológico CARTIF [2] y financiado por el programa Horizonte Europa, cuyo objetivo es estudiar la calidad del aire en interiores y su impacto a la salud. Se analizan varios entornos y fuentes de contaminación para proponer soluciones que mejoren esta calidad del aire. El ámbito doméstico es el escogido para el análisis.

Gran parte de la población mundial permanece en interiores entre un 80% y 90% de su vida demostrando la importancia del aire que respiramos. A pesar de ello, se suele descuidar el entorno y permitir acumulación de contaminantes disminuyendo el rendimiento cognitivo y la salud física de las personas [3].

3.2 millones de personas al año fallecen debido a la contaminación provocada mediante la calefacción o al cocinar con combustibles sólidos y queroseno según la Organización Mundial de la Salud [4], estos, penetran en el sistema respiratorio y debilitan tanto el sistema inmunológico como la capacidad de oxigenar la sangre.

Se tratarán varias muestras de datos para comprobar la interacción de contaminantes y agentes meteorológicos, entre sí para los tres países miembros del estudio (Austria, España y Polonia), frente a tipos de calefacción para Austria y Polonia y frente a hogares con presencia de fumadores en Austria y España.

1.2. Objetivos

Los objetivos clave de este trabajo se centran en los parámetros existentes en el aire:

- Identificar cómo se relacionan las variables asociadas a la calidad del aire usando la estadística descriptiva en forma de gráficos y correlaciones.
- Reducir dimensionalidad y encontrar patrones ocultos mediante el uso del Análisis de Correlaciones Principales.
- Analizar las relaciones de factores influyentes en la calidad del aire mediante redes bayesianas.
- Evaluar el impacto de los factores externos, analizar como el tipo de calefacción, el hábito de fumar y la estación del año afectan en el comportamiento de estos agentes a través de estudiar sus diferencias usando pruebas de Kruskal-Wallis y pruebas de Dunn.

1.3. Entorno de Trabajo

En la elaboración del código para el análisis y su posterior ejecución se ha utilizado:

R: Versión 4.4.1 (2024-06-14 ucrt). Es un lenguaje de programación y entorno software diseñado para el análisis estadístico, tratamiento de datos y elaboración de gráficos. Al ser de código abierto permite a gente externa la elaboración de paquetes generando gran variedad de funciones y también es fuerte en el manejo de grandes volúmenes de datos, lo que incentiva su uso académico y profesional [5].

RStudio: Versión 2024.04.2+764. Es un IDE, entorno de desarrollo integrado, para R. Ofrece una interfaz gráfica que facilita la escritura de código, visualización de datos y seguimiento de las ejecuciones, y aunque R se pueda usar de forma independiente, RStudio mejora la experiencia mediante scripts, consola, entorno de variables y espacio para gráficos [6].

1.4. Estructura

EL TFG está ordenado en varias secciones para facilitar la comprensión:

Capítulo 1. Introducción: Se presenta el trabajo contextualizando e informando sobre el problema a analizar, para poder asentar unas bases de conocimiento y discutir el objetivo. También se muestra el entorno de trabajo usado y la estructura de la memoria.

Capítulo 2. Metodología: En esta sección se muestran las diferentes técnicas utilizadas, junto con una descripción teórica de cada una. Estas son: análisis descriptivo, pruebas de normalidad, análisis de componentes principales, redes bayesianas y prueba de Kruskal-Wallis con análisis post-hoc.

Capítulo 3. Datos: Los datos son clasificados y descritos, explicando los distintos grupos utilizados, de donde provienen, las variables que contienen, una pequeña descripción de estas y la importación y preprocesado de los datos.

Capítulo 4. Análisis de los datos y resultados: Se interpretan en profundidad los resultados de las diferentes técnicas mencionadas tras aplicarlas a los datos. Por cada país y comparando entre ellos.

Capítulo 5. Conclusiones: Los objetivos del estudio son comentados frente a los resultados y las posibles utilidades de dicho estudio.

Capítulo 6. Bibliografía: Las referencias realizadas a lo largo del TFG según las normas APA.

Anexo: El código del tratamiento y análisis de los datos

Capítulo 2

Metodología

El objetivo de los distintos métodos aplicados es el de facilitar la comprensión del comportamiento, los patrones y las relaciones de las variables a estudiar. La fundamentación teórica abordará tanto los métodos como las inferencias asociadas a cada uno para justificar la elección y su contribución a los objetivos del trabajo.

2.1. Estadística descriptiva

Es importante comprobar las correlaciones entre variables, ya que permite detectar si dos variables están relacionadas y en qué medida. Cuando una variable aumenta y la segunda también tiende a aumentar, se habla de correlación positiva; si en cambio tiende a descender, se considera correlación negativa. Si no existe relación aparente es correlación nula. Analizar estas correlaciones también ayuda a tener una visión más clara de patrones y tendencias presentes en los datos. Para la visualización se utiliza la librería “corrplot” [10].

La correlación de Pearson [11], de las más utilizadas, estima la relación principalmente lineal entre dos variables en el intervalo $[-1, 1]$. Sin embargo, dado que la naturaleza de los datos puede incluir relaciones no lineales, se usará la correlación de Spearman [12]. Esta también produce valores en intervalo $[-1, 1]$ pero es más adecuada para detectar relaciones monótonas no lineales al usar rangos para calcular rho (ρ), siendo los rangos las posiciones relativas que ocupan los valores al ordenar el conjunto de datos, lo que la hace más robusta en otros escenarios y no necesitar de normalidad de los datos. El coeficiente se calcula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia de rangos ($R(x_i) - R(y_i)$) y n es el número de pares de datos.

Se usarán técnicas descriptivas para así obtener una visión general del comportamiento de los datos. La media, la mediana o la asimetría son ejemplos de medidas que se usarán como apoyo para los gráficos de densidades, estos representan la distribución de los valores de variable continua, utilizando una curva suave que estima la densidad de la probabilidad. Se han utilizado las librerías “ggplot2” [7], “patchwork” [8] y “moments” [9] en R para estos análisis. Para las variables categóricas se usan gráficos que resumen el número de observaciones por cada interacción entre ellas.

2.2. Pruebas de normalidad

Dado que algunas variables pueden presentar distribuciones asimétricas, se ha considerado el uso de la transformación Box-Cox con el objetivo de reducir dichas asimetrías y así aproximarse a una distribución normal [13]. Se usa la librería “car” [14] en R. Esta transformación genera una nueva variable dependiente del parámetro lambda (λ):

$$y(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \log(\lambda), & \text{si } \lambda = 0 \end{cases}$$

donde $x > 0$ y λ parámetro estimado mediante máxima log-verosimilitud.

Se usarán dos pruebas para detectar normalidad, la primera es el test de Shapiro-Wilk [15], ampliamente reconocida por su potencia en detectar desviaciones. Se usará la librería de R “stats” [16]. Su estadístico se define:

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

donde $x_{(i)}$ son los valores ordenados de mayor a menor y a_i los coeficientes calculados a partir de la media y la varianza según una teórica distribución normal.

La segunda prueba realizada es el test de Lilliefors [17]. Trata de una modificación de la prueba de Kolmogorov-Smirnov donde se supone que la media y la varianza de la población no son conocidas y se estiman a partir de la muestra. En R se usará la librería “nortest” [18]. El estadístico se basa en una diferencia y se define:

$$D = \sup_x |F_n(x) - \Phi_{\hat{\mu}, \hat{\sigma}}(x)|$$

donde $F_n(x)$ es la distribución empírica acumulada y $\Phi_{\hat{\mu}, \hat{\sigma}}(x)$ la distribución normal estimada.

Como complemento visual a estas dos pruebas, se imprimen los diagramas cuantil-cuantil (Q-Q plots), los cuales representan los cuantiles teóricos de la distribución normal en el eje horizontal y los cuantiles observados en el vertical. Si los puntos se alinean sobre la diagonal, se considera que los datos se asemejan a una distribución normal.

2.3. Análisis de Componentes Principales

También denominada ACP o PCA en inglés, es una técnica estadística multivariante en la que se busca estudiar relaciones entre variables e individuos, en la que no existe variable respuesta, propuesto por primera vez por Pearson [19]. Se buscan combinaciones lineales de las variables originales que maximicen la información del conjunto de datos creando nuevas variables dando la posibilidad de disminuir la dimensión del conjunto de datos [20].

Las nuevas variables son llamadas componentes principales y son ortogonales entre sí. La construcción es llevada a cabo tal que la primera componente captura la mayor proporción de varianza de los datos, y el resto son construidas siendo ortogonales a las anteriores y que expliquen la mayor cantidad de variabilidad. Los autovalores provenientes de la matriz de correlaciones tras escalar los datos están asociados a cada una de las componentes indicando la cantidad de variabilidad explicada por cada una. Para llevar a cabo esta estrategia se utilizan las funciones otorgadas por la librería “stats” [16].

La elección de dimensionalidad final puede guiarse por varios criterios [20], los que se tendrán en cuenta son:

- Regla de Kaiser: Las componentes elegidas son las que su autovalor es mayor que 1, ya que explican más varianza que una variable original promedio.
- Varianza acumulada: Se selecciona el número mínimo de componentes cuya suma de varianzas supere un porcentaje umbral.
- Gráfico de sedimentación o “scree plot”: Se representa el valor de los autovalores en orden descendente en un gráfico y se busca un el punto de inflexión a partir del que las ganancias son insignificantes.

2.4. Redes Bayesianas

Son modelos probabilísticos basados en grafos dirigidos acíclicos que representan relaciones y dependencias entre variables, generalmente categóricas. Los nodos del grafo representan las variables y los arcos las relaciones de dependencia. Dependiendo de la dirección de los arcos de un nodo existen tres tipos: nodos en secuencia ($X \rightarrow Y \rightarrow Z$), nodos divergentes ($X \leftarrow Y \rightarrow Z$) y nodos convergentes ($X \rightarrow Y \leftarrow Z$) [21]. Se usará la librería “bnlearn” de R [22].

Se utilizan métodos de inferencia aproximada, en el estudio se usará el muestreo Monte Carlo que permite estimar probabilidades por simulaciones que permite calcular probabilidades sin resolver la red de forma exacta. Una vez definida la estructura, se ajustan los parámetros, calculando las probabilidades condicionales de las variables dado su conjunto de padres. Es llamado el aprendizaje de parámetros y el objetivo es estimar la distribución conjunta:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pa}(X_i))$$

Siendo un conjunto de variables aleatorias $X = \{X_1, X_2, \dots, X_n\}$ donde $\text{Pa}(X_i)$ es el conjunto de variables padre de X_i .

Las probabilidades condicionales se estiman principalmente usando dos métodos: máxima verosimilitud (MLE) mediante las frecuencias relativas y la estimación bayesiana que incorpora conocimiento previo (a priori).

2.5. Kruskal-Wallis y análisis post-hoc

El test de Kruskal-Wallis es una prueba que permite comparar si existen diferencias significativas entre grupos independientes de forma no paramétrica, lo que permite a la variable ordinal no cumplir el requisito de normalidad [23]. Es un test basado en rangos lo que lo hace robusto ante asimetrías y presencia de outliers. Las hipótesis correspondientes son:

- H_0 : Misma distribución entre grupos (no hay diferencias).
- H_1 : Al menos un grupo es diferente de otro.

Dadas N observaciones divididas en k grupos con tamaños n_i y rangos R_i , el estadístico, el cual sigue una distribución chi-cuadrado con $k-1$ grados de libertad, se calcula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

donde $\bar{R}_i$ es la media de los rangos del grupo i . Si se usan muchas observaciones, la mínima diferencia puede resultar significativa, por ello usaremos η^2 , una medida de cuánta proporción de la variabilidad se debe a las diferencias entre grupos. La cantidad de efecto que se considera significativa son los valores superiores a 0.14. Se calcula como:

$$\eta^2 = \frac{H - (k - 1)}{N - 1}$$

Si el test de Kruskal-Wallis resulta significativo, se procede al análisis post-hoc para comprobar que grupos difieren entre sí. Para ello utilizaremos el test de Dunn de la librería “dunn.test” de R [24], que compara los pares de grupos posibles y calcula la diferencia entre las medias de los rangos [25]. El estadístico sigue la distribución normal $Z \sim N(0,1)$ y se calcula como:

$$Z = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

En caso de contar con gran cantidad de observaciones, se usa el tamaño del efecto. Para el test de Dunn es el r de efecto, el cual, se calcula a partir del estadístico para comprobar si la diferencia estadísticamente significativa es también influyente en términos prácticos. La obtención de la r de efecto es:

$$r = \frac{Z}{\sqrt{N}}$$

Capítulo 3

Datos

El estudio recoge cuatro conjuntos de datos base provenientes de tres países, Austria, España, Polonia; y la unión de todos los países. La obtención de todos los datos todavía no se encuentra disponible al público, pero lo estarán con una solicitud previa a través de la plataforma ZENODO, dentro del perfil del proyecto K-HEALTHinAIR. Este TFG tiene el correspondiente permiso de la utilización de los datos.

3.1. Características de los datos

Las variables usadas han sido medidas mediante dos dispositivos de monitoreo de la calidad del aire interior, los modelos inBiot (MICA) [26] y Kaiterra (Sensedge Mini) [27] ambos diseñados para medir contaminantes y parámetros ambientales en espacios interiores. Los conjuntos de datos contienen las siguientes variables:

- X_time: Fecha y hora en las que se recogió cada observación, en formato POSIXct de R.
- season: Estación del año (en inglés) correspondiente a: invierno “Winter”, primavera “Spring”, verano “Summer” y otoño “Autumn”.
- sub_scenariio: Tipo de vivienda. “HOM01” es el caso español, sin distinción entre casas. Para Austria y Polonia “HOM02” corresponde a un hogar con calefacción de combustibles sólidos, “HOM04” para combustibles de gas y “HOM05” para calefacción eléctrica.
- area: Identificador de cada hogar, corresponde con su sub_scenariio seguido de un número asociado, por ejemplo: “HOM01005”, “HOM01111” o “HOM04010”.
- Sensor_id: Identificador de cada sensor utilizado.
- co2: Concentración de dióxido de carbono (CO₂) expresada en partes por millón (ppm).
- humidity: Humedad relativa del ambiente, medido en porcentaje.
- pm1, pm2.5, pm10: Concentración de material particulado con diámetros aerodinámicos menores o iguales a 1 µm, 2.5 µm y 10 µm respectivamente, expresadas en microgramos por metro cúbico (µg/m³).
- temperature: Temperatura ambiental en grados Celsius (°C).
- tvocs: Compuestos Orgánicos Volátiles Totales, gases emitidos por ciertos sólidos o líquidos medidos en partes por billón (ppb).
- formaldehyde: Concentración de formaldehído (CH₂O), gas incoloro medido en partes por billon (ppb). Solo disponible para España.

AQI_pm25, AQI_co2 y AQI_formaldehyde (esta última únicamente considerada para el caso español), son variables que clasifican la calidad del aire de los tres compuestos en tres niveles: “GOOD”, moderada “MODERATE” y mala “UNHEALTHY”.

La última variable, HOM_smoking únicamente disponible para Austria y España, diferencia las observaciones de hogares con presencia de un fumador (“YES”) y sin fumador (“NO”). Los valores se encuentran en dos archivos .xlsx independientes de las lecturas de los sensores. Cada conjunto de datos de cada país tiene sus características propias, tanto en estructura como en cobertura temporal. No todos los sensores empiezan a recopilar datos al mismo tiempo, el intervalo de tiempo de muestreo también varía ya que no fueron instalados y activados todos al mismo tiempo como podemos observar en la Figura 3.1.

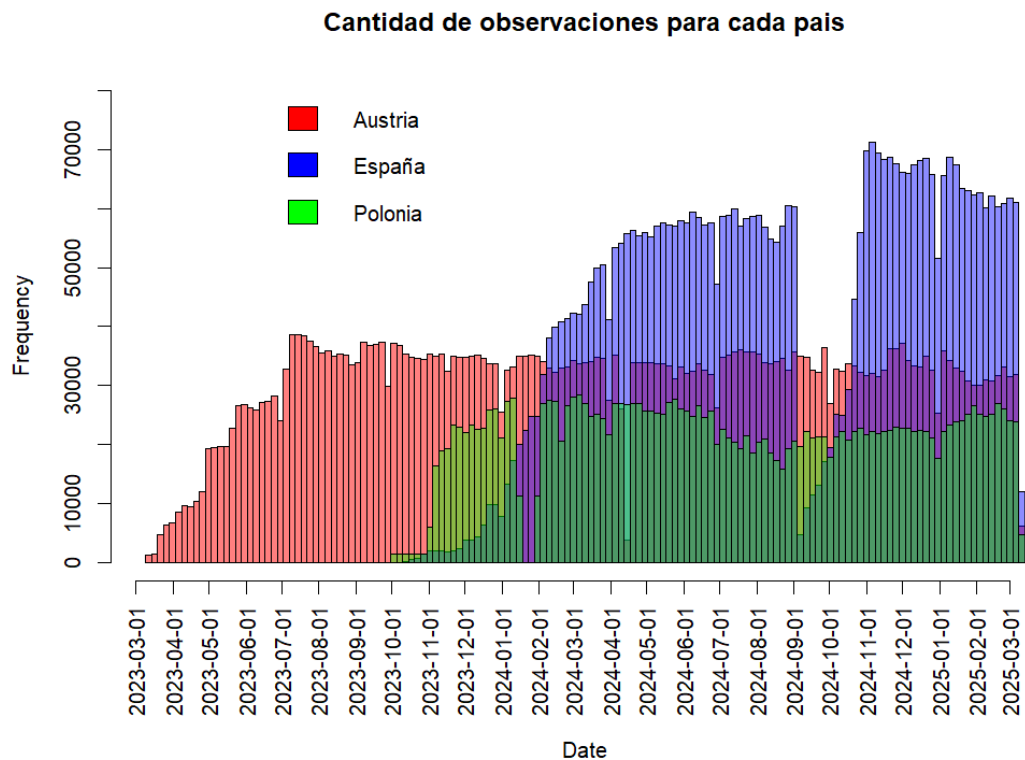


Figura 3.1: Histograma sobre la cantidad de muestras a lo largo del tiempo

Con el objetivo de comparar los conjuntos de los países la Tabla 3.1 resume las principales características como el número de observaciones y filas incompletas (en las que falta algún dato) junto a su proporción sobre el total de datos, el número de hogares del estudio en cada país y el periodo cubierto. El tratamiento de valores missing consiste en eliminarlos o descartarlos, debido a que representan menos del 1% del total y no habrá pérdida de información sustancial.

País	Código	Nº Observaciones	Nº Filas Incompletas	Hogares	Fecha Inicio	Fecha Fin
Austria	AT	4 511 169	22 (0.0005%)	29	2023-03-10 12:02:44	2025-03-09 23:58:43
España	ES	4 412 582	21923 (0.4968%)	100	2023-10-13 08:19:57	2025-03-09 23:59:54
Polonia	PL	2 230 110	50 (0.0022%)	27	2023-10-01 00:02:49	2025-03-09 23:58:28

Tabla 3.1: Resumen de los conjuntos de datos

3.2. Importar y preprocesar los datos

Los datos no provienen de un solo archivo, y tampoco tienen el formato correcto. El formato original es .csv, y los datos de un mismo país vienen separados por fechas. Se importan a R todos los archivos con `fread()` desde la librería “data.table” [28], que permite un manejo más rápido de grandes conjuntos de datos, se limpian variables y se unifican en un solo conjunto por país. El tratamiento de outliers es nulo debido a que estos corresponden a eventos importantes de la vida diaria, por ejemplo, fumar. El código de todo el apartado se encuentra en Anexo A.1.

3.2.1. Recodificación de variables

La variable `X_time` se transforma a formato `POSIXct` en zona horaria UTC para facilitar el futuro manejo de los datos. Creamos la variable `season`, usando la fecha (`X_time`) y la opción de la librería “lubridate” [29], siguiendo los equinoccios y solsticios del hemisferio norte.

3.2.2. Identificación de hogares fumadores

Los archivos “IAQ_summary.xlsx” y “HOMES_Smokers.xlsx” contienen los códigos de los hogares para clasificar dentro de la variable `HOM_smoking` si se fuma o no, esta toma el valor “YES” y “NO” respectivamente. La librería “readxl” [30] se ha utilizado para leer los archivos en R.

3.2.3. Unificación y promedio horario

Para futuros análisis es interesante tener un conjunto de datos con todas las muestras diferenciando a los países creando la variable `Country`, para ello es útil la librería “dplyr” [31]. Los hogares generan mucha variabilidad como se puede comprobar en la Tabla 3.2, que enfrenta las desviaciones típicas de las variables del conjunto de datos original frente a las desviaciones típicas del conjunto de datos al promediar por hora. Como el comportamiento individual es muy diverso, se realizarán algunos estudios en base al promedio horario por variable de cada país. Dado que algunas variables numéricas presentan asimetría pronunciada, en los futuros análisis se evalúa una transformación Box-Cox junto al promedio horario correspondiente.

Variables	Austria		España		Polonia		Total	
	Por casa	Promedio	Por casa	Promedio	Por casa	Promedio	Por casa	Promedio
CO ₂	268.87	99.2	327.73	123.55	375.77	225.5	318.92	108.68
Humedad	10.2	7.78	7.72	6.05	8.95	5.64	9.18	5.81
PM ₁	32.83	11.65	72.58	15.58	192.98	48.7	100.04	15.24
PM _{2.5}	63.31	21.79	89	18.18	322.19	81.4	160.11	26.46
PM ₁₀	54.71	19.42	81.7	17.17	321.63	81.17	157.01	24.9
Temperatura	2.19	1.51	3.49	2.89	2.77	1.93	2.88	2
TVOCs	5198.71	1967.23	6951.93	3003.36	7462.07	2068.57	6425.52	1167.52
Formaldehído	NA	NA	29.37	10.59	NA	NA	29.37	10.59

Tabla 3.2: Desviaciones típicas originales frente el promedio.

3.2.4. Clasificación de la calidad del aire

Se calculan indicadores cualitativos de la calidad del aire (AQI: Air Quality Index), permiten traducir concentraciones numéricas de contaminantes en categorías más sencillas. Dependiendo de los umbrales de la Tabla 3.3 para cada contaminante provenientes de un estudio previo denominado GOAQs (Global Open Air Quality Standards) [32], se han clasificado CO₂, PM_{2.5} y formaldehído en tres niveles:

- “GOOD”: La concentración del contaminante es baja y representa riesgo bajo. Es el nivel ideal en el que mantener las concentraciones.
- “MODERATE”: Señala niveles de contaminación, que aunque no son peligrosos, pueden representar un riesgo para personas más débiles
- “UNHEALTHY”: Representa calidad del aire deficiente, con concentraciones que, si expuestos a largo plazo, pueden afectar a todo tipo de personas.

Para el formaldehído, debido al cambio de unidades (de ppb a µg/m³) se ha procedido a dividir estos umbrales entre 1.228 y mantener así la coherencia en las comparaciones.

Nivel	CO ₂ (ppm)	PM _{2.5} (µg/m ³)	Formaldehído (µg/m ³)
GOOD	≤ 800	≤ 8	≤ 50
MODERATE	(800, 2000]	(8, 36]	(50, 120]
UNHEALTHY	> 2000	> 36	> 120

Tabla 3.3: Umbrales clasificadores del AQI.

Capítulo 4

Análisis estadístico de la base de datos

4.1. Estadística descriptiva

Para entender mejor las características que rodean a las variables tanto numéricas como categóricas, aplicamos las correlaciones, histogramas, resúmenes numéricos y gráficos de barras. Se observarán los comportamientos de las variables y un primer análisis de las relaciones entre ellas. Todo el código se encuentra en el Anexo A.2.

Las correlaciones permiten una vista preliminar de las relaciones entre parámetros, dependiendo del valor del coeficiente de Spearman podremos interpretar patrones generales de los datos y formular unas hipótesis iniciales. Además, el comportamiento entre países es distinto debido a diferentes hábitos de ventilación o el aire externo y puede verse reflejado en las interacciones. Se usan los datos promediados por hora porque disminuyen el ruido provocado por las casas y los patrones se verán más reflejados. Debido al tamaño de los datos, la significancia de correlación entre variables es dada por el umbral de 0.33 en valor absoluto.

El dióxido de carbono presenta una estructura similar en los tres países, una correlación negativa con la temperatura de -0.55 en Austria, -0.48 en Polonia y -0.57 en España. Por lo que el aumento del CO₂ conlleva una disminución de la temperatura. Aunque parezca contradictorio porque al ventilar ambos compuestos disminuyen, en invierno no se ventila tanto y el CO₂ sube con temperaturas bajas, y en verano se ventila más, entra calor y baja el CO₂. Este compuesto, las partículas en suspensión y los compuestos volátiles presentan unas correlaciones positivas entre sí demostrando que el aumento de uno de los contaminantes suele indicar que los demás están aumentando también resultando en una calidad del aire baja.

La temperatura y la humedad presentan un coeficiente de correlación alto sobre todo en Austria del 0.63 (Tabla 4.1). Sin embargo, su comportamiento es distinto para Polonia y España (Tablas 4.2 y 4.3), esta correlación no es tan alta como el caso austriaco y la temperatura parece mantener alguna relación con el resto de contaminantes, no solo el CO₂, evento que no ocurre en Austria. Esto podría indicar que la parte de la temperatura que se correlaciona con la humedad no es la misma que la correlacionada con los contaminantes de forma negativa.

En cuestión con los volátiles, salvo para el caso del CO₂ en los tres países con un coeficiente positivo, no parece generar mucha correlación con la temperatura y la humedad salvo el caso español, cuyos coeficientes negativos nos indican que a mayor cantidad de volátiles menor cantidad de temperatura/humedad. El formaldehído presenta una total contradicción frente a los compuestos volátiles ya que, aunque su coeficiente de correlación no es significativo (-0.18) las correlaciones con los demás parámetros son contrarios. Lo que nos conduce a que aun siendo volátil su comportamiento es totalmente distinto al resto.

Por último, destaca el caso de las correlaciones entre partículas. Al medir fracciones del mismo contaminante, presentan coeficientes de 1 entre sí, dando a indicar que su comportamiento es idéntico. PM_{2.5} resulta el más usado en estudios [33] por lo que se comparará solo este compuesto y extrapolaremos para los demás en algunos análisis.

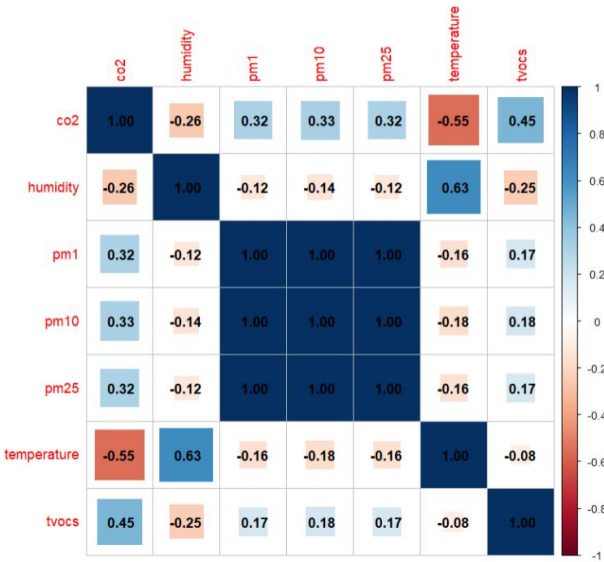


Tabla 4.1: Matriz de correlaciones entre variables observadas en Austria.



Tabla 4.2: Matriz de correlaciones entre variables observadas en Polonia.

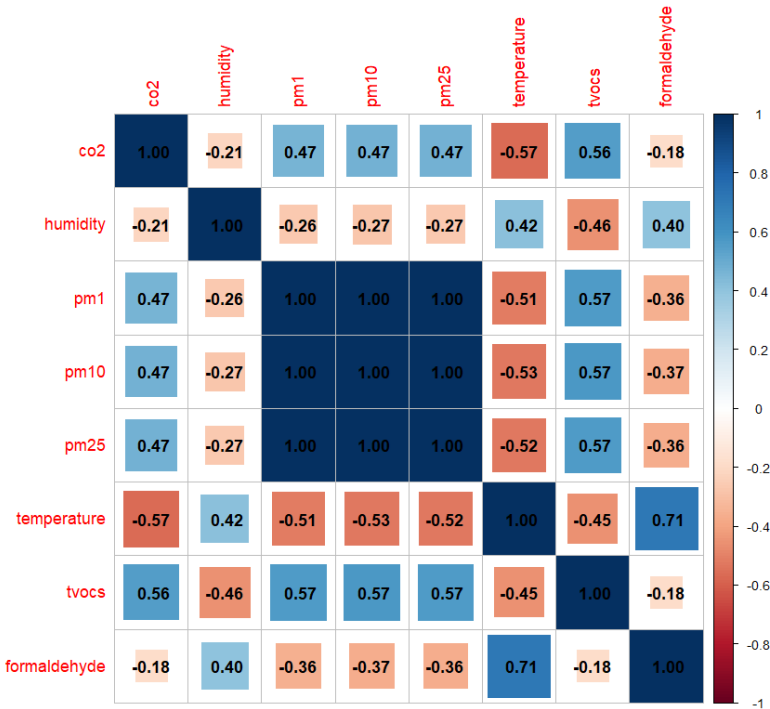


Tabla 4.2: Matriz de correlaciones entre variables observadas en España.

Para comprobar distribuciones y conocer un poco más de cada variable se generan gráficos de densidades por país, por estaciones y su total, además de estadísticos básicos que nos ayuden a entender las diferencias entre países.

El dióxido de carbono presenta para todos los países un pico en verano, dando a entender que varía menos durante el verano, es decir, tiene muchos valores cercanos a la media. En general, donde más varía el CO₂ es en la primavera, siendo la curva menos alta que el resto de estaciones. España y Austria presentan colas ligeras, la Tabla 4.1 nos indica que la curtosis no se aleja mucho de 3, y una asimetría moderada cercana a 0. Polonia, sin embargo, presenta una distribución mucho más asimétrica (3.62) y con colas muy largas, como muestra su curtosis (33.12). Para el total, al hacer la media de todas las mediciones, la mayor parte de las colas se compensan con los valores de los otros países y por ello presenta menor curtosis y asimetría. En la Figura 4.1 también se comprueba que, en las estaciones de frío, el CO₂ es mayor que en las de calor, verano y primavera presentan valores más bajos que invierno y otoño.

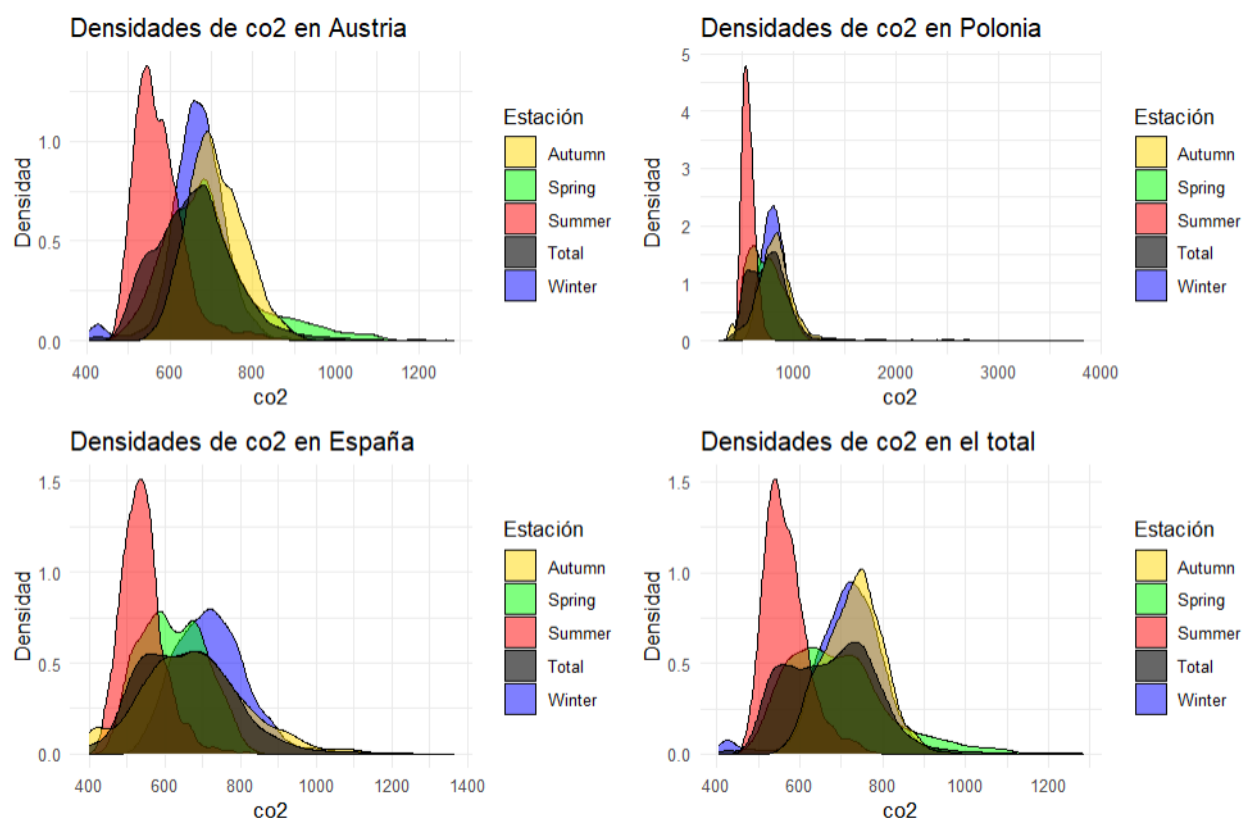


Figura 4.1: Distribuciones del CO₂.

País	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
Austria	404.6	592.8	657.3	661.1	716	1286.6	0.789	4.99
España	400	563.6	651.3	658.3	737.5	1365	0.596	3.66
Polonia	272.4	611.7	749.9	760.1	863.3	3832.4	3.624	33.12
TOTAL	404.6	587.2	676.9	675.1	750.6	1286.6	0.414	3.5

Tabla 4.4: Resumen del CO₂.

La variación estacional también se presenta en la humedad, siendo ahora el invierno la estación con valores más bajos y verano con los valores más altos. En la tabla vemos que, aunque las distribuciones de los países varían tanto, la media tiene valores similares y que la diferencia de las costumbres, por ejemplo, de ventilación, no genera tanta diferencia en valor medio.

Todas las distribuciones tienden a ser simétricas, destacando un poco más la asimetría positiva. La humedad no presenta tantas colas dando a entender que es difícil encontrar valores extremos de esta medida, y por ello la curtosis no es tan alta como el caso del CO₂. Austria y el total, al presentar distribuciones más aplanadas tienen una curtosis baja demostrando una mayor dispersión, al contrario de España y Polonia cuyos valores parecen no desviarse tanto de la media.

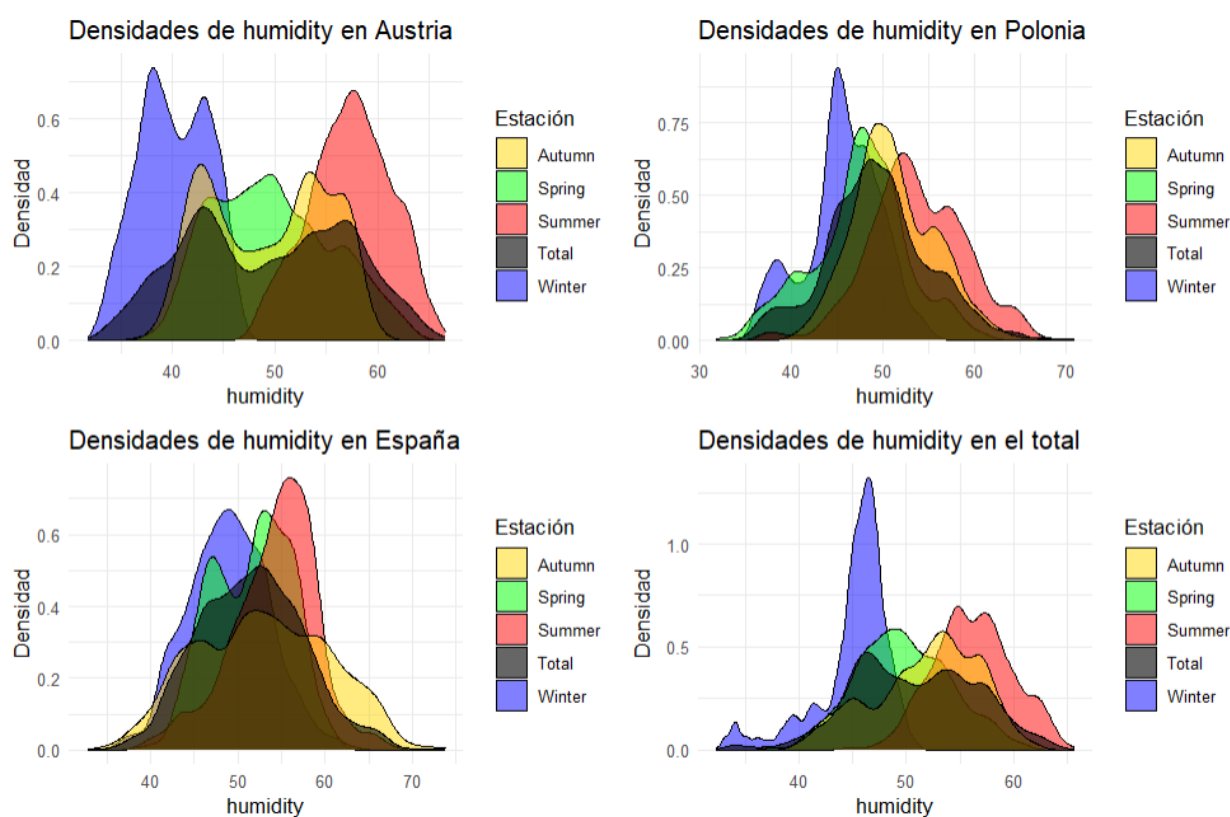


Figura 4.2: Distribuciones de la humedad.

País	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
Austria	31.79	42.74	49.47	49.25	55.97	66.6	-0.008	1.91
España	32.8	47.09	51.58	51.56	55.7	73.8	0.168	2.82
Polonia	31.79	45.71	49.07	49.3	52.62	70.92	0.142	3.16
TOTAL	32.25	46.35	50.61	50.75	55.18	65.76	-0.098	2.64

Tabla 4.5: Resumen de la humedad.

Las concentraciones de $PM_{2.5}$ están fuertemente sesgadas a la derecha. Los valores generalmente son cercanos a la media, pero hay eventos puntuales de alta concentración, por ejemplo, al fumar. Esto ocurre sobre todo en Polonia donde en otoño hay generada una segunda moda. En todos los países el verano presenta la menor de las dispersiones, en Polonia, además del verano la primavera destaca por este hecho.

Esta presencia de colas grandes se ve reflejada tanto en la asimetría, con valores lejanos al 0, en España salvo verano las demás estaciones parecen relajarse y tener una densidad más simétrica y por ello el valor es el más bajo de todos los países. Debido a estas colas, la curtosis se aleja mucho de la normal (3) y recoge valores muy altos cercanos al 10.

En este caso, como todas las distribuciones son parecidas, el total sufre también de extremas asimetría y curtosis.

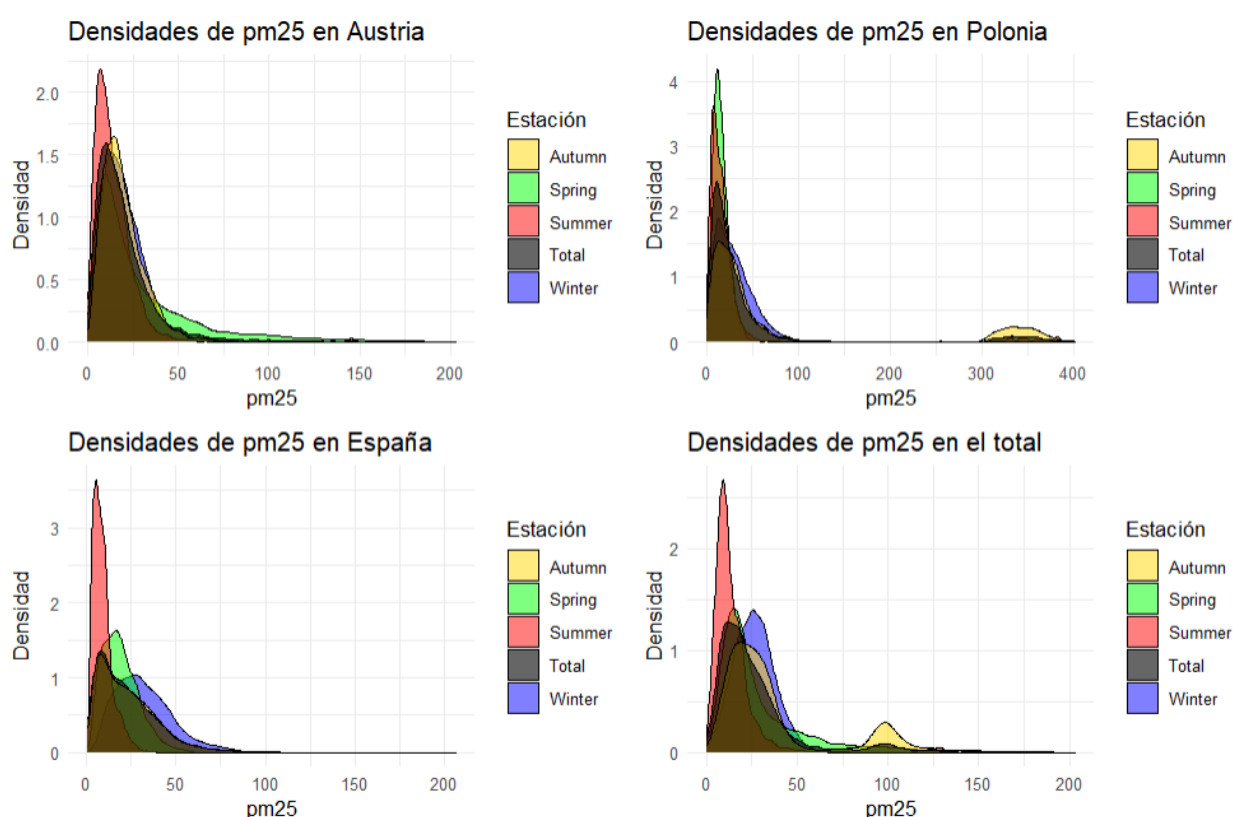


Figura 4.3: Distribuciones de $PM_{2.5}$.

País	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
Austria	0	9.36	15.97	20.71	25.44	203.51	3.352	19.86
España	0.5	9.78	19.15	22.89	31.66	207.13	1.692	9.21
Polonia	0	11.44	20.4	45.17	35.92	401.6	3.238	12.17
TOTAL	0	12.24	20.31	27.18	31.86	203.51	2.463	10.1

Tabla 4.6: Resumen de $PM_{2.5}$.

La estacionalidad vuelve a reflejarse con la temperatura, picos en invierno y otoño, donde la ventilación es menor y por tanto la temperatura interior no varía tanto. También se demuestra que el exterior es un factor influyente, es meses fríos la temperatura interior es menor que en los meses calurosos.

La curtosis es menos acentuada indicando eventos extremos poco probables, sin embargo, la diferencia estacional genera un aumento de los valores en verano, siendo más relevantes las temperaturas bajas que las altas, provocando una asimetría positiva.

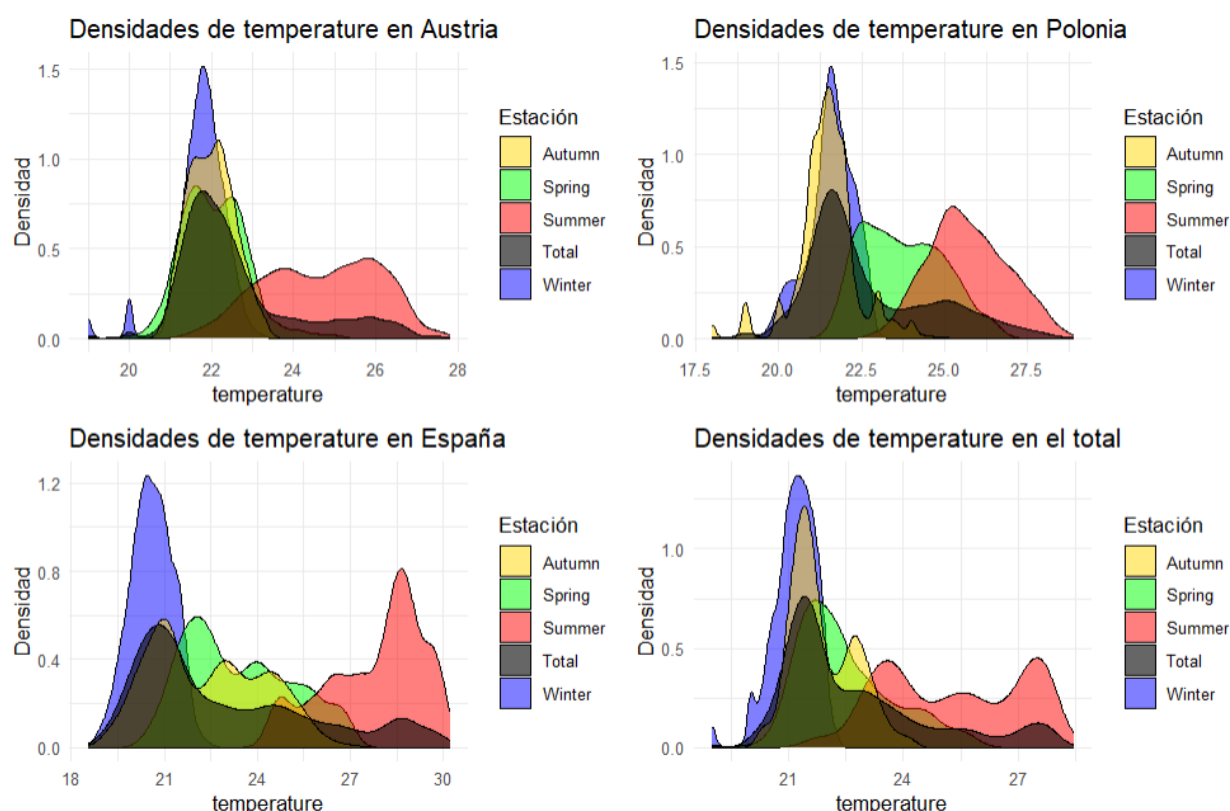


Figura 4.4: Distribuciones de la temperatura.

País	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
Austria	19	21.66	22.2	22.65	23.08	27.82	1.183	3.84
España	18.55	20.72	21.84	22.98	24.83	30.21	0.845	2.59
Polonia	18	21.39	22.01	22.7	24	28.89	0.815	2.99
TOTAL	19	21.32	21.97	22.74	23.64	28.46	1.125	3.41

Tabla 4.7: Resumen de la temperatura.

Al igual que CO_2 y $\text{PM}_{2.5}$, los volátiles presentan menor varianza en verano y por tanto picos en las densidades. Las estaciones frías parecen generar valores más altos de TVOCs debido a que la campana en estos meses está presente en valores superiores que los de verano o primavera.

La aparición de valores extremos es clara debido a las colas hacia la derecha, esto se demuestra en los altos valores de curtosis confirmando la presencia de valores atípicos. Al realizar el total, parece que los valores atípicos se ven reducidos por los valores de los demás países resultando en una menor asimetría.

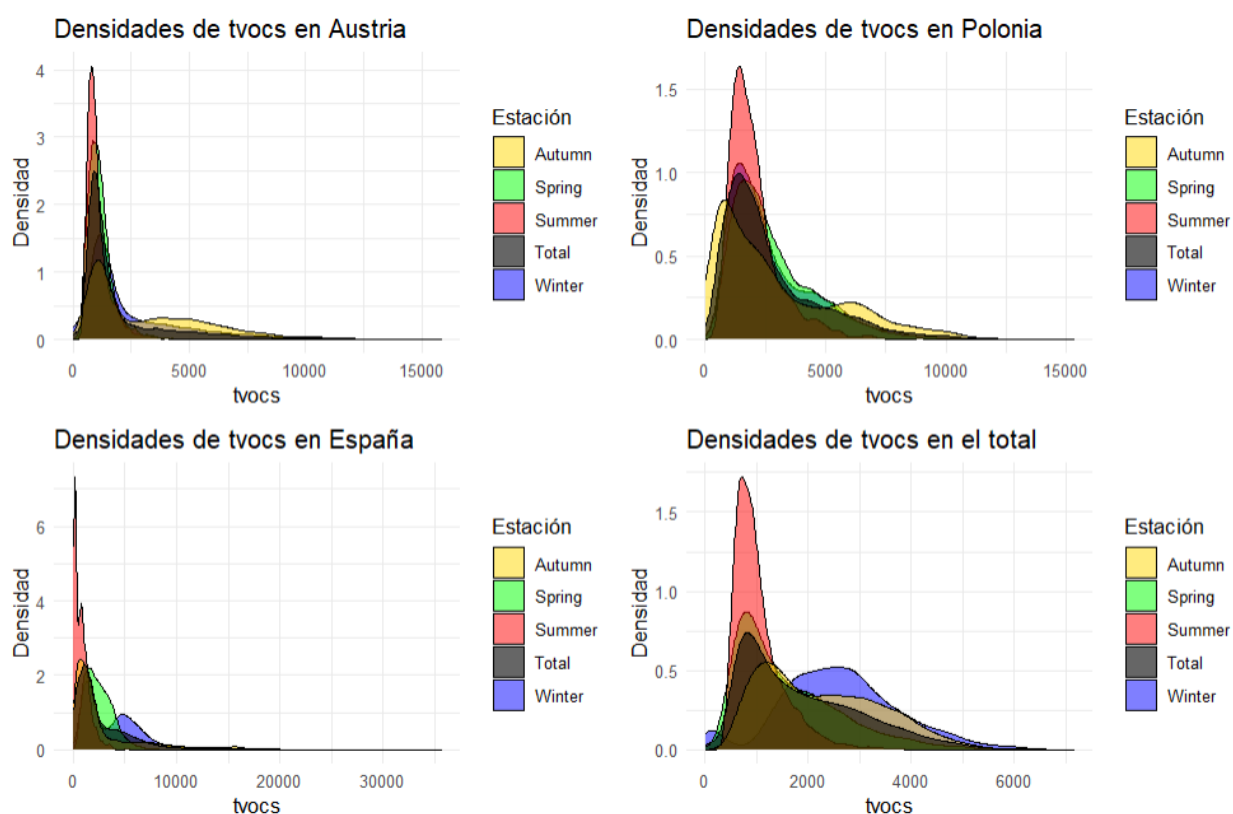


Figura 4.5: Distribuciones de los volátiles.

País	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
Austria	7.17	858.31	1184.35	1967.85	2039.35	15879.87	2.343	8.89
España	3.67	903.36	1703.7	2783.18	3877.35	35736.62	2.486	12.09
Polonia	25.25	1311.83	2103.92	2752.67	3602.33	15335.58	1.529	5.56
TOTAL	7.17	918.54	1572.99	1875.63	2641.45	7171.68	0.896	3.23

Tabla 4.8: Resumen de los volátiles (TVOCs).

La estacionalidad del formaldehído presenta valores más bajos y agrupados en invierno, mismo patrón que para temperatura y humedad. Presenta una muy baja asimetría (-0.056), pero con una curtosis lejana al 3, posiblemente debido a la cola de la derecha y al cúmulo de valores cercanos al 0 en otoño.

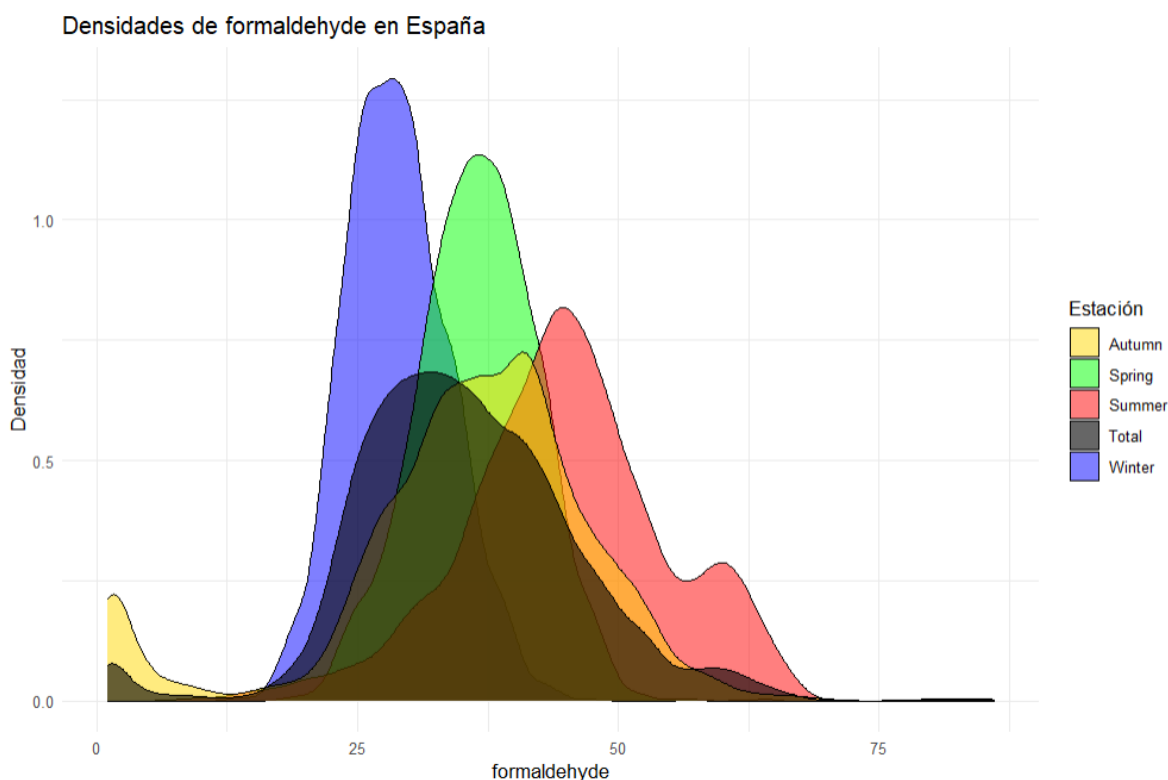


Figura 4.6: Distribuciones del formaldehído.

Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Máximo	Asimetría	Curtosis
1	28.45	34.64	35.23	41.65	86.12	-0.056	4.39

Tabla 4.9: Resumen del formaldehído.

Estas distribuciones junto con la ayuda de las correlaciones los comportamientos son distinguibles en dos grupos:

Los contaminantes CO₂, PM_{2.5} y TVOCs comparten distribuciones fuertemente sesgadas hacia la derecha, asimetría positiva. Las curtosis altas para partículas y volátiles sugieren una mayor concentración de valores alrededor de la media y presencia mayor de eventos extremos.

El segundo grupo está representado por temperatura, humedad y formaldehído (para España), eventos extremos menos probables con valores más cercanos a la media, sin embargo, la estacionalidad de estos parámetros es más evidente ya que es donde se concentra la mayor parte de su varianza, no como el primer grupo cuya estacionalidad puede verse oculta debido a los valores atípicos. Esto indica que el segundo grupo suele estar relacionado a procesos más estables y controlados que el primero.

Para las variables cualitativas, el análisis descriptivo se realiza con diagramas de barras que permiten visualizar de manera efectiva la mayor cantidad de información sobre las interacciones.

Austria

La Figura 4.7 muestra los resultados en Austria. Los diagramas de barras muestran el porcentaje de observaciones de cada tipo de calefacción en cada nivel de AQI. Por cada barra generada, se muestra la proporción de fumadores de las observaciones que la componen. Así, si se quiere comprobar cuantas observaciones tiene HOM02 con buenos niveles de CO₂, se suman el 16.5% que no fuman más 11.3% que fuman, resultando en 27.8% de las observaciones en esa interacción de niveles. Si se suman todos los porcentajes resulta en el 100% de las observaciones. Esto se repite para los niveles de AQI de CO₂ y los niveles de AQI de PM_{2.5}.

Según la altura de las barras, se observa que la calefacción con gas (HOM04) es la más predominante en las observaciones, seguida de los combustibles sólidos (HOM02) y calefacción eléctrica (HOM05).

La calidad del aire en función del CO₂ es generalmente buena, la cantidad de observaciones es significativamente mayor que para moderada y no saludable. La distribución de fumadores también varía, resaltando que aproximadamente la mitad de los hogares con combustibles sólidos presentan fumadores contrastando con HOM05 que presenta el menor porcentaje de fumadores. Según empeora la calidad del aire el número de observaciones con fumadores respecto al de no fumadores incrementa, mostrando que fumar tiene un mayor impacto en la calidad del CO₂.

Si se evalúa la calidad del aire mediante PM_{2.5}, hay un descenso claro de la misma. Si bien sigue predominando buena calidad, hay un notable incremento en las observaciones de moderada y no saludable. Este deterioro también viene asociado de forma más clara con la presencia de fumadores, las observaciones en hogares fumadores son tres veces más frecuentes que las no fumadoras para la categoría no saludable.

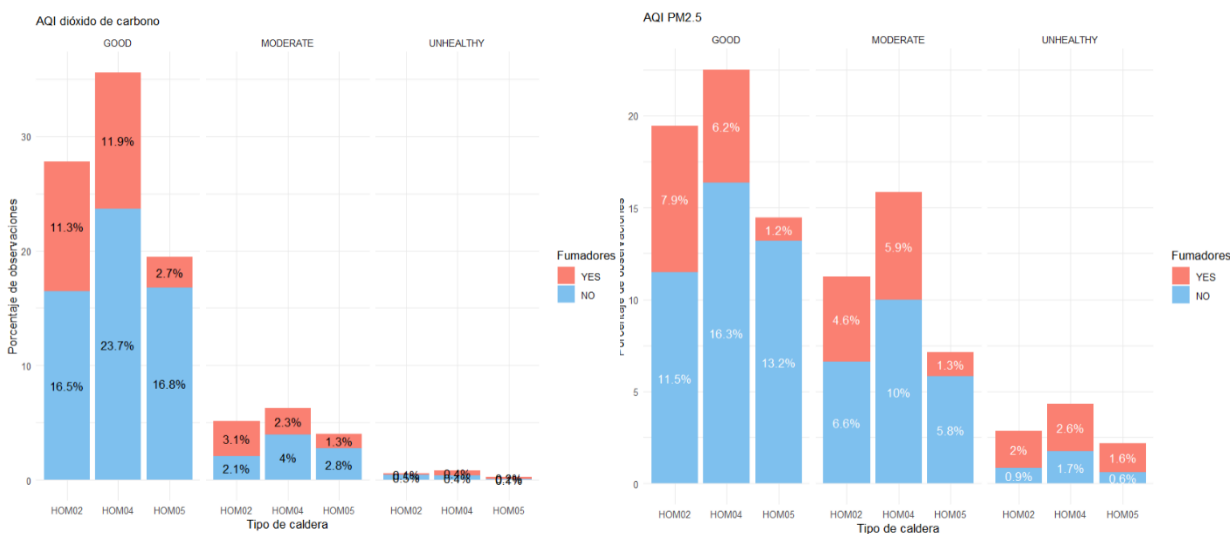


Figura 4.7: Distribución de las observaciones por las interacciones de las variables categóricas en Austria.

España

En el caso español (Figura 4.8), al no haber diferenciación de calderas simplemente dividimos por categorías de las variables AQI. La buena calidad del aire sigue predominante, siendo el CO₂ el que mantiene mejores niveles. En el caso de las partículas, el 9.4% de las observaciones se encuentran en malas condiciones y para el formaldehído un 30.2% del tiempo en una calidad del aire moderada resultando en menos observaciones con buena clasificación.

Es observable que proporcionalmente CO₂ y formaldehído mantienen una distribución de fumadores aproximadamente similar a lo largo de las categorías. Si comparamos esto con AQI de PM_{2.5}, se comprueba que para niveles bajos de AQI la presencia del hábito de fumar es baja, pero según aumenta la concentración de partículas, aumenta el porcentaje de observaciones asociadas a fumadores. Si se compara con lo visto en Austria, se puede comprobar que sobre todo para PM_{2.5}, fumar constituye un efecto importante a la hora de comprobar la calidad.

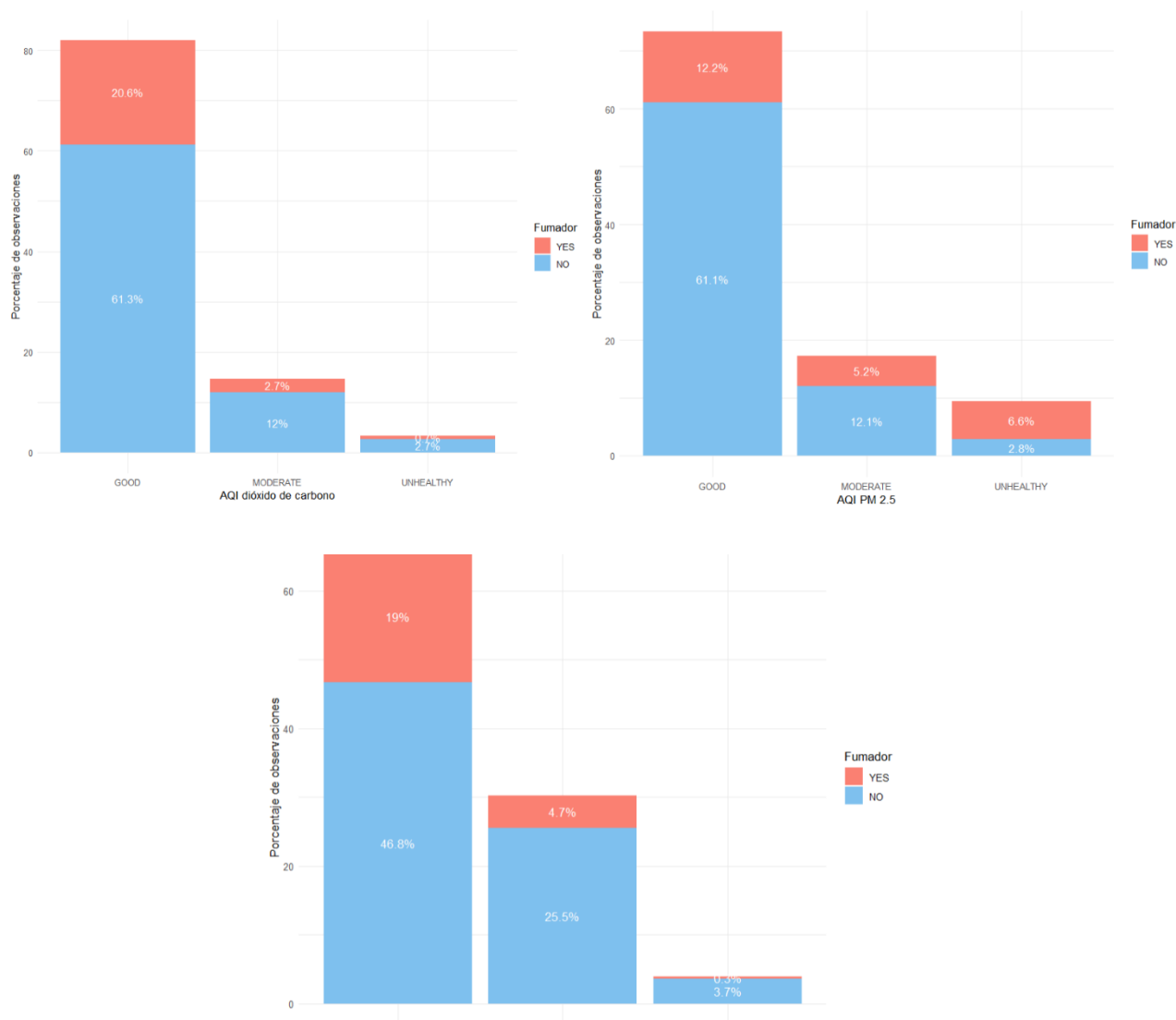


Figura 4.8: Distribución de las observaciones por las interacciones de las variables categóricas en España

Polonia

La Figura 4.9 representando a Polonia, toma el mismo gráfico que Austria, sin embargo, como Polonia no contiene la variable de fumadores (HOM_smoking) ha sido reemplazada con la variable AQI del CO₂. Por lo que los distintos colores ahora representan la cantidad de observaciones que hay en cada barra de los niveles de AQI_co2

A diferencia de Austria, el sistema de calefacción principal es la electricidad con un 42.7% de los casos, seguido de combustible gaseoso con un 42.1% y finalmente el combustible sólido representando el 15.2% de los datos.

En términos generales, predomina un AQI del CO₂ bueno, recogiendo un 70% de los datos para esta categoría. No obstante, analizando el AQI del PM_{2.5}, Polonia presenta niveles significativamente menores que los de España o Austria. La mayoría de las observaciones se encuentran en la categoría moderada superando a las de buena calidad. Según empeora la calidad de PM_{2.5}, la calidad relativa del CO₂ también disminuye concordando con lo visto en las variables numéricas.

El gráfico también pone en evidencia que el tipo de caldera está asociado con los niveles de AQI, demostrando que las casas con combustibles sólidos presentan la misma cantidad de datos en niveles no saludables de PM_{2.5} que saludables, contrastando con Austria donde no había diferencia aparente entre tipos de calefacción.

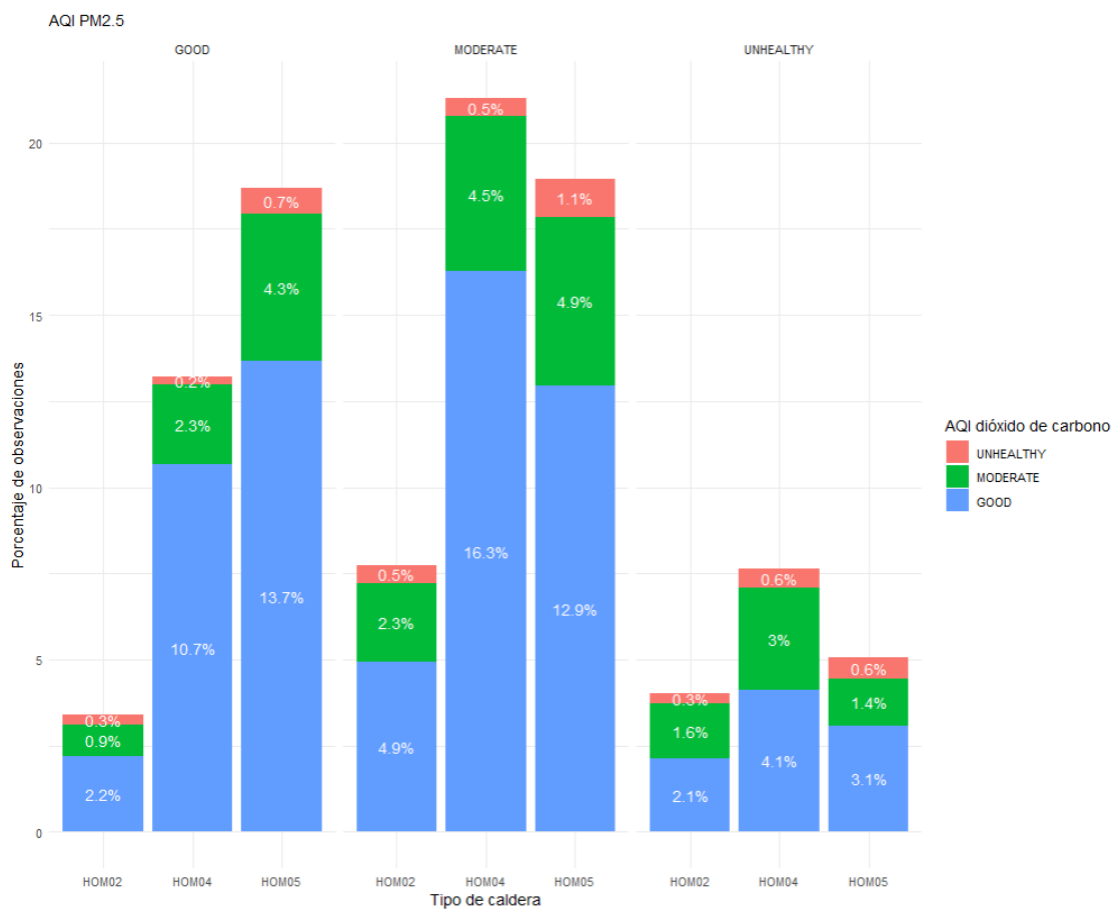


Figura 4.9: Distribución de las observaciones por las interacciones de las variables categóricas en Polonia.

4.2. Normalidad de los datos

Las pruebas de normalidad como Shapiro-Wilk o Lilliefors son muy sensibles al tamaño muestral, desviaciones muy pequeñas de la normalidad hacen que los test rechacen la hipótesis nula. Dado el gran volumen de los datos, los test se aplicaron sobre muestras aleatorias de tamaño 5000. Se realizaron 1000 iteraciones del muestreo para comprobar la estabilidad de los resultados. Para garantizar la comparabilidad entre variables, se utilizaron las mismas muestras aleatorias en todos los casos, obteniendo control sobre el efecto aleatorio.

4.2.1. CO₂

Se realiza una transformación Box-Cox para intentar corregir la asimetría positiva del CO₂, como podemos observar en la Figura 4.10, en Austria el valor de lambda asociado a la transformación es tanto negativo como distinto de 0 en un intervalo del 95%. Debido a las grandes cantidades de datos, el estimador tiende a estabilizarse y el intervalo de confianza es mínimo.

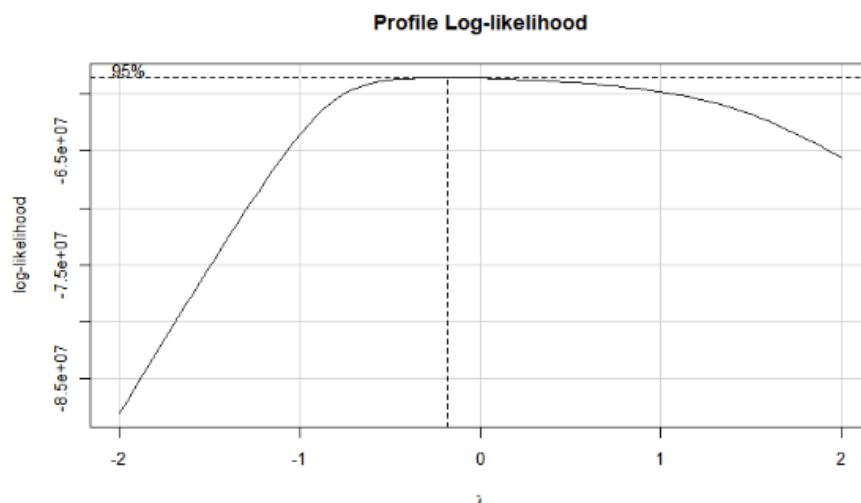


Figura 4.10: Log-verosimilitud perfilada de CO₂ en Austria.

En los tres países, el estimador λ resulta negativo. Esta característica implica que los valores grandes sean más pequeños reduciendo su impacto, como podemos ver en los diagramas cuantil-cuantil, donde los valores atípicos ahora se encuentran en los valores bajos tratando de corregir la asimetría original. Aún con la transformación, en los tres países se rechaza la posibilidad de atribuir una distribución normal al CO₂ debido a que en ninguna de las muestras se alcanza un p-valor mayor a una confianza del 95%.

País	Lambda	Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
Austria	-0.1818	$1.443 * 10^{-13}$	0 %	$1.369 * 10^{-8}$	0 %
España	-0.1414	$1.472 * 10^{-24}$	0 %	$1.595 * 10^{-45}$	0 %
Polonia	-0.2626	$7.199 * 10^{-10}$	0 %	$2.011 * 10^{-6}$	0 %

Tabla 4.10: Resumen normalidad de CO₂.

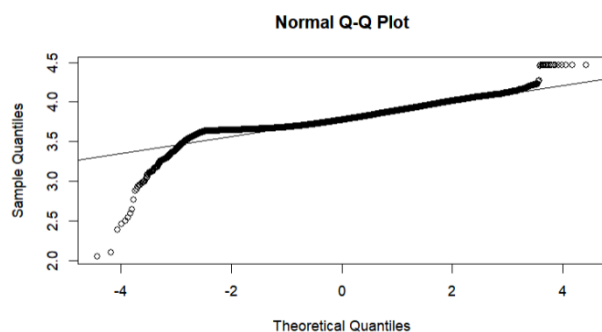


Figura 4.11: Q-Q plot de CO₂ en Austria.

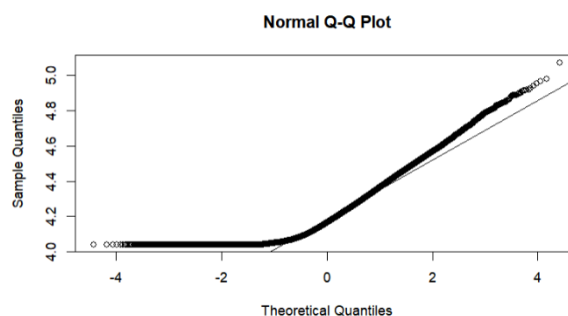


Figura 4.12: Q-Q plot de CO₂ en España.

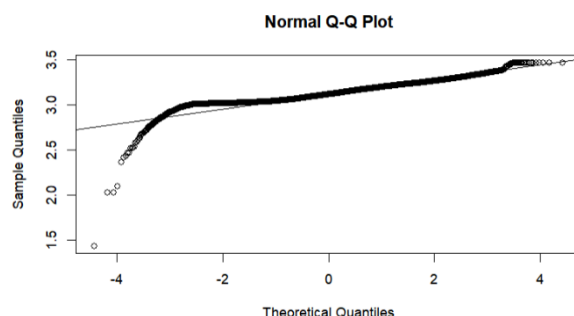


Figura 4.13: Q-Q plot de CO₂ en Polonia.

4.2.2. Humedad

La humedad no tiene una asimetría clara, por lo que no se usa Box-Cox. En Austria, los resultados de las pruebas de normalidad demuestran p-valores muy bajos lo que indica una desviación respecto a la normalidad. Sin embargo, en España y Polonia, los resultados son considerablemente mejores, alcanzando un 80% de las muestras españolas un p-valor superior a 0.05 y un 38% y 53% en el caso polaco. Estas conclusiones se ven reforzadas por las Figuras 4.15 y 4.16 que muestran un alineamiento más cercano a la recta teórica. Por tanto, en España y Polonia la humedad tiene una distribución normal, caso no asumible en Austria.

País	Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
Austria	$2.795 \cdot 10^{-5}$	0 %	$1.867 \cdot 10^{-4}$	0 %
España	0.285	81.5%	0.273	80%
Polonia	0.089	38.1%	0.141	53.4%

Tabla 4.11: Resumen normalidad de humedad.

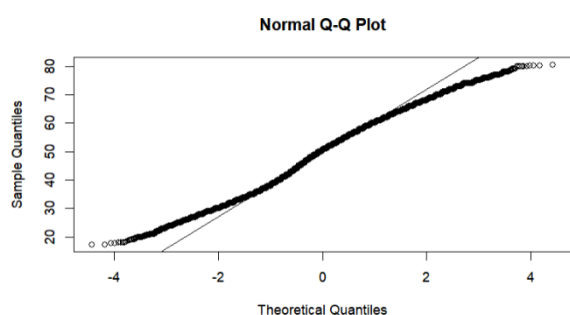


Figura 4.14: Q-Q plot de humedad en Austria.

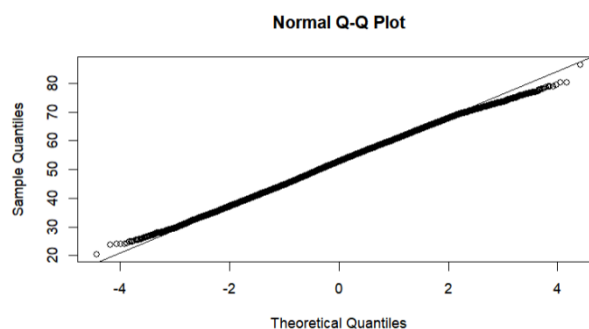


Figura 4.15: Q-Q plot de humedad en España.

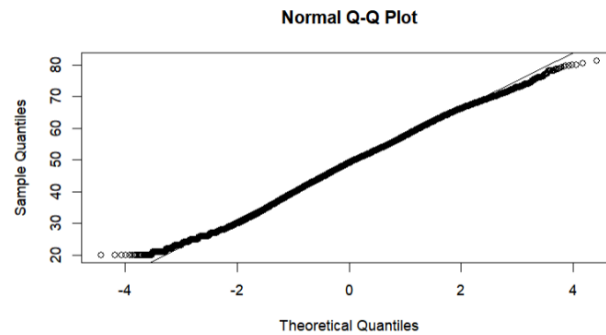


Figura 4.16: Q-Q plot de humedad en Polonia.

4.2.3. Partículas

Al igual que el CO_2 , se aplica la transformación Box-Cox a las partículas. Pero a diferencia del primero, para Austria como Polonia λ es positivo distinto de 0. En España, λ es distinto de 0 pero es lo suficientemente pequeño como para poder aproximar la transformación a logaritmo, como podemos observar en la Figura 4.17.

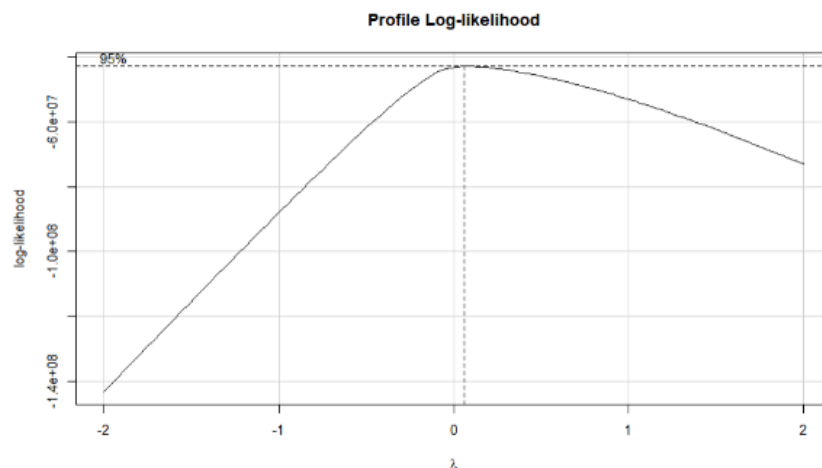


Figura 4.17: Log-verosimilitud perfilada de $\text{PM}_{2.5}$ en España.

Aún con la transformación Box-Cox, ninguna de las distribuciones de las partículas resulta normal. Teniendo λ positivo menor a 1 permite expandir los valores pequeños y suavizar los grandes. Si comprobamos el gráfico de densidad de las partículas, esto provoca un aumento de la influencia de los datos agrupados en la moda y debilita la cola de la distribución. El logaritmo simplemente suaviza la cola reduciendo el peso de los outliers. Sin embargo, esta transformación no parece suficiente como para aceptar normalidad, ninguna muestra alcanza el nivel de significancia y los diagramas cuantil-cuantil no están alineados con la recta teórica en ninguno de los tres casos.

País	Lambda	Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
Austria	0.2222	$1.602 * 10^{-24}$	0 %	$5.640 * 10^{-60}$	0 %
España	0.0606	$3.088 * 10^{-28}$	0 %	$3.642 * 10^{-52}$	0 %
Polonia	0.1818	$2.969 * 10^{-28}$	0 %	$6.462 * 10^{-36}$	0 %

Tabla 4.12: Resumen normalidad de partículas.

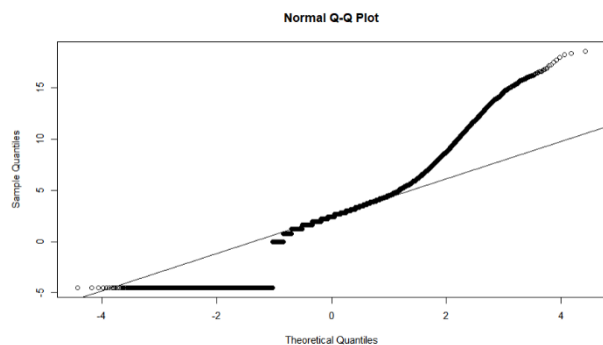


Figura 4.18: Q-Q plot de PM_{2.5} en Austria.

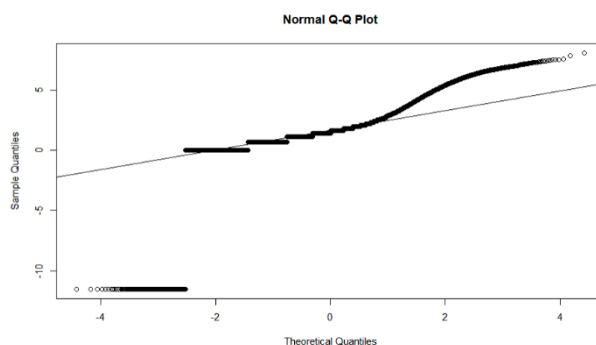


Figura 4.19: Q-Q plot de PM_{2.5} en España.

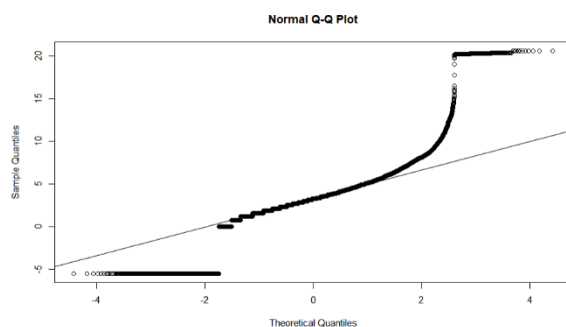


Figura 4.20: Q-Q plot de PM_{2.5} en Polonia.

4.2.4. Temperatura

En las densidades de temperatura (Figura 4.4) muestran una asimetría positiva, pero esto sucede por una bimodalidad en los datos provocada por las observaciones en verano. La transformación Box-Cox no corrige este suceso por lo que las pruebas de normalidad son aplicadas sobre la variable original. Debido a que todas las muestras rechazan normalidad y los diagramas correspondientes se ven influidos por los valores atípicos, tanto bajos como altos, la temperatura no es una variable con distribución normal.

País	Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
Austria	$1.621 * 10^{-7}$	0 %	$5.234 * 10^{-12}$	0 %
España	$5.021 * 10^{-12}$	0 %	$9.273 * 10^{-12}$	0 %
Polonia	$7.249 * 10^{-5}$	0 %	$5.210 * 10^{-5}$	0 %

Tabla 4.13: Resumen normalidad de temperatura.

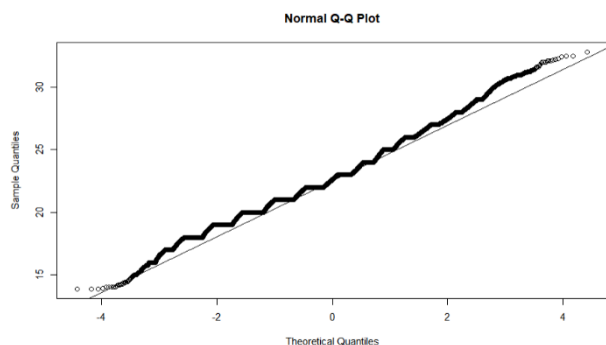


Figura 4.21: Q-Q plot de temperatura en Austria.

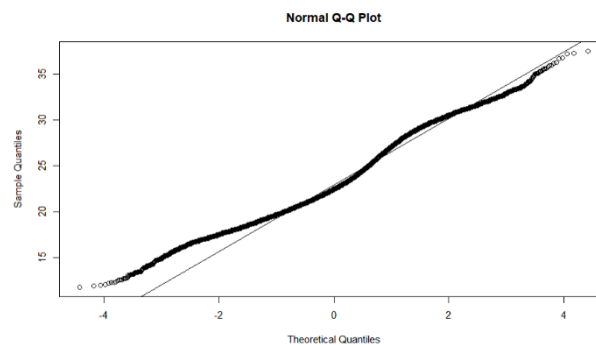


Figura 4.22: Q-Q plot de temperatura en España.

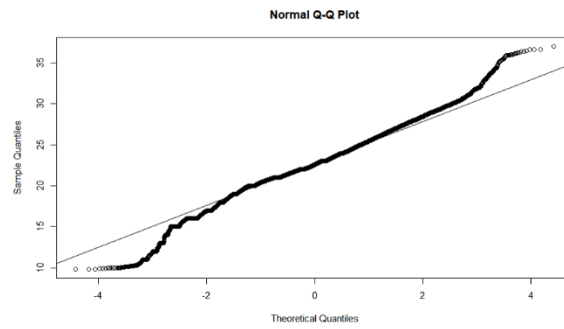


Figura 4.23: Q-Q plot de temperatura en Polonia.

4.2.5. TVOCs

Como se observó en la comparación de densidades, los compuestos volátiles también poseen asimetría positiva, por lo que se realiza la transformación Box-Cox para buscar normalidad.

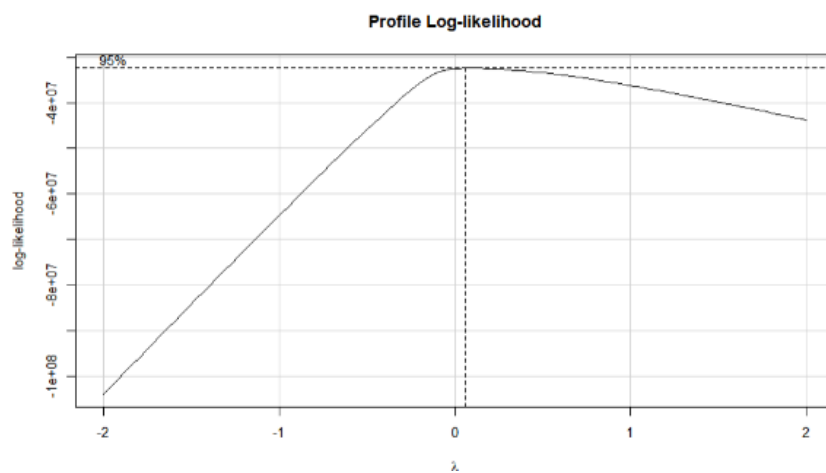


Figura 4.24: Log-verosimilitud perfilada de TVOCs en Polonia.

El caso de los volátiles parece evidente aplicar el logaritmo en Austria y Polonia, ya que los valores de λ obtenidos son cercanos a cero (0.0606 en ambos casos). Aun aplicando la transformación, ninguno de los tres países se les atribuye normalidad a sus valores de TVOCs. Las medias menores a 0.05, los porcentajes prácticamente nulos y las desviaciones respecto a la línea teórica de los diagramas cuantil-cuantil respaldan rechazar la normalidad del parámetro.

País	Lambda	Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
Austria	0.0606	$5.536 * 10^{-6}$	0 %	$5.471 * 10^{-4}$	0 %
España	0.1414	$2.045 * 10^{-24}$	0 %	$5.260 * 10^{-42}$	0 %
Polonia	0.0606	$6.063 * 10^{-5}$	0 %	$2.033 * 10^{-3}$	0.9 %

Tabla 4.14: Resumen normalidad de TVOCs.

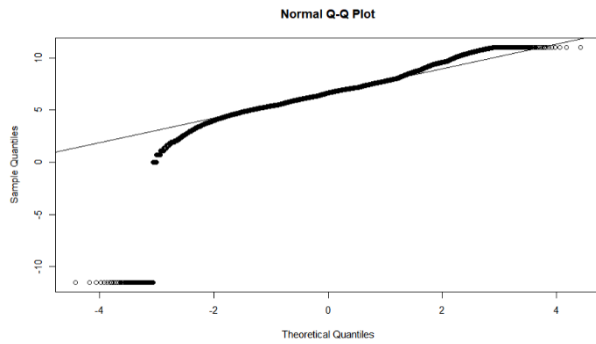


Figura 4.25: Q-Q plot de TVOCs en Austria.

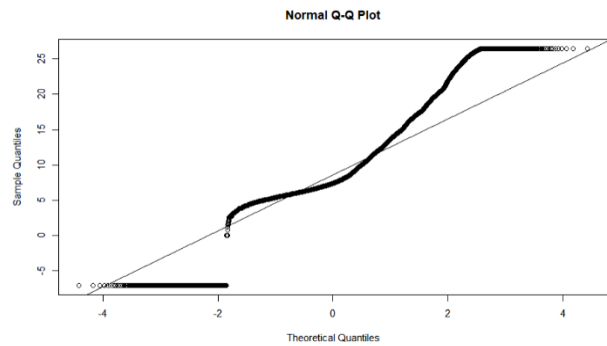


Figura 4.26: Q-Q plot de TVOCs en España.

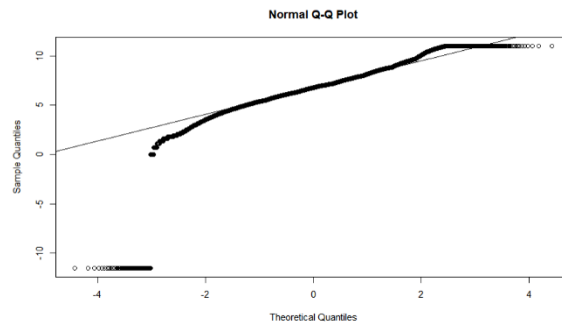


Figura 4.27: Q-Q plot de TVOCs en Polonia.

4.2.6. Formaldehído

Al igual que humedad, las pruebas de normalidad se aplican directamente sobre los datos originales, ya que, en la Figura 4.6 puede percibirse la densidad de una distribución normal. Tras aplicar el muestreo y las pruebas correspondientes, rechazamos normalidad del formaldehído como se puede comprobar en el gráfico

Media p-valor Shapiro-Wilk	Porcentaje p-valor > 0.05	Media p-valor Lilliefors	Porcentaje p-valor > 0.05
$4.362 * 10^{-24}$	0%	$1.489 * 10^{-22}$	0%

Tabla 4.15: Resumen normalidad de formaldehído.

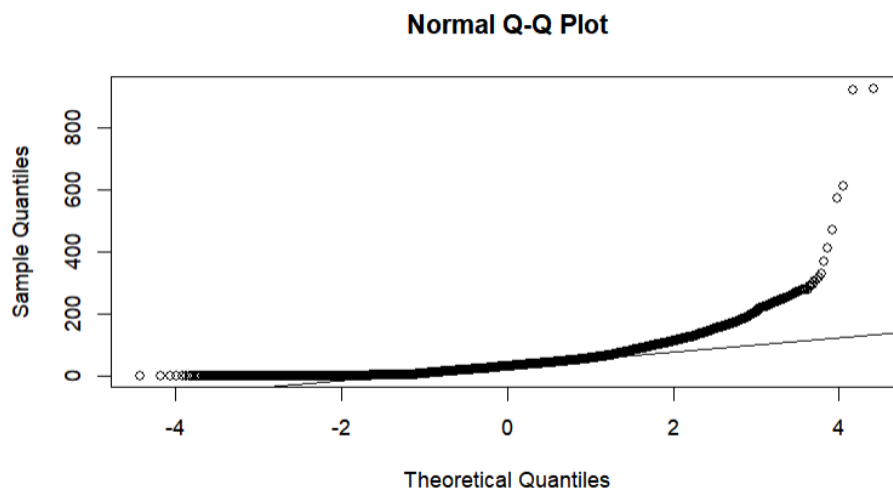


Figura 4.28: Q-Q plot de formaldehído en España.

4.3. Análisis de Componentes Principales

Para conseguir incorporar al estudio más posibles relaciones entre las variables, se realiza ACP para cada país. Esto servirá también para reducir el tamaño del número de variables procurando mantener la mayor cantidad de información. Usaremos los datos promediados por horas tras realizar un escalado. Para mayor rigor, se aplica este mismo método a los datos transformados mediante Box-Cox provenientes de las pruebas de normalidad y comprobar la función de suavizar los valores atípicos.

4.3.1. Datos originales

Austria

En la Tabla 4.16 se aprecia la cantidad de información explicada por cada componente principal en Austria. Podemos comprobar que con tan solo 5 componentes principales podemos explicar aproximadamente el 100% de los datos. Siguiendo los criterios de elección de dimensionalidad, en específico la regla de Kaiser, se usarían las dos primeras dimensiones siendo las únicas que superan el valor unitario. Explican un 74.33% de la información total, siendo un porcentaje de varianza alto.

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Desviación estándar	1.8597	1.3209	0.949	0.8158	0.4774	0.0525	0
Proporción de varianza	0.4941	0.2492	0.1287	0.0951	0.0326	0.0004	0
Proporción acumulada	0.4941	0.7433	0.872	0.967	0.9996	1	1

Tabla 4.16: Importancia de las componentes principales en Austria.

Usando el gráfico de sedimentación de la Figura 4.29 como apoyo, se observa que el punto de inflexión o “codo” se encuentra en el autovalor de la segunda componente, dando a entender que la elección de dos componentes para Austria es suficiente para explicar de forma correcta las variables.

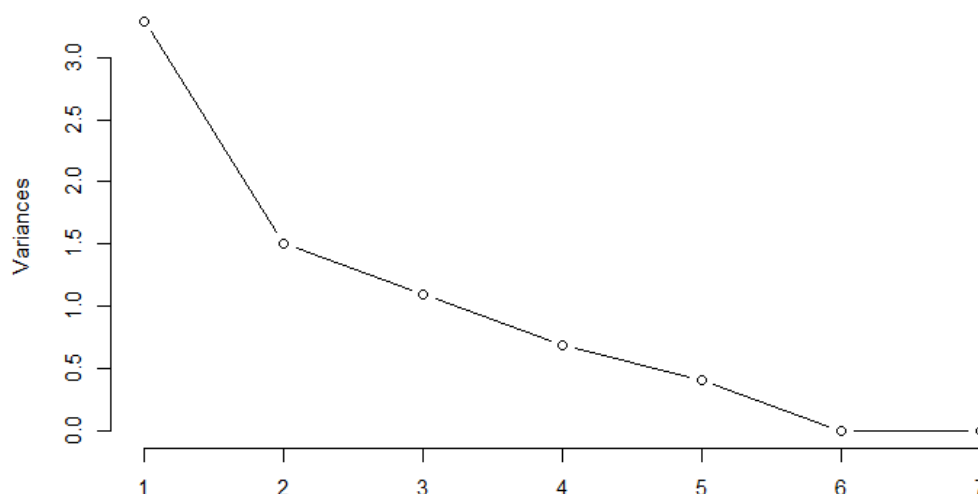


Figura 4.29: Scree plot de Austria.

A través de las cargas de las variables, vemos cómo conforman las variables cada componente principal. Las variables de partículas, como ya se comentó previamente, parecen comportarse de manera similar al estar tan próximas en el biplot y cargas similares.

La primera componente principal separa las variables en los dos grupos mencionados antes. CO₂, partículas y volátiles, al tener un coeficiente positivo, crecen cuando la primera variable crece. Humedad y temperatura, al ser contrarias, decrecen si CP1 aumenta. Tras la primera relación, la segunda parece distinguir las variables según las estaciones. Como vimos en las densidades y ahora en el biplot, temperatura y humedad mantienen valores más altos en verano, al contrario de volátiles y el dióxido de carbono que crecen cuando hace más frío. Las partículas están muy representadas en el primer eje, esto se comprueba por los valores altos de las cargas para la primera componente. Mismo caso sucede con temperatura y humedad para el segundo eje.

Variable	CP1	CP2
CO ₂	0.359	-0.248
Humedad	-0.17	0.576
PM ₁	0.507	0.239
PM ₁₀	0.508	0.232
PM _{2.5}	0.507	0.239
Temperatura	-0.25	0.553
TVOCs	0.096	-0.365



Figura 4.30: Biplot de Austria.

Tabla 4.17: Cargas de las variables originales en Austria

España

En el caso español añadimos una nueva variable, el formaldehído, que provoca un aumento de la varianza por explicar y por tanto menos porcentaje explicado por cada componente. Aun así, la tercera componente no alcanza el valor 1 como para poder considerar su uso, y con tan solo dos dimensiones la varianza acumulada alcanza un 70% según la Tabla 4.18.

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8
Desviación estándar	2.0585	1.1637	0.984	0.898	0.6534	0.452	0.0494	0.0021
Proporción de varianza	0.5297	0.1693	0.121	0.1008	0.0534	0.0255	0.0003	0
Proporción acumulada	0.5297	0.6989	0.82	0.9208	0.9741	0.9997	1	1

Tabla 4.18: Importancia de las componentes principales en España.

El scree plot también apoya el uso de dos componentes principales ya que el punto de inflexión vuelve a encontrarse en el segundo autovalor.

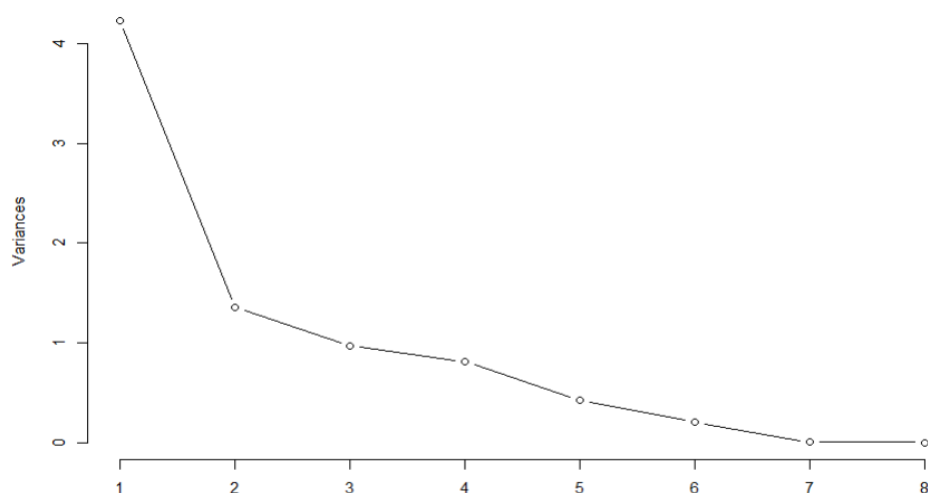


Figura 4.31: Scree plot de España.

En el biplot de España (Figura 4.32), las partículas se mantienen similares, pero ahora volátiles y el CO_2 son más cercanos pudiendo significar comportamientos parecidos. La primera componente vuelve a separar los contaminantes, ahora con influencia negativa, de la temperatura, humedad y formaldehído, con influencia positiva. Ahora parece que es la primera dimensión la que divide las observaciones por estaciones, estando hacia el eje positivo las observaciones de meses más calurosos, y hacia el negativo los meses más fríos. Ahora el segundo eje es una componente de tamaño muy influenciada por el formaldehído y de la que CO_2 no tiene prácticamente representación. Esto implica una tendencia de crecimiento general excepto para CO_2 que parece estar muy representada por el formaldehído.

Variable	CP1	CP2
CO_2	-0.31	0.0002
Humedad	0.217	0.3793
PM_{10}	-0.447	0.2537
$\text{PM}_{2.5}$	-0.448	0.248
Temperatura	0.341	0.4713
TVOCs	-0.323	0.1003
Formaldehído	0.185	0.6631

Tabla 4.19: Cargas de las variables originales en España.

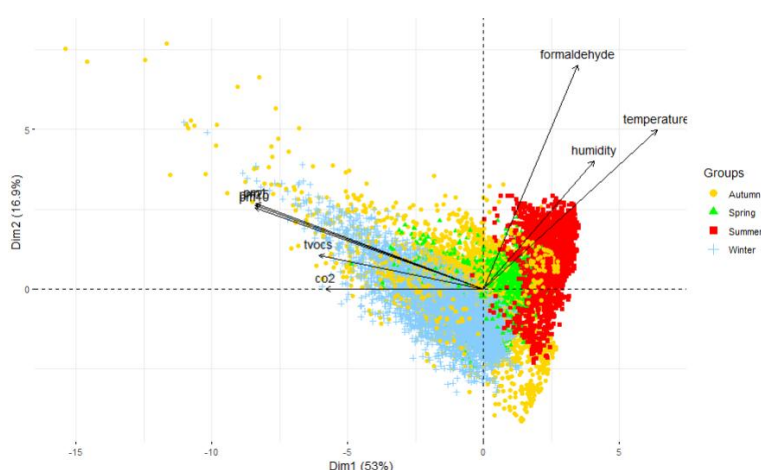


Figura 4.32: Biplot de España.

Polonia

La elección de las componentes se complica en Polonia, ya que según la regla de Kaiser deberíamos elegir las 3 primeras componentes, sin embargo, para los demás países se han elegido únicamente las dos primeras y se explicaba un porcentaje similar que el explicado con dos en este caso, en la Tabla 4.20 aproximadamente un 69%, como el caso español.

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Desviación estándar	1.814	1.2292	1.0468	0.834	0.6382	0.0178	0
Proporción de varianza	0.47	0.2159	0.1565	0.0994	0.0582	0.00005	0
Proporción acumulada	0.47	0.6859	0.8424	0.9418	0.9999	1	1

Tabla 4.20: Importancia de las componentes principales en Polonia.

El punto de inflexión del gráfico del gráfico de sedimentación 4.33 sigue ocurriendo en el segundo autovalor, dando a entender una similitud a Austria y España. Por ello aun sobrepasando el valor 1, usaremos las dos primeras componentes principales y simplificar así las comparaciones entre países.

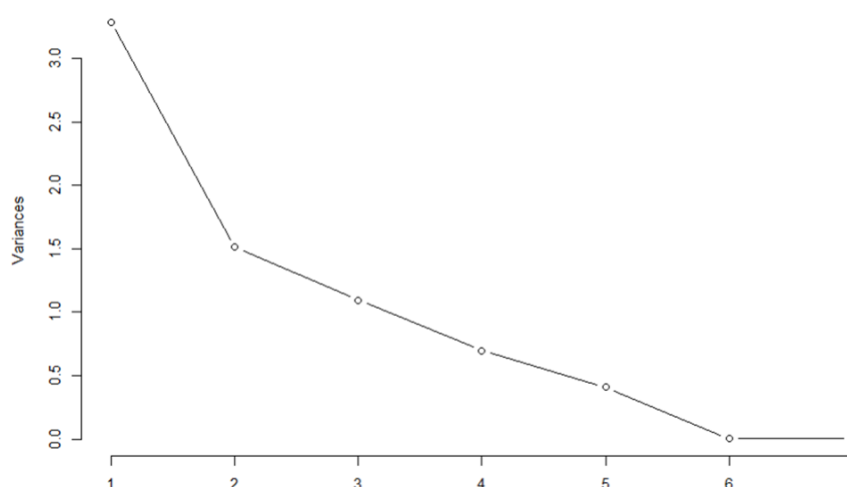


Figura 4.33: Scree plot de Polonia.

En el biplot de España (Figura 4.32), las partículas se mantienen similares, pero ahora volátiles y el CO₂ son más cercanos pudiendo significar comportamientos parecidos. La primera componente vuelve a separar los contaminantes, ahora con influencia negativa, de la temperatura, humedad y formaldehído, con influencia positiva. Ahora parece que es la primera dimensión la que divide las observaciones por estaciones, estando hacia el eje positivo las observaciones de meses más calurosos, y hacia el negativo los meses más fríos. Ahora el segundo eje es una componente de tamaño muy influenciada por el formaldehído y de la que CO₂ no tiene prácticamente representación. Esto implica una tendencia de crecimiento general excepto para CO₂ que parece estar muy representada por el formaldehído.

Al igual que Austria, la segunda variable parece ser la encargada de separar el análisis por las estaciones. Esta primera componente parece querer explicar los outliers presentes en otoño y parte del verano en Polonia, estos valores atípicos están muy influenciados por las partículas dando a entender que son las variables culpables de esta formación, se puede comprobar en los gráficos de densidades de las partículas (Figura 4.3). La diferenciación ahora de los grupos mencionados en los otros dos países ha recaído en la segunda componente, mostrando positividad frente a temperatura y humedad, por el contrario, TVOCs y CO₂ influyen a esta componente de forma negativa. Al representar mucha de la información en la primera componente, esta segunda apenas se ve influenciada por las partículas.

Variable	CP1	CP2
CO ₂	0.1672	-0.568
Humedad	0.1389	0.3792
PM ₁	0.5368	0.1062
PM ₁₀	0.5375	0.0974
PM _{2.5}	0.5368	0.1062
Temperatura	-0.1035	0.6812
TVOCs	0.2777	-0.1932

Tabla 4.21: Cargas de las variables originales en Polonia.



Figura 4.34: Biplot de Polonia.

4.3.2. Datos transformados

Las gráficas y tablas del Análisis de Componentes Principales tras realizar la transformación Box-Cox en CO₂, en partículas (PM₁, PM_{2.5} y PM₁₀) y en TVOCs. Se comprueba si los valores atípicos tienen mucha influencia en las interacciones entre variables, de manera positiva o negativa, y si es recomendable reducir esta influencia.

Austria

La tercera componente sí registra un autovalor mayor a 1 obteniendo casi un 90% de varianza explicada, y con tan solo dos componentes alcanza un porcentaje aproximado del 75%.

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Desviación estándar	1.7554	1.4628	1.0209	0.74294	0.42264	0.07718	0
Proporción de varianza	0.4402	0.3057	0.1489	0.07885	0.02552	0.00085	0
Proporción acumulada	0.4402	0.7459	0.8948	0.97363	0.99915	1	1

Tabla 4.22: Importancia de las componentes principales en Austria con datos transformados.

El codo del gráfico se encuentra en la tercera componente, afirmando que la mejor elección de dimensión final son tres componentes, no dos como resultaba la primera ACP.

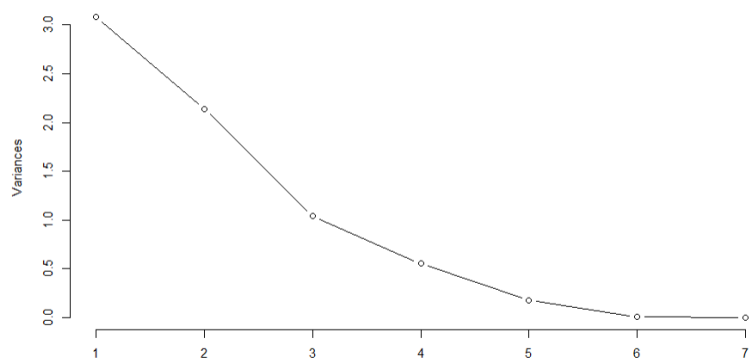


Figura 4.35: Scree plot de Austria con datos transformados.

La primera componente principal, se convierte en una componente de tamaño, indicando un crecimiento general de los datos, además de resumir la variable de partículas casi en su totalidad debido a los bajos coeficientes en las otras dos componentes. La distinción de grupos que en el análisis original correspondía a la primera componente, la ha tomado la segunda mostrando un tercer grupo aislando las partículas de CO₂ y de los volátiles. Las dos últimas componentes también parecen diferenciar juntas a las variables por estaciones, humedad y temperatura para verano, las partículas parecen no diferenciarse por estaciones, y las dos restantes otoño e invierno.

Variable	CP1	CP2	CP3
CO ₂	0.0941	-0.5474	0.3565
Humedad	0.0708	0.4748	0.5351
PM ₁	0.5656	0.0075	-0.1113
PM ₁₀	0.5644	-0.0111	-0.1155
PM _{2.5}	0.5656	0.0075	-0.1113
Temperatura	0.111	0.5721	0.3394
TVOCs	0.1245	-0.3839	0.6582

Tabla 4.23: Cargas de las variables transformadas en Austria.

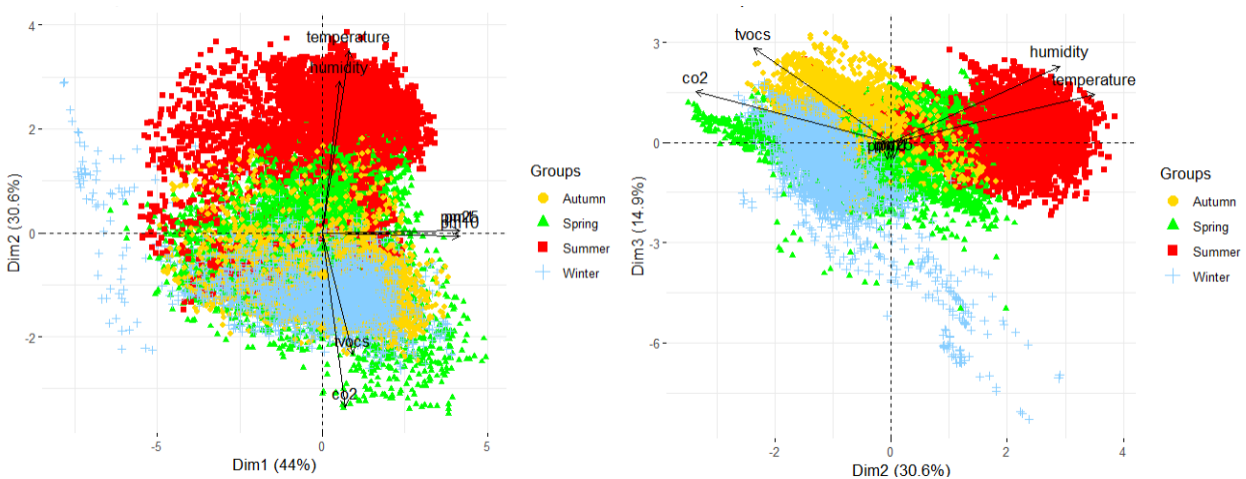


Figura 4.36: Biplots de Austria con datos transformados.

España

El autovalor de la tercera componente en este caso resulta inferior a 1 y la cantidad de información explicada por los dos primeros componentes es del 71%. Mostrando una mejoría frente a los datos originales,

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8
Desviación estándar	1.9024	1.4373	0.9566	0.8545	0.6553	0.4251	0.23439	0.0693
Proporción de varianza	0.4524	0.2582	0.1144	0.0913	0.0537	0.0226	0.0069	0.0006
Proporción acumulada	0.4524	0.7106	0.825	0.9163	0.9699	0.9925	0.9994	1

Tabla 4.24: Importancia de las componentes principales en España con datos transformados.

El scree plot A.3 muestra el punto de inflexión en la tercera componente. Debido a una mayor proporción de varianza acumulada que supera el 82%, se eligen las 3 primeras componentes como dimensión final. Al contrario de los sin la transformación Box-Cox.

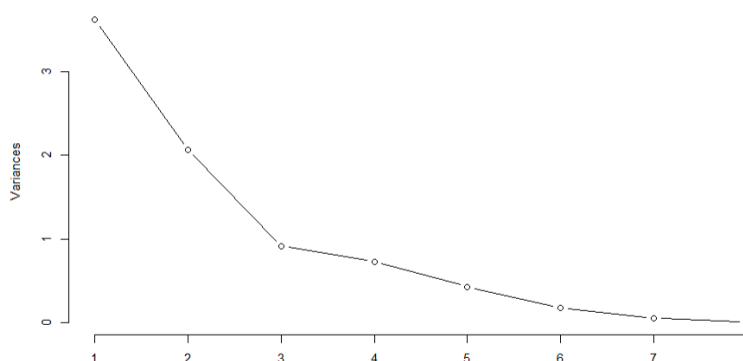


Figura 4.37: Scree plot de España con datos transformados.

La primera componente en España vuelve a tomar el papel de diferenciación entre grupos. Las cargas de la segunda componente parecen distinguir ahora las partículas de los otros dos contaminantes y relacionándolas más con las componentes ambientales. La tercera componente, muy influida por el formaldehído, reduce la participación de las partículas y separa el efecto de la humedad, con una correlación más débil (Tabla 4.2), del efecto de la temperatura y del formaldehído.

Variable	CP1	CP2	CP3
CO ₂	-0.3565	0.1947	-0.3784
Humedad	0.1303	-0.4412	0.2591
PM ₁	-0.4477	-0.3255	0.0978
PM ₁₀	-0.4642	-0.2956	0.1086
PM _{2.5}	-0.4626	-0.3088	0.1064
Temperatura	0.3106	-0.4602	-0.1951
TVOCs	-0.3312	0.2876	-0.4208
Formaldehído	0.1412	-0.4293	-0.7362

Tabla 4.25: Cargas de las variables transformadas en España.

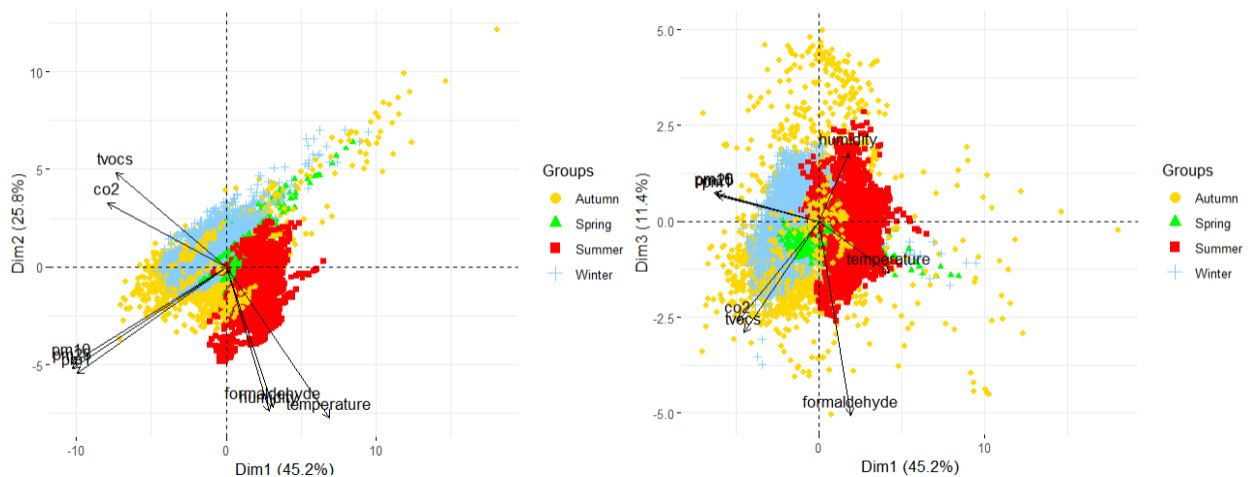


Figura 4.38: Biplots de España con datos transformados.

Polonia

Al igual que con los datos originales la tercera componente su autovalor es mayor que 1, sin embargo, esta toma más importancia al verse incrementado el valor de la componente. La varianza explicada con 3 componentes es ahora de 87% y con dos componentes un 68%. Ha resultado una mayor dimensión con mayor cantidad de varianza explicada, pero que si tomamos dos dimensiones como en los datos originales hay una caída en el porcentaje.

Métrica	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Desviación estándar	1.8381	1.1787	1.1536	0.7736	0.5455	0.0739	0
Proporción de varianza	0.4827	0.1985	0.1901	0.0855	0.0425	0.0008	0
Proporción acumulada	0.4827	0.6811	0.8712	0.9567	0.9992	1	1

Tabla 4.26: Importancia de las componentes principales en Polonia con datos transformados.

Según el scree plot, se toman dos dimensiones, el codo ocurre en el segundo autovalor, pero a partir del tercero hay una clara caída pronunciada. Se decide elegir tres componentes de nuevo para mantener el porcentaje umbral elegido en el resto de países, también debido a que el autovalor supera el valor unidad.

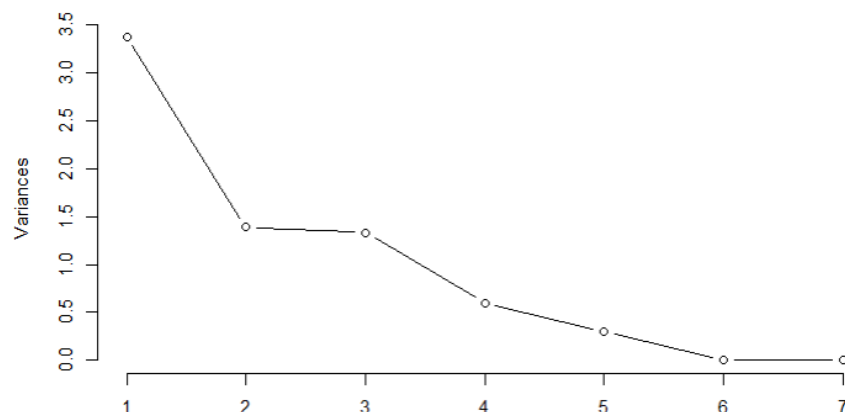


Figura 4.39: Scree plot de Polonia.

Al suavizar outliers, la primera componente ahora diferencia los dos grupos persistentes en los análisis, a diferencia del caso original, aunque con poca representación de la humedad. Sin embargo, la segunda componente ahora separa el CO₂ del resto de contaminantes, dando a entender que, en Polonia, los volátiles están más asociados a las partículas. Si se transforman las variables, ahora Polonia es más similar a los análisis de los otros países. Aunque, ambos ACP polacos coinciden en que la segunda dimensión es la encargada de la distinción estacional.

Variable	CP1	CP2	CP3
CO ₂	0.3	0.3526	-0.5255
Humedad	-0.0182	-0.5875	-0.4258
PM ₁	0.5181	-0.1611	0.2053
PM ₁₀	0.5227	-0.1285	0.194
PM _{2.5}	0.5181	-0.1611	0.2053
Temperatura	-0.2343	-0.6575	0.1271
TVOCs	0.2113	-0.1726	-0.636

Tabla 4.27: Cargas de las variables transformadas en Polonia.

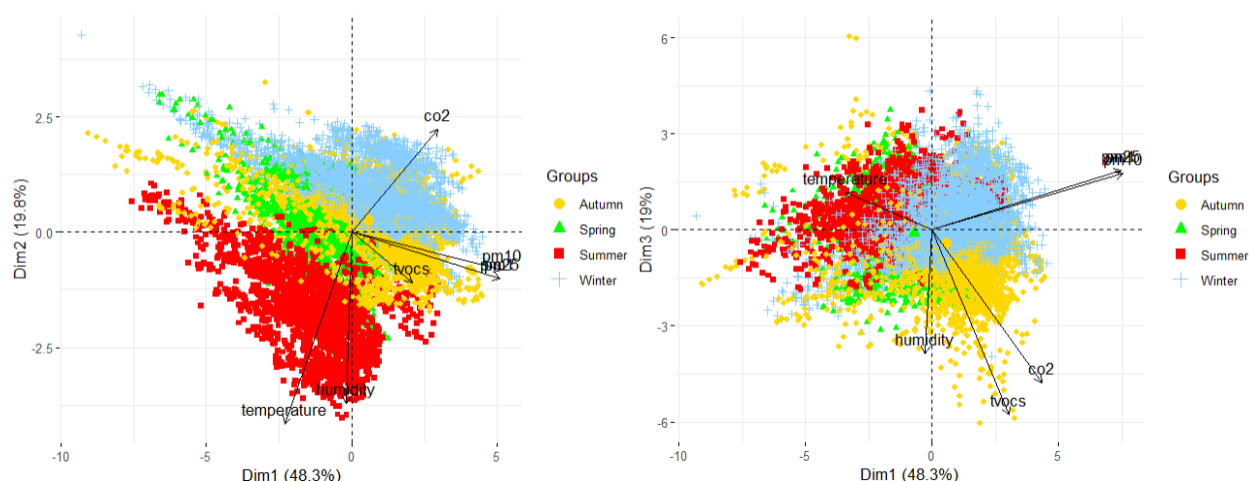


Figura 4.40: Biplots de Polonia con datos transformados.

Cuando se aplica una transformación a las variables para reducir el efecto de los valores atípicos, además de ligeramente incrementar la varianza explicada con dos dimensiones, provoca un aumento significancia de la tercera componente principal. Provocando un aumento de la proporción de varianza acumulada general de aproximadamente un 15% pero elevando en uno el número de dimensiones. La disminución de la influencia de los outliers provoca que las variables de Polonia no estén tan influenciadas y se puedan asemejar sus resultados del análisis con los otros países. En general siempre hay una variable que diferencia las variables según la época del año dando a entender que las relaciones entre variables también sufren de estacionalidad.

4.4. Redes bayesianas

El análisis de las relaciones entre variables categóricas (las variables AQI, season, HOM_smoking, sub_scenário) se llevó a cabo mediante redes bayesianas, las cuales permiten representar las relaciones de dependencia a través de un grafo acíclico. Esta técnica resulta especialmente útil frente al conjunto de datos del estudio caracterizado por múltiples factores implicados y el gran volumen de observaciones, ya que sobresale frente a otros métodos estadísticos que pueden resultar lentos y computacionalmente costosos.

Con el objetivo de simplificar las interacciones y centrar el análisis en relaciones relevantes, se han limitado el número de nodos padres a un máximo de uno y dos por nodo hijo. Esto se aplica en los tres países y en el conjunto de datos donde se unieron las observaciones y se creó una nueva variable que diferencia el país para cada observación (country). Así se permite mejorar la calidad del modelo evitando reducir la capacidad explicativa de las redes. La ejecución de esta técnica estadística en cada país por separado y en conjunto permite identificar patrones particulares en cada país y patrones generales. Las interacciones clave serán usadas para comprobar diferencias entre grupos de las variables numéricas.

Austria

Con un máximo de un nodo padre, las estaciones se comportan como un nodo divergente dando a entender que influyen en las proporciones de los niveles de la calidad de aire tanto del CO₂ como del PM_{2.5}. Este último a su vez es un nodo en secuencia siendo padre de la variable que indica si hay fumadores o no en el hogar, esta relación ya se observó en la Figura 4.7, donde la proporción de observaciones de fumadores iba aumentando a medida que disminuía el nivel de AQI_pm25. Hom_smoking a su vez es un nodo padre de sub_scenário indicando que fumar se asocia con el tipo de calefacción.

Si se vuelve a realizar la red bayesiana pero ahora con 2 nodos padres, season sigue siendo un nodo divergente pero ahora afecta al tipo de calefacción, mostrando que, por características de la distribución de sensores, la estación afecta al número de observaciones por calefacción. Tanto fumar como la calidad del aire según PM_{2.5} son nodos hijos de sub_scenário cuando antes este nodo era convergente, dando a entender que estas relaciones son más sutiles que la relación de season con AQI_pm25 y a su vez AQI_pm25 con presencia de fumador. Mismo caso ocurre con AQI_co2 que se mantiene como nodo convergente.

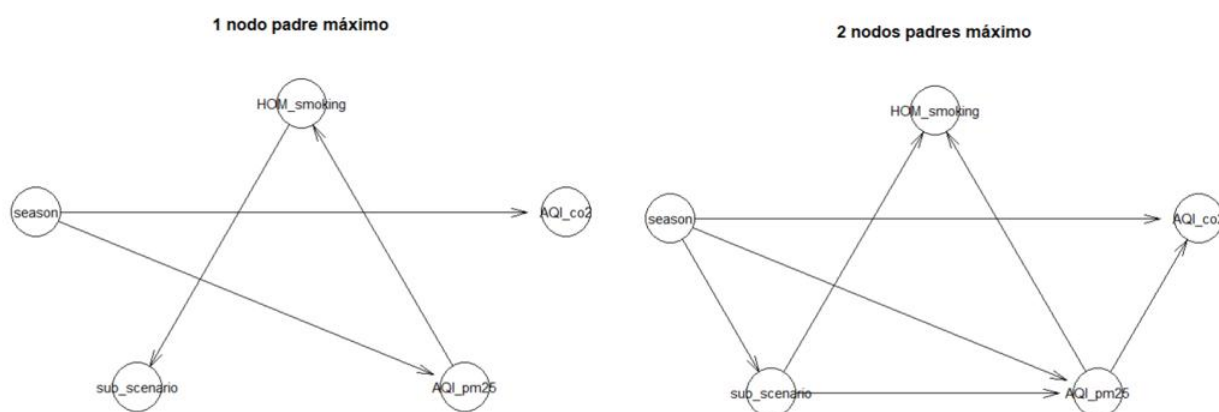


Figura 4.41: Redes bayesianas para Austria.

España

En España las relaciones más directas vuelven a resultar los niveles de la calidad del aire dependientes del nodo padre seasons. Esto resulta evidente dado la estacionalidad de las variables numéricas observada en el análisis descriptivo como en la ACP. La variable HOM_smoking vuelve a estar condicionada por la calidad del aire de las partículas dando a entender una relación entre fumar y la cantidad de PM_{2.5} en el aire.

Cuando se aplican los dos nodos padres, la variable season mantiene su rol de nodo divergente condicionando también HOM_smoking, a diferencia de Austria, mostrando que la estacionalidad afecta también al tabaquismo. En esta estructura más compleja los niveles de formaldehído se ven influidos por la presencia de fumadores mostrando una relación que parecía nula según la Figura 4.8, ya que la proporción de fumadores se mantenía estable a lo largo de los niveles de AQI_formaldehyde. Sin embargo, hay un cambio en la dirección de dependencia entre HOM_smoking y los niveles de partículas, esto puede inferir en una relación de mutua dependencia posiblemente debido a la alta emisión de partículas al fumar.

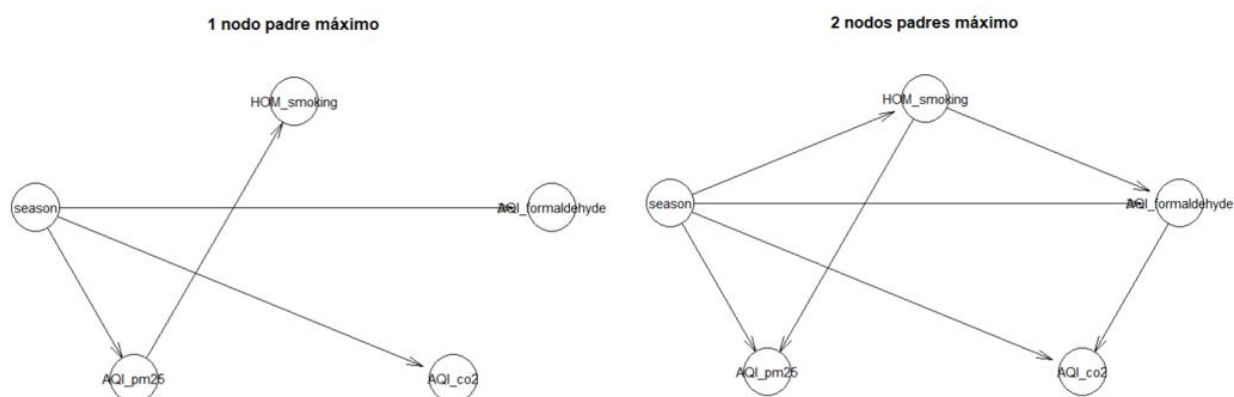


Figura 4.42: Redes bayesianas para España.

Polonia

Al tener menor número de variables, Polonia mantiene una estructura sencilla aún con dos nodos padres. La variable season todavía como nodo raíz influyendo en los niveles de calidad del aire refuerza que en los tres países la estacionalidad es un factor clave en la contaminación de los hogares. Al igual que en Austria, el tipo de calefacción está asociado con el nivel de partículas por lo que es claro que la cantidad de partículas varía según el tipo de calefacción como muestra la Figura 4.9 y las diferencias de alturas de las barras en un mismo tipo de calefacción.

Cuando el límite de dos nodos padres es usado, vuelve a aparecer la diferencia de número de observaciones de las clases de sub_scenariio por cada estación posiblemente debido a la forma de instalación de los sensores. También se muestra una relación entre el AQI del CO₂ con el de PM_{2.5}, mostrando una clara similaridad con la red bayesiana austriaca si la variable relacionada con el tabaco se elimina del grafo.

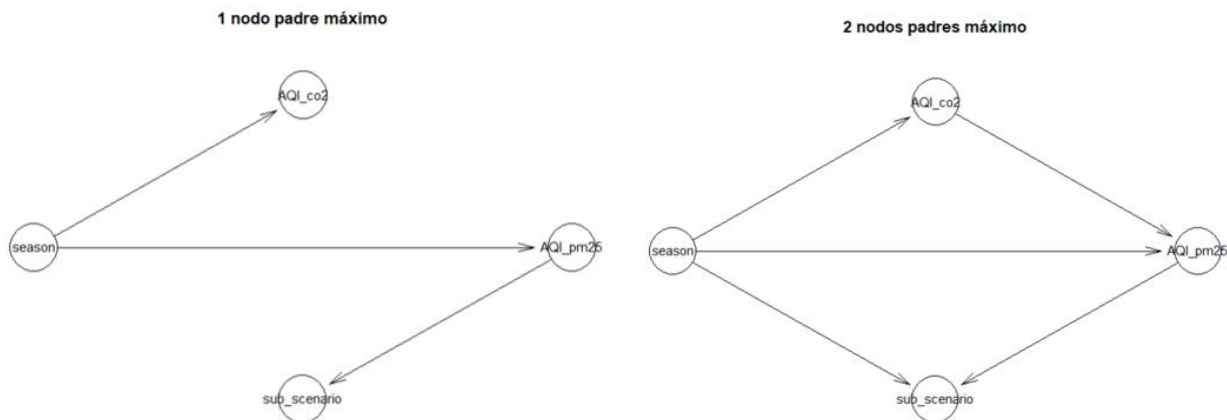


Figura 4.43: Redes bayesianas para Polonia.

Total

En ambas configuraciones la variable que distingue cada país actúa como nodo divergente, es bastante coherente ya que la combinación de datos de los países produce diferencias estructurales. Ambos modelos poseen las mismas relaciones para esta variable, influye directamente sobre HOM_smoking, dando a entender que España y Austria no tienen la misma distribución de fumadores (24 casas de 100 para España y 9 de 29 para Austria). Mismo caso para el tipo de calefacción y el AQI de PM_{2.5}, una variable muy relacionada con el tabaco. La variable season sigue siendo influyente en los casos de contaminantes influyendo a CO₂ mediante las relaciones de dependencia de otras variables.

La ampliación del número de nodos padre, provoca una mayor densidad de la red en la relación de las variables de la calidad del aire según los compuestos. El resto de relaciones se mantienen igual por lo que las interacciones vistas en el primer grafo son las más representativas.

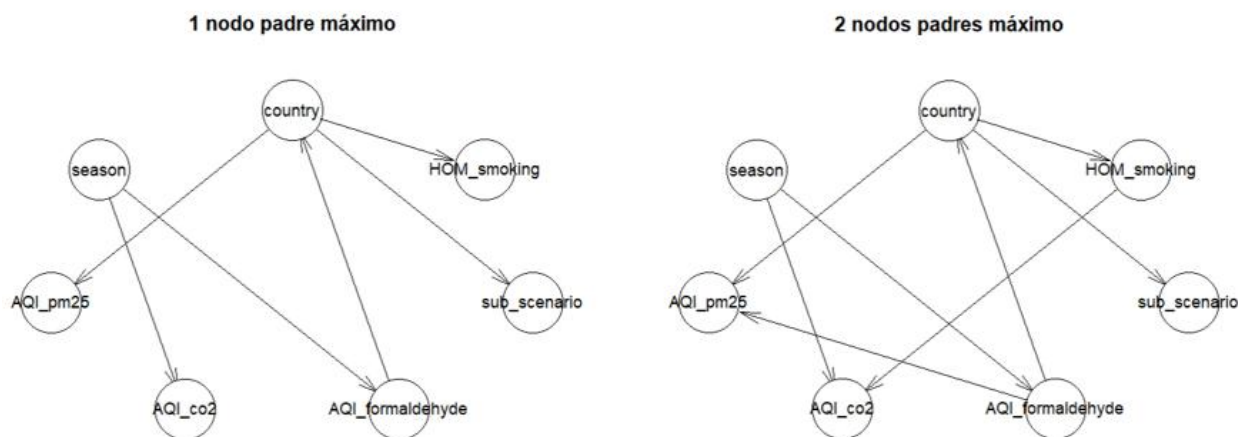


Figura 4.43: Redes bayesianas para el total.

En conjunto, los distintos países como estaciones son un efecto clave en las relaciones de las variables categóricas, dentro de estas variables tienen interacciones únicas entre sí no tan específicas a nivel nacional, es decir, la relación del tabaco con las partículas y el tipo de calefacción se han visto presentes, en Austria una relación de las tres variables, en España tabaco con partículas (al no diferenciar calefacción) y Polonia partículas con calefacción (al no distinguir hogares con presencia de fumadores).

4.5. Pruebas de Kruskal-Wallis y de Dunn

Anteriormente se comprobó que las variables numéricas, salvo la humedad no cumplen con los criterios de normalidad, por ello para comprobar diferencias entre grupos categóricos se requieren pruebas no paramétricas. Por ello se utiliza el test de Kruskal-Wallis, que, a diferencia de ANOVA, usa rangos para comprobar las diferencias entre grupos sin asumir normalidad ni homogeneidad de varianzas. Esta cuadrada (η^2) diferencia como de importante es esta distinción entre grupos, cuanto más cercano a 1 más importante, si se aproxima a 0, las diferencias son mínimas.

Una vez comprobadas las diferencias se aplica el test de Dunn como análisis post-hoc para comprobar en qué grupos y en qué medida difieren las variables numéricas. Al usar rangos, se consigue comprobar que un grupo es mayor que otro evitando mucha influencia de los valores atípicos. Esta diferencia entre grupos se comprueba con el signo del estadístico, si es positivo, los valores del primer grupo son mayoritariamente más grandes que los del segundo, si es negativo, viceversa. El p-valor y la r de efecto indicarán si esta diferencia es significativa o no. Al tener un gran número de datos el test se vuelve computacionalmente costoso, y debido al tamaño del n se utiliza el promedio por hora dentro de cada grupo/relación de grupos relevantes.

4.5.1. *Kruskal-Wallis*

En la tabla 4.28 están resumidos los resultados de la prueba de Kruskal-Wallis, para cada combinación entre variables categóricas y numéricas muestra el p-valor del test, así como eta cuadrada. Estadísticamente todos los factores causan diferencias en las variables numéricas, sin embargo, observando la importancia de ese efecto comprobamos que no todas las interacciones resultan relevantes. Estacionalmente hay diferencias tanto significativas como relevantes, el menor valor de η es de aproximadamente 0.2 en las partículas, representando una gran diferencia de los compuestos según la estación.

La variable country, sin embargo, aunque muestre diferencias significativas el tamaño del efecto de estas es muy bajo, como se observó en análisis anteriores, los resultados entre países rara vez diferían. En contraste, HOM_smoking si tiene un efecto notable frente a las partículas y al formaldehído, demostrando que fumar produce partículas. Mismo caso de country sucede con el tipo de calefacción, sorprende el poco efecto que este produce en el PM_{2.5} cuando se comprobó que en las redes bayesianas AQI_pm25 y sub_scenario estaban altamente relacionadas.

Los índices de calidad del aire (AQI) muestran alta asociación únicamente con sus respectivos contaminantes. Para CO₂ y PM_{2.5} sí aparece un poco de efecto frente a volátiles mostrando la correlación inicial que obtenían las variables numéricas. Este tipo de análisis demuestra que no todos los factores influyen significativamente y que grandes cantidades de datos provocan que variaciones pequeñas se vuelvan influyentes.

Variables categóricas	Medida	CO₂	Humedad	PM_{2.5}	Temperatura	TVOCs	Formaldehído
season	p-valor	0	0	0	0	0	0
	η^2	0.420	0.500	0.194	0.596	0.320	0.359
country	p-valor	0	0	0	0	0	-
	η^2	0.06	0.02	0.019	0.006	0.055	-
HOM smoking	p-valor	0	0	0	0	0	0
	η^2	0.006	0.019	0.329	0.021	0.032	0.246
sub_scenariio	p-valor	0	0	0	0	0	-
	η^2	0.012	0.009	0.007	0.069	0.034	-
AQI_co2	p-valor	0	0	0	0	0	0
	η^2	0.883	0.174	0.018	0.005	0.303	0.045
AQI_pm25	p-valor	0	0	0	0	0	0
	η^2	0.140	0.025	0.889	0.013	0.114	0.009
AQI formaldehído	p-valor	0	0	0	0	0	0
	η^2	0.069	0	0.077	0.062	0.065	0.883

Tabla 4.28: Resumen de las pruebas de Kruskal-Wallis

4.5.2. Dunn

Se tratan las diferencias significativas con grandes efectos mediante el test de Dunn para tener una mayor precisión de cómo afectan los grupos de las variables categóricas. El valor de la r de efecto indica tanto si tomaremos la diferencia como significativa y en qué niveles.

La primera variable categórica en la que se aplica el test es season, siendo esta la que mayor significancia contaba en el test de Kuskal-Wallis. Todas las variables muestran tener el efecto más significativo en verano frente a invierno y otoño y en menos nivel con primavera demostrando que el calor provoca un efecto notorio en los comportamientos de los hogares, posiblemente debido a una mayor ventilación. Destacan las bajas r de efecto en PM_{2.5} apoyando ser el valor de la eta cuadrado de este contaminante como la más baja.

Comparando los valores del estadístico cuando se enfrenta una estación cualquiera con verano, todos los estadísticos tienen signo negativo en los contaminantes demostrando que verano es la época del año con mejor calidad del aire, a contrario, invierno representa la estación con una menor calidad, siendo las r de efecto entre estas las mayores indicando que invierno y verano son las estaciones más dispares.

Grupos	Medida	CO₂	Humedad	PM_{2.5}	Temperatura	TVOCs	Formaldehído
Primavera	Z	-16.84	38.121	-8.348	47.025	-42.717	36.391
vs	p-valor	0	0	0	0	0	0
Invierno	r	-0.128	0.289	-0.063	0.357	-0.324	0.329
Primavera	Z	54.12	-53.802	39.602	-52.097	25.657	-23.661
vs	p-valor	0	0	0	0	0	0
Verano	r	0.41	-0.408	0.301	-0.395	0.195	-0.214
Primavera	Z	22.701	18.669	12.187	-18.109	28.96	-1.173
vs	p-valor	0	0	0	0	0	0.1204
Otoño	r	0.172	0.142	0.092	-0.137	0.22	-0.011

Otoño	Z	5.806	56.111	3.801	28.55	-13.568	41.096
vs	p-valor	0	0	0.0001	0	0	0
Invierno	r	0.044	0.426	0.029	0.217	-0.103	0.372
Otoño	Z	76.175	-34.434	51.31	-69.578	54.335	-27.756
Vs	p-valor	0	0	0	0	0	0
Verano	r	0.578	-0.261	0.389	-0.528	0.412	-0.251
Verano	Z	-70.345	91.34	-47.494	98.571	-68.129	63.409
vs	p-valor	0	0	0	0	0	0
Invierno	r	-0.534	0.693	-0.36	0.748	-0.517	0.574

Tabla 4.29: Resumen del test de Dunn sobre season.

Hay una clara diferencia en partículas entre hogares con presencia de fumador y los hogares con ausencia de este. El valor de la r de efecto y por tanto del estadístico es muy grande y con signo negativo, por lo que se puede decir con confianza que fumar provoca una subida notable en las partículas. Al formaldehído le sucede lo contrario, el valor de este cuando el hábito del tabaquismo es ausente crece en grandes medidas, esto resulta especialmente interesante ya que uno de los subproductos de fumar es el formaldehído, sin embargo el acto de ventilar tras fumar puede tanto evitar el aumento de este producto como la disminución de este.

Grupos	Medida	PM _{2.5}	Formaldehído
No fumador	Z	-123.677	77.080
vs	p-valor	0	0
Fumador	r	-0.574	0.4965

Tabla 4.30: Resumen del test de Dunn sobre HOM_smoking.

Como era de esperar, en la Tabla 4.31, el nivel de CO₂ va aumentando según empeora la calidad del aire, mismo caso ocurre con la humedad y los compuestos volátiles, aunque, para el último caso, el efecto de la diferencia cuando la calidad del aire es mala frente a moderada parece ser nulo, por lo que niveles bajos de volátiles está asociado a buena calidad del aire, los niveles altos pueden indicar tanto moderada como mala calidad.

El caso de la humedad es más curioso, debido a que originalmente este compuesto y el CO₂ obtenían correlaciones negativas, hay que resaltar que el único efecto más o menos significativo sucede cuando la buena calidad del aire es enfrentada con la perjudicial. Esta relación puede referirse a que cuando hay altas concentraciones de CO₂ la humedad es mayor debido a la poca ventilación, pero cuando el contaminante sube, la humedad baja, sin embargo, esta puede reducirse desde valores muy altos en un principio y por ello provocar esta relación tan peculiar.

Grupos	Medida	CO₂	Humedad	TVOCs
Buena	Z	-116.648	-63.341	-98.613
vs	p-valor	0	0	0
Moderada	r	-0.539	-0.288	-0.451
Buena	Z	-201.609	-87.878	-106.537
vs	p-valor	0	0	0
Perjudicial	r	-0.931	-0.4	-0.487
Moderada	Z	-92.816	-28.951	-15.016
vs	p-valor	0	0	0
Mala	r	-0.429	-0.132	-0.069

Tabla 4.31: Resumen del test de Dunn sobre AQI_co2.

En la calidad del aire según partículas los tres contaminantes parecen diferenciarse entre grupos en la misma proporción, siendo el efecto entre buena calidad y mala calidad, el doble del efecto de las otras dos diferencias, Recalcando que para CO₂ y volátiles el r entre grupos es muy similar.

Grupos	Medida	CO₂	PM_{2.5}	TVOCs
Buena	Z	-42.574	-108.47	-35.297
vs	p-valor	0	0	0
Moderada	r	-0.186	-0.482	-0.157
Buena	Z	-85.876	-212.05	-75.963
vs	p-valor	0	0	0
Perjudicial	r	-0.374	-0.942	-0.337
Moderada	Z	-43.602	-104.372	-40.9
vs	p-valor	0	0	0
Mala	r	-0.190	-0.464	-0.182

Tabla 4.32: Resumen del test de Dunn sobre AQI_pm25.

Capítulo 5

Conclusiones

Todos los análisis y sus resultados han guiado a una serie de características del conjunto de datos relevantes para la comprensión de la calidad del aire en entornos domésticos:

El estudio ha demostrado que, pese a algunas diferencias entre países, existe una estructura común en el comportamiento de los parámetros ambientales. Las matrices de correlación y los gráficos de densidad, reflejan patrones similares, reforzados por los resultados del Análisis de Componentes Principales. Al disminuir la influencia de valores extremos, en todos los países se observa una estructura similar, componentes que separan las variables con características similares instando que el comportamiento de estas entre países es similar, y otra componente que distingue las observaciones según la estación del año.

Continuando con la estacionalidad, es observable el efecto de la época del año en todos los análisis presentes, demostrando una diferencia significativa de las variables. Las estaciones frías (invierno y otoño) están más asociadas a un empeoramiento de la calidad del aire, principalmente porque hay una mayor acumulación de contaminantes y posiblemente menor ventilación. En cambio, primavera y verano, caracterizadas por un aumento de la temperatura y humedad interior, tienen una mayor calidad del aire, destacando el caso del verano.

Se ha identificado una relación fuerte y persistente entre los principales contaminantes: CO₂, PM_{2.5} y TVOCs. A lo largo de los análisis realizados, estas variables aparecen fuertemente correlacionadas sugiriendo compartir fuentes de emisión comunes. Las partículas en especial, muestran cierta independencia respecto a los otros dos contaminantes en determinados contextos, posiblemente debido a su mayor sensibilidad al tabaquismo, como evidencian los valores más altos en hogares con presencia de fumadores.

La falta de normalidad en la gran parte de las variables estudiadas, indica la necesidad de emplear métodos estadísticos no paramétricos en este tipo de estudios ambientales. Esto es necesario aun incluyendo la posibilidad de la transformación Box-Cox, específicamente en los contaminantes debidos a las colas provocadas por outliers, aunque esta transformación ha permitido reducir el impacto de estos valores y mostrar relaciones ocultas. Con la premisa de no normalidad, estudios como ANOVA, modelos lineales o algunos análisis multivariantes no son de utilidad al no cumplir este supuesto.

Estudiar las relaciones de las variables categóricas ha permitido identificar relaciones de dependencia entre ellas. Se ha comprobado como condicionan la calidad del aire interior por el valor que toman, aspectos como la estación del año o la presencia de fumadores afectan directamente en el nivel de contaminación en los hogares. Esta visión conjunta permite facilitar la visualización de cadenas causa-efecto potenciales entre variables que pueden ser útiles para el diseño de estrategias de intervención.

5.1 Trabajos futuros

Este estudio ofrece múltiples posibilidades de continuar la investigación en el campo de la calidad del aire en interiores:

- **Desarrollo de protocolos de monitoreo y actuación en hogares:** Establecer sistemas de control en tiempo real, que, basados en los datos previos, generen respuestas automáticas de alerta como recomendaciones de ventilación o purificación del aire.
- **Modelos de predicción del hábito de fumar en interiores:** Usando algoritmos de machine learning sobre los parámetros ambientales, poder obtener un análisis predictivo capaz de inferir la posibilidad de presencia de fumadores en un hogar. Útil en estudios epidemiólogos o auditorías de salud pública. Un trabajo en el que estoy colaborando.
- **Series temporales para la predicción de parámetros ambientales:** Inferir más en el tema temporal y estacional con el objetivo de modelar la evolución de contaminantes. Esto permitirá capturar las dinámicas estacionales o patrones horarios mediante modelos ARIMA o modelos de suavizado exponencial permitiendo anticipar episodios de mala calidad del aire y optimizar así los protocolos de actuación
- **Ampliación geográfica y social del estudio:** Complementar este estudio con más países y comprobar si, como en este caso, comparten resultados. También incluir la posibilidad de diferenciar niveles socioeconómicos y zonas rurales y urbanas para comprobar si la calidad del aire es un aspecto influido por estos factores.

Capítulo 6

Bibliografía

- [1] K-HealthinAir. (2025, mayo 28). *Proyecto europeo K-HealthinAir*. <https://k-healthinair.eu/>
- [2] CARTIF. (2025, mayo 28). *K-HealthinAir*. <https://www.cartif.es/k-healthinair/>
- [3] Ensaco. (2025, mayo 28). *Importancia de la calidad del aire*. <https://www.ensaco.es/importancia-calidad-aire>
- [4] Organización Mundial de la Salud (OMS). (2025, mayo 29). *Contaminación del aire en los hogares y salud*. <https://www.who.int/es/news-room/fact-sheets/detail/household-air-pollution-and-health>
- [5] R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- [6] Posit, PBC. (2025, junio 2). *Descargas para RStudio y Posit*. <https://posit.co/downloads/>
- [7] Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.
[CRAN: <https://cran.r-project.org/web/packages/ggplot2/index.html>]
- [8] Pedersen, T. L. (2020). *patchwork: The composer of plots*.
[CRAN: <https://cran.r-project.org/web/packages/patchwork/index.html>]
- [9] Komsta, L., & Novomestky, F. (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*.
[CRAN: <https://cran.r-project.org/web/packages/moments/index.html>]
- [10] Wei, T., & Simko, V. (2021). *corrplot: Visualization of a correlation matrix*.
[CRAN: <https://cran.r-project.org/web/packages/corrplot/index.html>]
- [11] Datatab. (2025, mayo 31). *Correlación de Pearson*.
<https://datatab.es/tutorial/pearson-correlation>
- [12] Datatab. (2025, mayo 31). *Correlación de Spearman*.
<https://datatab.es/tutorial/spearman-correlation>
- [13] Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–252.
- [14] Fox, J., & Weisberg, S. (2019). *car: Companion to applied regression*.
[CRAN: <https://cran.r-project.org/web/packages/car/index.html>]
- [15] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611.
- [16] R Core Team. (2024). *stats: The R Stats Package*.
[Base R: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html>]

- [17] Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399–402.
- [18] Gross, J., & Ligges, U. (2015). *nortest: Tests for normality*.
[CRAN: <https://cran.r-project.org/web/packages/nortest/index.html>]
- [19] IBM. (2025, junio 3). *Principal Component Analysis (PCA)*. <https://www.ibm.com/es-es/think/topics/principal-component-analysis>
- [20] Fernández, M. A. (2024). *Análisis en componentes principales*. Universidad de Valladolid.
- [21] INAOE. (2025, junio 3). *Redes Navesianas*.
<https://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf>
- [22] Scutari, M. (2025). *bnlearn: Bayesian network structure learning, parameter learning and inference*. [CRAN: <https://cran.r-project.org/web/packages/bnlearn/index.html>]
- [23] Datatab. (2025, junio 3). *Kruskal-Wallis Test*.
<https://datatab.es/tutorial/kruskal-wallis-test>
- [24] Dinno, A. (2024). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*.
[CRAN: <https://cran.r-project.org/web/packages/dunn.test/index.html>]
- [25] StackExchange. (2025, junio 3). *Dunn Test*.
<https://stats.stackexchange.com/tags/dunn-test/info>
- [26] inBiot. (2025, mayo 30). *Monitorización y Mejora de la Calidad del Aire Interior*.
<https://www.inbiot.es/es>
- [27] Kaiterra. (2025, mayo 30). *Smart Air Quality Monitor for Commercial Use*.
<https://www.kaiterra.com>
- [28] Barret, T. (2025). *data.table: Extension of 'data.frame'*.
[CRAN: <https://cran.r-project.org/web/packages/data.table/index.html>]
- [29] Vitalie, S. (2024). *lubridate: Make Dealing with Dates a Little Easier*.
[CRAN: <https://cran.r-project.org/web/packages/lubridate/index.html>]
- [30] Wickham, H. (2025). *readxl: Read Excel Files*.
[CRAN: <https://cran.r-project.org/web/packages/readxl/index.html>]
- [31] Wickham, H. (2023). *dplyr: A Grammar of Data Manipulation*.
[CRAN: <https://cran.r-project.org/web/packages/dplyr/index.html>]
- [32] GOAQS. (2025, mayo 30). <https://goaqs.org/>
- [33] EPA. (2025, junio 5). *Particulate Matter (PM) Basics*.
<https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>

Anexo

Apéndice A

Contiene el código de R utilizado para elaborar el estudio.

A.1. Importar y preprocesar los datos

```
#####  
##### IMPORTAR DATOS #####  
#####  
library(data.table)  
  
files <- list.files(path = "DATOS", pattern = "\\*.csv$", full.names =  
TRUE)  
  
# Crear listas vacías para almacenar los indoor y outdoor por grupo  
(ATHOM, PLHOM, ESHOM)  
indoor_list <- list()  
  
# Lista para almacenar las columnas de cada conjunto de datos  
all_columns_in <- character()  
  
# Obtener los objetos y organizarlos por nombre de grupo  
for (file in files) {  
  # Importar archivo como data.table  
  data <- fread(file)  
  
  # Obtener el nombre del archivo (sin la extensión y con guion bajo)  
  name <- gsub(".csv$", "", basename(file))  
  name <- gsub("-", "_", name)  
  colnames(data)[1] <- "X_time"  
  # Obtener las variables de outdoor  
  var_out <- grep("out", colnames(data), value = TRUE)  
  var_out <- which(colnames(data) %in% var_out)  
  
  # Separar observaciones indoor  
  indoor <- data[!apply(is.na(data[, !c(1, 2, 3, 4, var_out)], with =  
FALSE)], 1, function(x) all(x)), !var_out, with = FALSE]  
  
  # Actualizar lista de columnas para todos los datasets  
  all_columns_in <- unique(c(all_columns_in, colnames(indoor)))  
  
  # Asignar a las listas indoor y outdoor por nombre base (ATHOM,  
PLHOM, ESHOM)  
  base_name <- gsub("_.*", "", name)
```

```

    if (grepl("ATHOM", base_name)) {
      indoor_list[["ATHOM"]] <- append(indoor_list[["ATHOM"]],
list(indoor))
    } else if (grepl("PLHOM", base_name)) {
      indoor_list[["PLHOM"]] <- append(indoor_list[["PLHOM"]],
list(indoor))
    } else if (grepl("ESHOM", base_name)) {
      indoor_list[["ESHOM"]] <- append(indoor_list[["ESHOM"]],
list(indoor))
    }
  }
}

# Asegurar que todas las columnas estén presentes, agregando NAs si
faltan
ensure_columns <- function(df, all_columns) {
  missing_cols <- setdiff(all_columns, colnames(df))
  for (col in missing_cols) {
    df[[col]] <- NA
  }
  return(df[, ..all_columns]) # Reorganizar las columnas en el mismo
orden que all_columns
}

# Asegurar de que todas las columnas estén presentes antes de hacer
rbind
ATHOM_in <- rbindlist(lapply(indoor_list[["ATHOM"]], ensure_columns,
all_columns = all_columns_in), use.names = TRUE, fill = TRUE)
ESHOM_in <- rbindlist(lapply(indoor_list[["ESHOM"]], ensure_columns,
all_columns = all_columns_in), use.names = TRUE, fill = TRUE)
PLHOM_in <- rbindlist(lapply(indoor_list[["PLHOM"]], ensure_columns,
all_columns = all_columns_in), use.names = TRUE, fill = TRUE)

# Vaciar espacio
remove(files, indoor, outdoor, data, indoor_list, outdoor_list,
all_columns_in, all_columns_out, base_name, name, file)

ATHOM_in$X_time<-as.POSIXct(ATHOM_in$X_time, format = "%Y-%m-%d
%H:%M")
PLHOM_in$X_time<-as.POSIXct(PLHOM_in$X_time, format = "%Y-%m-%d
%H:%M")
ESHOM_in$X_time<-as.POSIXct(ESHOM_in$X_time, format = "%Y-%m-%d
%H:%M")

#####
##### DIFERENCIAR ESTACIONES #####
#####
library(lubridate)

```

```

# Para ATHOM_in
ATHOM_in[, season := fifelse(X_time >= as.Date(paste0(year(X_time), "-03-20")) & X_time < as.Date(paste0(year(X_time), "-06-21")), "Spring",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-06-21")) & X_time <
as.Date(paste0(year(X_time), "-09-23")), "Summer",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-09-23")) & X_time <
as.Date(paste0(year(X_time), "-12-21")), "Autumn", "Winter")))]
ATHOM_in <- ATHOM_in[, c(1, 13, 2:12)] # Reordenar las columna

# Para ESHOM_in
ESHOM_in[, season := fifelse(X_time >= as.Date(paste0(year(X_time), "-03-20")) & X_time < as.Date(paste0(year(X_time), "-06-21")), "Spring",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-06-21")) & X_time <
as.Date(paste0(year(X_time), "-09-23")), "Summer",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-09-23")) & X_time <
as.Date(paste0(year(X_time), "-12-21")), "Autumn", "Winter")))]
ESHOM_in <- ESHOM_in[, c(1, 13, 2:12)] # Reordenar las columnas

# Para PLHOM_in
PLHOM_in[, season := fifelse(X_time >= as.Date(paste0(year(X_time), "-03-20")) & X_time < as.Date(paste0(year(X_time), "-06-21")), "Spring",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-06-21")) & X_time <
as.Date(paste0(year(X_time), "-09-23")), "Summer",
                             fifelse(X_time >=
as.Date(paste0(year(X_time), "-09-23")) & X_time <
as.Date(paste0(year(X_time), "-12-21")), "Autumn", "Winter")))]
PLHOM_in <- PLHOM_in[, c(1, 13, 2:12)] # Reordenar las columnas

#####
##### FUMADORES #####
#####
library(readxl)
IAQ_summary <- read_excel("IAQ_summary.xlsx")
IAQ_summary$area <- substr(IAQ_summary$ID, 3, 10)
IAQ_summary$HOM_smoking <- ifelse(IAQ_summary$smoking_status=="YES" |
IAQ_summary$passive_smoking=="YES", "YES", "NO")

ESHOM_in <- merge(ESHOM_in, IAQ_summary[, 28:29], by="area")

AT_smokers <- read_excel("HOMES_Smokers.xlsx")
ATHOM_in[, HOM_smoking := fifelse(area %in% substr(AT_smokers$`Homes-
CODES`, 3, 10), "YES", "NO")]

```

```
#####
##### UNIFICAR Y PROMEDIAR DATOS #####
#####
library(dplyr)
# Unificar datos totales
HOM_in <- bind_rows(ATHOM_in, ESHOM_in, PLHOM_in)
HOM_in$country <- c(rep("Austria", nrow(ATHOM_in)),
rep("Spain", nrow(ESHOM_in)), rep("Poland", nrow(PLHOM_in)))
m_in <- HOM_in[, lapply(.SD, mean, na.rm = TRUE), by =
.(X_time=format(X_time, "%Y-%m-%d %H"), season), .SDcols = 6:13]
m_in$X_time <- as.POSIXct(m_in$X_time, format="%Y-%m-%d %H")

# Unificar datos por hora Austria
ATHOM_in<-ATHOM_in[,-13] #Eliminar formaldehido vacío
m_AT_in <- ATHOM_in[, lapply(.SD, mean, na.rm = TRUE), by =
.(X_time=format(X_time, "%Y-%m-%d %H"), season), .SDcols = 6:12]
m_AT_in$X_time<-as.POSIXct(m_AT_in$X_time, format="%Y-%m-%d %H")

# Unificar datos por hora Espana
m_ES_in <- ESHOM_in[, lapply(.SD, mean, na.rm = TRUE), by =
.(X_time=format(X_time, "%Y-%m-%d %H"), season), .SDcols = 6:13]
m_ES_in$X_time<-as.POSIXct(m_ES_in$X_time, format="%Y-%m-%d %H")

# Unificar datos por hora Polonia
PLHOM_in<-PLHOM_in[,-13]
m_PL_in <- PLHOM_in[, lapply(.SD, mean, na.rm = TRUE), by =
.(X_time=format(X_time, "%Y-%m-%d %H"), season), .SDcols = 6:12]
m_PL_in$X_time<-as.POSIXct(m_PL_in$X_time, format="%Y-%m-%d %H")

#####
##### GOAQs #####
#####
# Bucle para aplicar AQI en cada archivo
for (nombre in ls(pattern = "_in$")) {
  df <- get(nombre)
  r_co2 <- round(df$co2)
  r_pm25 <- round(df$pm25)

  # AQI para PM2.5
  df$AQI_pm25 <- ifelse(r_pm25 <= 8,
                        "GOOD", ifelse(r_pm25 > 8 & r_pm25 <= 35,
                        "MODERATE", "UNHEALTHY"))
}
```

```

# AQI para CO2
df$AQI_co2 <- ifelse(r_co2 <= 850,
                    "GOOD", ifelse(r_co2 > 800 & r_co2 <= 1400,
                                    "MODERATE", "UNHEALTHY"))

assign(nombre, df) # Reasignar el objeto modificado
}

for(nombre in c("ESHOM_in", "m_ES_in", "HOM_in", "m_in")){
  df <- get(nombre)
  # AQI del formaldehido
  formaldehyde <- round(df$formaldehyde)
  # Para pasar de ppb a microgramos/m^3 dividimos entre 1.228
  df$AQI_formaldehyde <- ifelse(formaldehyde <= 50/1.228,
                                "GOOD", ifelse(formaldehyde >
50/1.228 & formaldehyde <= 120/1.228,
                                                "MODERATE",
"UNHEALTHY"))
  assign(nombre, df) # Reasignar el objeto modificado
}
rm(df, r_co2, r_pm25, nombre, formaldehyde)

#####
##### CARACTERISTICAS DE LOS DATOS #####
#####
par(mar = c(8, 6, 4, 2))
fechas_completas <- seq.Date(as.Date("2023-03-01"), as.Date("2025-03-
10"), by = "month")

hist(date(ATHOM_in$X_time), breaks =
length(unique(date(ATHOM_in$X_time)))/5, col = rgb(1, 0, 0, 0.5),
     xlab = "", yaxt = "n", xaxt = "n", ylim=c(0,80000), freq = T, main
= "Cantidad de observaciones para cada pais")
hist(date(ESHOM_in$X_time), add = TRUE, breaks =
length(unique(date(ESHOM_in$X_time)))/5,
     col = rgb(0, 0, 1, 0.45), xaxt = "n", yaxt = "n", freq = T)
hist(date(PLHOM_in$X_time), add = TRUE, breaks =
length(unique(date(PLHOM_in$X_time)))/5,
     col = rgb(0, 1, 0, 0.45), xaxt = "n", yaxt = "n", freq = T)

axis(1, at = fechas_completas, labels = fechas_completas, las = 2)
axis(2, at = seq(0, 80000, by = 10000))
mtext("Date", side = 1, line = 6)
legend("topleft", inset = c(0.15, 0), legend = c("Austria", "España",
"Polonia"),
     fill = c("red", "blue", "green"), bty = "n")

```

```

#Numero de casos incompletos y su porcentaje del total
(nrow(ATHOM_in)-sum(complete.cases(ATHOM_in)))/nrow(ATHOM_in)*100
nrow(ATHOM_in)-sum(complete.cases(ATHOM_in))

(nrow(ESHOM_in)-sum(complete.cases(ESHOM_in)))/nrow(ESHOM_in)*100
nrow(ESHOM_in)-sum(complete.cases(ESHOM_in))

(nrow(PLHOM_in)-sum(complete.cases(PLHOM_in)))/nrow(PLHOM_in)*100
nrow(PLHOM_in)-sum(complete.cases(PLHOM_in))

length(unique(ATHOM_in$area))
length(unique(ESHOM_in$area))
length(unique(PLHOM_in$area))

min(ATHOM_in$X_time);max(ATHOM_in$X_time)
min(ESHOM_in$X_time);max(ESHOM_in$X_time)
min(PLHOM_in$X_time);max(PLHOM_in$X_time)

```

A.2. Estadística descriptiva

```

#####
##### VAR CUANTITATIVAS #####
#####

# CORRELACIONES
library(corrplot)
corrplot(cor(m_AT_in[3:9], use = "pairwise.complete.obs", method =
"spearman"),
         method = "square", addCoef.col = T)

corrplot(cor(m_ES_in[3:10], use = "pairwise.complete.obs", method =
"spearman"),
         method = "square", addCoef.col = T)

corrplot(cor(m_PL_in[3:9], use = "pairwise.complete.obs", method =
"spearman"),
         method = "square", addCoef.col = T)

# Función para generar los histogramas por variable y país
library(ggplot2)
crear_graficos <- function(df, pais) {
  lista_graficos <- list()
  # Variables numericas
  var_cuant <- names(df)[sapply(df, is.numeric)]

```

```

for (var in var_cuant) {
  var_sym <- sym(var)

  p <- ggplot() +
    geom_density(data = subset(df, season == "Winter"), aes(y =
after_stat(count) / sum(after_stat(count)) * 100,
x = !!var_sym, fill = "Winter"), alpha = 0.5,
na.rm = TRUE) +
    geom_density(data = subset(df, season == "Spring"), aes(y =
after_stat(count) / sum(after_stat(count)) * 100,
x = !!var_sym, fill = "Spring"), alpha = 0.5,
na.rm = TRUE) +
    geom_density(data = subset(df, season == "Summer"), aes(y =
after_stat(count) / sum(after_stat(count)) * 100,
x = !!var_sym, fill = "Summer"), alpha = 0.5,
na.rm = TRUE) +
    geom_density(data = subset(df, season == "Autumn"), aes(y =
after_stat(count) / sum(after_stat(count)) * 100,
x = !!var_sym, fill = "Autumn"), alpha = 0.5,
na.rm = TRUE) +
    geom_density(data = df, aes(y = after_stat(count) /
sum(after_stat(count)) * 100,
x = !!var_sym, fill = "Total"), alpha = 0.6,
na.rm = TRUE) +
    theme_minimal()+
    scale_fill_manual(name = "Estación",
values = c("Winter" = "blue", "Spring" = "green", "Summer" =
"red", "Autumn" = "gold", "Total"="black")) +
    ggtitle(paste("Densidades de", var, "en", pais)) + xlab(var) +
ylab("Densidad")

  lista_graficos[[var]] <- p
}

return(lista_graficos)
}

# Crear listas para cada dataset
lista_graficos_AT <- crear_graficos(m_AT_in, "Austria")
lista_graficos_ES <- crear_graficos(m_ES_in, "España")
lista_graficos_PL <- crear_graficos(m_PL_in, "Polonia")
lista_graficos <- crear_graficos(m_in, "el total")

# Imprimir cada grafico y ver estadísticos
library(moments)
library(patchwork)

print(lista_graficos_AT$co2)/print(lista_graficos_ES$co2) |
print(lista_graficos_PL$co2)/print(lista_graficos$co2)
summary(m_AT_in$co2); skewness(m_AT_in$co2, na.rm=T);

```

```

kurtosis(m_AT_in$co2, na.rm=T)
summary(m_ES_in$co2); skewness(m_ES_in$co2, na.rm=T);
kurtosis(m_ES_in$co2, na.rm=T)
summary(m_PL_in$co2); skewness(m_PL_in$co2, na.rm=T);
kurtosis(m_PL_in$co2, na.rm=T)
summary(m_in$co2); skewness(m_in$co2, na.rm=T); kurtosis(m_in$co2,
na.rm=T)

```

```

print(lista_graficos_AT$humidity)/print(lista_graficos_ES$humidity) |
print(lista_graficos_PL$humidity)/print(lista_graficos$humidity)
summary(m_AT_in$humidity); skewness(m_AT_in$humidity, na.rm=T);
kurtosis(m_AT_in$humidity, na.rm=T)
summary(m_ES_in$humidity); skewness(m_ES_in$humidity, na.rm=T);
kurtosis(m_ES_in$humidity, na.rm=T)
summary(m_PL_in$humidity); skewness(m_PL_in$humidity, na.rm=T);
kurtosis(m_PL_in$humidity, na.rm=T)
summary(m_in$humidity); skewness(m_in$humidity, na.rm=T);
kurtosis(m_in$humidity, na.rm=T)

```

```

print(lista_graficos_AT$pm25)/print(lista_graficos_ES$pm25) |
print(lista_graficos_PL$pm25)/print(lista_graficos$pm25)
summary(m_AT_in$pm25); skewness(m_AT_in$pm25, na.rm=T);
kurtosis(m_AT_in$pm25, na.rm=T)
summary(m_ES_in$pm25); skewness(m_ES_in$pm25, na.rm=T);
kurtosis(m_ES_in$pm25, na.rm=T)
summary(m_PL_in$pm25); skewness(m_PL_in$pm25, na.rm=T);
kurtosis(m_PL_in$pm25, na.rm=T)
summary(m_in$pm25); skewness(m_in$pm25, na.rm=T); kurtosis(m_in$pm25,
na.rm=T)

```

```

print(lista_graficos_AT$temperature)/print(lista_graficos_ES$temperatu
re) |
print(lista_graficos_PL$temperature)/print(lista_graficos$temperature)
summary(m_AT_in$temperature); skewness(m_AT_in$temperature, na.rm=T);
kurtosis(m_AT_in$temperature, na.rm=T)
summary(m_ES_in$temperature); skewness(m_ES_in$temperature, na.rm=T);
kurtosis(m_ES_in$temperature, na.rm=T)
summary(m_PL_in$temperature); skewness(m_PL_in$temperature, na.rm=T);
kurtosis(m_PL_in$temperature, na.rm=T)
summary(m_in$temperature); skewness(m_in$temperature, na.rm=T);
kurtosis(m_in$temperature, na.rm=T)

```

```

print(lista_graficos_AT$tvocs)/print(lista_graficos_ES$tvocs) |
print(lista_graficos_PL$tvocs)/print(lista_graficos$tvocs)
summary(m_AT_in$tvocs); skewness(m_AT_in$tvocs, na.rm=T);
kurtosis(m_AT_in$tvocs, na.rm=T)

```

```

summary(m_ES_in$tvocs); skewness(m_ES_in$tvocs, na.rm=T);
kurtosis(m_ES_in$tvocs, na.rm=T)
summary(m_PL_in$tvocs); skewness(m_PL_in$tvocs, na.rm=T);
kurtosis(m_PL_in$tvocs, na.rm=T)
summary(m_in$tvocs); skewness(m_in$tvocs, na.rm=T);
kurtosis(m_in$tvocs, na.rm=T)

print(lista_graficos_ES$formaldehyde)
summary(m_ES_in$formaldehyde); skewness(m_ES_in$formaldehyde,
na.rm=T); kurtosis(m_ES_in$formaldehyde, na.rm=T)

#####
##### VAR CUALITATIVAS #####
#####

# Austria
df <- ATHOM_in %>% filter(complete.cases(.)) %>%
  count(sub_scenario, AQI_co2, HOM_smoking) %>%
  mutate(pct = n / sum(n) * 100)
ggplot(df, aes(x = sub_scenario, y = pct, fill = factor(HOM_smoking,
levels = c("YES", "NO")))) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~ AQI_co2) +
  scale_fill_manual(values = c("NO" = "skyblue2", "YES" = "salmon")) +
  geom_text(aes(label = paste0(round(pct, 1), "%"),
    position = position_stack(vjust = 0.5), color = "black",
size = 4) +
  labs(x = "Tipo de caldera", y = "Porcentaje de observaciones", fill
= "Fumadores", subtitle = "AQI dióxido de carbono") +
  theme_minimal()

df <- ATHOM_in %>% filter(complete.cases(.)) %>%
  count(sub_scenario, AQI_pm25, HOM_smoking) %>%
  mutate(pct = n / sum(n) * 100)
ggplot(df, aes(x = sub_scenario, y = pct, fill = factor(HOM_smoking,
levels = c("YES", "NO")))) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~ AQI_pm25) +
  scale_fill_manual(values = c("NO" = "skyblue2", "YES" = "salmon")) +
  geom_text(aes(label = paste0(round(pct, 1), "%"),
    position = position_stack(vjust = 0.5), color = "white",
size = 4) +
  labs(x = "Tipo de caldera", y = "Porcentaje de observaciones", fill
= "Fumadores", subtitle = "AQI PM2.5") +
  theme_minimal()

```

España

```
df <- ESHOM_in %>% filter(complete.cases(.)) %>%  
  count(AQI_co2, HOM_smoking) %>%  
  mutate(pct = n / sum(n) * 100)  
ggplot(df, aes(x = AQI_co2, y = pct, fill = factor(HOM_smoking, levels  
= c("YES", "NO")))) +  
  geom_bar(stat = "identity", position = "stack") +  
  scale_fill_manual(values = c("NO" = "skyblue2", "YES" = "salmon")) +  
  geom_text(aes(label = paste0(round(pct, 1), "%")),  
            position = position_stack(vjust = 0.5), color = "white",  
size = 4) +  
  labs(x = "AQI dióxido de carbono", y = "Porcentaje de  
observaciones", fill = "Fumador") +  
  theme_minimal()
```

```
df <- ESHOM_in %>% filter(complete.cases(.)) %>%  
  count(AQI_pm25, HOM_smoking) %>%  
  mutate(pct = n / sum(n) * 100)  
ggplot(df, aes(x = AQI_pm25, y = pct, fill = factor(HOM_smoking,  
levels = c("YES", "NO")))) +  
  geom_bar(stat = "identity", position = "stack") +  
  scale_fill_manual(values = c("NO" = "skyblue2", "YES" = "salmon")) +  
  geom_text(aes(label = paste0(round(pct, 1), "%")),  
            position = position_stack(vjust = 0.5), color = "white",  
size = 4) +  
  labs(x = "AQI PM 2.5", y = "Porcentaje de observaciones", fill =  
"Fumador") +  
  theme_minimal()
```

```
df <- ESHOM_in %>% filter(complete.cases(.)) %>%  
  count(AQI_formaldehyde, HOM_smoking) %>%  
  mutate(pct = n / sum(n) * 100)  
ggplot(df, aes(x = AQI_formaldehyde, y = pct, fill =  
factor(HOM_smoking, levels = c("YES", "NO")))) +  
  geom_bar(stat = "identity", position = "stack") +  
  scale_fill_manual(values = c("NO" = "skyblue2", "YES" = "salmon")) +  
  geom_text(aes(label = paste0(round(pct, 1), "%")),  
            position = position_stack(vjust = 0.5), color = "white",  
size = 4) +  
  labs(x = "AQI formaldehído", y = "Porcentaje de observaciones", fill  
= "Fumador") +  
  theme_minimal()
```

#Polonia

```
df <- PLHOM_in %>% filter(complete.cases(.)) %>%
```

```

count(sub_scenariio, AQI_co2, AQI_pm25) %>%
mutate(pct = n / sum(n) * 100)
ggplot(df, aes(x = sub_scenariio, y = pct, fill = factor(AQI_co2,
levels = c("UNHEALTHY", "MODERATE", "GOOD")))) +
geom_bar(stat = "identity", position = "stack") +
facet_wrap(~ AQI_pm25) +
geom_text(aes(label = paste0(round(pct, 1), "%"),
position = position_stack(vjust = 0.5), color = "white",
size = 4) +
labs(x = "Tipo de caldera", y = "Porcentaje de observaciones", fill
= "AQI dióxido de carbono", subtitle = "AQI PM2.5") +
theme_minimal()

```

A.3. Pruebas de normalidad

```

# Guardar Los datos para la transformación
ATHOM_in_tr <- ATHOM_in
ESHOM_in_tr <- ESHOM_in
PLHOM_in_tr <- PLHOM_in

# Crear los indices para muestrear
s_AT <- replicate(1000, sample(1:nrow(ATHOM_in), size = 5000))
s_ES <- replicate(1000, sample(1:nrow(ESHOM_in), size = 5000))
s_PL <- replicate(1000, sample(1:nrow(PLHOM_in), size = 5000))

# Función sobre realizar los test en las muestras
library(nortest)
norm_muestras <- function(variable, indices){
  p_1 <- c()
  p_2 <- c()
  for(i in 1:1000){
    x <- variable[indices[i,]]
    p_1 <- c(p_1, shapiro.test(x)$p.value)
    p_2 <- c(p_2, lillie.test(x)$p.value)
  }
  return(c(mean(p_1), mean(p_1 > .05), mean(p_2), mean(p_2 > .05)))
}

library(car)
##### CO2
# Austria
a<-boxCox(ATHOM_in$co2~1)
l<-a$x[which(a$y==max(a$y))]; l

```

```

ATHOM_in_tr$co2 <- (ATHOM_in$co2^1 - 1)/1

qqnorm(ATHOM_in_tr$co2[s_AT[1:100,]])
qqline(ATHOM_in_tr$co2[s_AT[1:100,]])
norm_muestras(ATHOM_in_tr$co2, s_AT)

# España
a<-boxCox(ESHOM_in$co2+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

ESHOM_in_tr$co2 <- ((ESHOM_in$co2+1e-5)^1 - 1)/1

qqnorm(ESHOM_in_tr$co2[s_ES[1:100,]])
qqline(ESHOM_in_tr$co2[s_ES[1:100,]])
norm_muestras(ESHOM_in_tr$co2, s_ES)

# Polonia
a<-boxCox(PLHOM_in$co2+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

PLHOM_in_tr$co2 <- (PLHOM_in$co2^1 - 1)/1

qqnorm(PLHOM_in_tr$co2[s_PL[1:100,]])
qqline(PLHOM_in_tr$co2[s_PL[1:100,]])
norm_muestras(PLHOM_in_tr$co2, s_PL)

##### Humidity
# Austria
qqnorm(ATHOM_in$humidity[s_AT[1:100,]])
qqline(ATHOM_in$humidity[s_AT[1:100,]])
norm_muestras(ATHOM_in$humidity, s_AT)

# España
qqnorm(ESHOM_in$humidity[s_ES[1:100,]])
qqline(ESHOM_in$humidity[s_ES[1:100,]])
norm_muestras(ESHOM_in$humidity, s_ES)

# Polonia
qqnorm(PLHOM_in$humidity[s_PL[1:100,]])
qqline(PLHOM_in$humidity[s_PL[1:100,]])
norm_muestras(PLHOM_in$humidity, s_PL)

##### PM25
# Austria
a<-boxCox(ATHOM_in$pm25+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

```

```

ATHOM_in_tr$pm1 <- (ATHOM_in$pm1^1 - 1)/1
ATHOM_in_tr$pm25 <- (ATHOM_in$pm25^1 - 1)/1
ATHOM_in_tr$pm10 <- (ATHOM_in$pm10^1 - 1)/1

qqnorm(ATHOM_in_tr$pm25[s_AT[1:100,]])
qqline(ATHOM_in_tr$pm25[s_AT[1:100,]])
norm_muestras(ATHOM_in_tr$pm25, s_AT)

# España
a<-boxCox(ESHOM_in$pm25+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

ESHOM_in_tr$pm1 <- log(ESHOM_in$pm1 + 1e-5)
ESHOM_in_tr$pm25 <- log(ESHOM_in$pm25 + 1e-5)
ESHOM_in_tr$pm10 <- log(ESHOM_in$pm10 + 1e-5)

qqnorm(ESHOM_in_tr$pm25[s_ES[1:100,]])
qqline(ESHOM_in_tr$pm25[s_ES[1:100,]])
norm_muestras(ESHOM_in_tr$pm25, s_ES)

# Polonia
a<-boxCox(PLHOM_in$pm25+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

PLHOM_in_tr$pm1 <- (PLHOM_in$pm1^1 - 1)/1
PLHOM_in_tr$pm25 <- (PLHOM_in$pm25^1 - 1)/1
PLHOM_in_tr$pm10 <- (PLHOM_in$pm10^1 - 1)/1

qqnorm(PLHOM_in_tr$pm25[s_PL[1:100,]])
qqline(PLHOM_in_tr$pm25[s_PL[1:100,]])
norm_muestras(PLHOM_in_tr$pm25, s_PL)

##### Temperature
# Austria
qqnorm(ATHOM_in_tr$temperature[s_AT[1:100,]])
qqline(ATHOM_in_tr$temperature[s_AT[1:100,]])
norm_muestras(ATHOM_in_tr$temperature, s_AT)

# España
qqnorm(ESHOM_in_tr$temperature[s_ES[1:100,]])
qqline(ESHOM_in_tr$temperature[s_ES[1:100,]])
norm_muestras(ESHOM_in_tr$temperature, s_ES)

# Polonia
qqnorm(PLHOM_in$temperature[s_PL[1:100,]])
qqline(PLHOM_in$temperature[s_PL[1:100,]])
norm_muestras(PLHOM_in$temperature, s_PL)

```

```
##### TVOCs
# Austria
a<-boxCox(ATHOM_in$tvocs + 1e-5 ~1)
l<-a$x[which(a$y==max(a$y))]; l

ATHOM_in_tr$tvocs <- log(ATHOM_in$tvocs + 1e-5)

qqnorm(ATHOM_in_tr$tvocs[s_AT[1:100,]])
qqline(ATHOM_in_tr$tvocs[s_AT[1:100,]])
norm_muestras(ATHOM_in_tr$tvocs, s_AT)

# España
a<-boxCox(ESHOM_in$tvocs+1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

ESHOM_in_tr$tvocs <- (ESHOM_in$tvocs^1 - 1)/1

qqnorm(ESHOM_in_tr$tvocs[s_ES[1:100,]])
qqline(ESHOM_in_tr$tvocs[s_ES[1:100,]])
norm_muestras(ESHOM_in_tr$tvocs, s_ES)

# Polonia
a<-boxCox(PLHOM_in$tvocs + 1e-5 ~ 1)
l<-a$x[which(a$y==max(a$y))]; l

PLHOM_in_tr$tvocs <- log(PLHOM_in$tvocs + 1e-5)

qqnorm(PLHOM_in_tr$tvocs[s_PL[1:100,]])
qqline(PLHOM_in_tr$tvocs[s_PL[1:100,]])
norm_muestras(PLHOM_in_tr$tvocs, s_PL)

##### Formaldehído
qqnorm(ESHOM_in$formaldehyde[s_ES[1:100,]])
qqline(ESHOM_in$formaldehyde[s_ES[1:100,]])
norm_muestras(ESHOM_in$formaldehyde, s_ES)
```

A.4. Análisis de Componentes Principales

```
library(stats)
library(factoextra)
#####
##### Original #####
#####

#### Austria
AT_PCA <- prcomp(scale(m_AT_in[, c(3:9)]))
```

```

summary(AT_PCA)
screeplot(AT_PCA, type = "lines")

AT_PCA$rotation
fviz_pca_biplot(AT_PCA, label = "var", habillage = m_AT_in$season,
                 palette = c("gold", "green", "red", "skyblue1"),
                 col.var = "black")

##### España
ES_PCA <- prcomp(scale(m_ES_in[, c(3:10)]))
summary(ES_PCA)
screeplot(ES_PCA, type = "lines")

ES_PCA$rotation
fviz_pca_biplot(ES_PCA, label = "var", habillage = m_ES_in$season,
                 palette = c("gold", "green", "red", "skyblue1"),
                 col.var = "black")

##### Polonia
PL_PCA <- prcomp(scale(m_PL_in[,c(3:9)]))
summary(PL_PCA)
screeplot(PL_PCA, type = "lines", main = "")

PL_PCA$rotation
fviz_pca_biplot(PL_PCA, label = "var", habillage = m_PL_in$season,
                 palette = c("gold", "green", "red", "skyblue1"),
                 col.var = "black")

##### Box-Cox #####
##### Austria
df_AT <- ATHOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                          by = .(X_time=format(X_time,"%Y-%m-%d %H"), season),
                          .SDcols = 6:12]
AT_PCA_w <- prcomp(scale(df_AT[,c(3:9)]))
summary(AT_PCA_w)
screeplot(AT_PCA_w, type = "lines")

AT_PCA_w$rotation
fviz_pca_biplot(AT_PCA_w, label = "var", habillage = m_AT_in$season,
                 palette = c("gold", "green", "red", "skyblue1"),
                 col.var = "black")
fviz_pca_biplot(AT_PCA_w, label = "var", habillage = m_AT_in$season,
                 axes = c(2,3),
                 palette = c("gold", "green", "red", "skyblue1"),

```

```
col.var = "black")

##### España
df_ES <- ESHOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                        by = .(X_time=format(X_time,"%Y-%m-%d %H"), season),
                        .SDcols = 6:13]
ES_PCA_w <- prcomp(scale(df_ES[,c(3:10)]))
summary(ES_PCA_w)
screeplot(ES_PCA_w, type = "lines")

ES_PCA_w$rotation
fviz_pca_biplot(ES_PCA_w, label = "var", habillage = m_ES_in$season,
                palette = c("gold", "green", "red", "skyblue1"),
                col.var = "black")

fviz_pca_biplot(ES_PCA_w, label = "var", habillage = m_ES_in$season,
                axes = c(1,3),
                palette = c("gold", "green", "red", "skyblue1"),
                col.var = "black")

##### Polonia
df_PL <- PLHOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                        by = .(X_time=format(X_time,"%Y-%m-%d %H"),
season), .SDcols = 6:12]
PL_PCA_w <- prcomp(scale(df_PL[,c(3:9)]))
summary(PL_PCA_w)
screeplot(PL_PCA_w, type = "lines", main = "")

PL_PCA_w$rotation
fviz_pca_biplot(PL_PCA_w, label = "var", habillage = m_PL_in$season,
                palette = c("gold", "green", "red", "skyblue1"),
                col.var = "black")

fviz_pca_biplot(PL_PCA_w, label = "var", habillage = m_PL_in$season,
                axes = c(1,3),
                palette = c("gold", "green", "red", "skyblue1"),
                col.var = "black")
```

A.5. Redes Bayesianas

```
library(bnlearn)

##### Austria
# Transformar columnas a factores
nombres_columnas <- c("season", "sub_scenario", "AQI_pm25", "AQI_co2",
```

```

"HOM_smoking")
df <- ATHOM_in[, lapply(.SD, as.factor), .SDcols = nombres_columnas]

# Aprender estructura
AT_RB <- hc(df, maxp = 2)
plot(AT_RB, main = "2 nodos padres máximo")

AT_RB <- hc(df, maxp = 1)
plot(AT_RB, main = "1 nodo padre máximo")

##### España
nombres_columnas <- c("season", "AQI_pm25",
"AQI_co2", "AQI_formaldehyde", "HOM_smoking")
df <- ESHOM_in[, lapply(.SD, as.factor), .SDcols = nombres_columnas]

# Aprender estructura
ES_RB <- hc(df, maxp = 2)
plot(ES_RB, main = "2 nodos padres máximo")

ES_RB <- hc(df, maxp = 1)
plot(ES_RB, main = "1 nodo padre máximo")

##### Polonia
nombres_columnas <- c("season", "sub_scenario", "AQI_pm25", "AQI_co2")
df <- PLHOM_in[, lapply(.SD, as.factor), .SDcols = nombres_columnas]

# Aprender estructura
PL_RB <- hc(df, maxp = 2)
plot(PL_RB, main = "2 nodos padres máximo")

PL_RB <- hc(df, maxp = 1)
plot(PL_RB, main = "1 nodo padre máximo")

##### Total
nombres_columnas <- c("season", "AQI_pm25",
"AQI_co2", "AQI_formaldehyde",
"sub_scenario", "HOM_smoking", "country")
df <- HOM_in[, lapply(.SD, as.factor), .SDcols = nombres_columnas]
df$sub_scenario[df$country=="Spain"] <- NA

# Aprender estructura
T_RB <- hc(df, maxp = 2)
plot(T_RB, main = "2 nodos padres máximo")

T_RB <- hc(df, maxp = 1)
plot(T_RB, main = "1 nodo padre máximo")

```

A.6. Pruebas de Kruskal-Wallis y de Dunn

```
library(dunn.test)

HOM_in_tr <- HOM_in
# Eliminar sub_scenariio de España
HOM_in_tr$sub_scenariio[HOM_in_tr$country=="Spain"] <- NA

##### Season
df <- m_in

d<- dunn.test(df$co2, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$co2)) - 1)
d$Z/sqrt(sum(!is.na(df$co2)))

d<- dunn.test(df$humidity, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$humidity)) - 1)
d$Z/sqrt(sum(!is.na(df$humidity)))

d<- dunn.test(df$pm25, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$pm25)) - 1)
d$Z/sqrt(sum(!is.na(df$pm25)))

d<- dunn.test(df$temperature, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$temperature)) - 1)
d$Z/sqrt(sum(!is.na(df$temperature)))

d<- dunn.test(df$tvocs, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$tvocs)) - 1)
d$Z/sqrt(sum(!is.na(df$tvocs)))

d<- dunn.test(df$formaldehyde, df$season)
(d$chi2 - 4 - 1)/(sum(!is.na(df$formaldehyde)) - 1)
d$Z/sqrt(sum(!is.na(df$formaldehyde)))

##### country
df2 <- HOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                      by = .(format(X_time,"%Y-%m-%d %H"), country), .SDcols
= 6:13]

d<- dunn.test(df2$co2, df2$country)
(d$chi2 - 3 - 1)/(sum(!is.na(df2$co2)) - 1)
d$Z/sqrt(sum(!is.na(df2$co2)))

d<- dunn.test(df2$humidity, df2$country)
(d$chi2 - 3 - 1)/(sum(!is.na(df2$humidity)) - 1)
d$Z/sqrt(sum(!is.na(df2$humidity)))
```

```

d<- dunn.test(df2$pm25, df2$country)
(d$chi2 - 3 - 1)/(sum(!is.na(df2$pm25)) -1)
d$Z/sqrt(sum(!is.na(df2$pm25)))

d<- dunn.test(df2$temperature, df2$country)
(d$chi2 - 3 - 1)/(sum(!is.na(df2$temperature)) -1)
d$Z/sqrt(sum(!is.na(df2$temperature)))

d<- dunn.test(df2$tvocs, df2$country)
(d$chi2 - 3 - 1)/(sum(!is.na(df2$tvocs)) -1)
d$Z/sqrt(sum(!is.na(df2$tvocs)))

##### HOM_smoking
df3 <- HOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                        by = .(format(X_time,"%Y-%m-%d %H"), HOM_smoking),
                        .SDcols = 6:13]

d<- dunn.test(df3$co2, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$co2)) -1)
d$Z/sqrt(sum(!is.na(df3$co2)))

d<- dunn.test(df3$humidity, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$humidity)) -1)
d$Z/sqrt(sum(!is.na(df3$humidity)))

d<- dunn.test(df3$pm25, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$pm25)) -1)
d$Z/sqrt(sum(!is.na(df3$pm25)))

d<- dunn.test(df3$temperature, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$temperature)) -1)
d$Z/sqrt(sum(!is.na(df3$temperature)))

d<- dunn.test(df3$tvocs, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$tvocs)) -1)
d$Z/sqrt(sum(!is.na(df3$tvocs)))

d<- dunn.test(df3$formaldehyde, df3$HOM_smoking)
(d$chi2 - 1 - 1)/(sum(!is.na(df3$formaldehyde)) -1)
d$Z/sqrt(sum(!is.na(df3$formaldehyde)))

##### sub_scenario
df4 <- HOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                    by = .(format(X_time,"%Y-%m-%d %H"), sub_scenario),
                    .SDcols = 6:12]

d<- dunn.test(df4$co2, df4$sub_scenario)
(d$chi2 - 3 - 1)/(sum(!is.na(df4$co2)) -1)

```

```

d$Z/sqrt(sum(!is.na(df4$co2)))

d<- dunn.test(df4$humidity, df4$sub_scenario)
(d$chi2 - 3 - 1)/(sum(!is.na(df4$humidity)) -1)
d$Z/sqrt(sum(!is.na(df4$humidity)))

d<- dunn.test(df4$pm25, df4$sub_scenario)
(d$chi2 - 3 - 1)/(sum(!is.na(df4$pm25)) -1)
d$Z/sqrt(sum(!is.na(df4$pm25)))

d<- dunn.test(df4$temperature, df4$sub_scenario)
(d$chi2 - 3 - 1)/(sum(!is.na(df4$temperature)) -1)
d$Z/sqrt(sum(!is.na(df4$temperature)))

d<- dunn.test(df4$tvocs, df4$sub_scenario)
(d$chi2 - 3 - 1)/(sum(!is.na(df4$tvocs)) -1)
d$Z/sqrt(sum(!is.na(df4$tvocs)))

##### AQI_co2
df5 <- HOM_in[, lapply(.SD, mean, na.rm = TRUE),
                    by = .(format(X_time,"%Y-%m-%d %H"), AQI_co2),
                    .SDcols = 6:13]

d<- dunn.test(df5$co2, df5$AQI_co2)
(d$chi2 - 4 - 1)/(sum(!is.na(df5$co2)) -1)
d$Z/sqrt(sum(!is.na(df5$co2)))

d<- dunn.test(df5$humidity, df5$AQI_co2)
(d$chi2 - 3 - 1)/(sum(!is.na(df5$humidity)) -1)
d$Z/sqrt(sum(!is.na(df5$humidity)))

d<- dunn.test(df5$pm25, df5$AQI_co2)
(d$chi2 - 3 - 1)/(sum(!is.na(df5$pm25)) -1)
d$Z/sqrt(sum(!is.na(df5$pm25)))

d<- dunn.test(df5$temperature, df5$AQI_co2)
(d$chi2 - 3 - 1)/(sum(!is.na(df5$temperature)) -1)
d$Z/sqrt(sum(!is.na(df5$temperature)))

d<- dunn.test(df5$tvocs, df5$AQI_co2)
(d$chi2 - 3 - 1)/(sum(!is.na(df5$tvocs)) -1)
d$Z/sqrt(sum(!is.na(df5$tvocs)))

d<- dunn.test(df5$formaldehyde, df5$AQI_co2)
(d$chi2 - 3 - 1)/(sum(!is.na(df5$formaldehyde)) -1)
d$Z/sqrt(sum(!is.na(df5$formaldehyde)))

##### AQI_pm25

```

```

df6 <- HOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                      by = .(format(X_time, "%Y-%m-%d %H"), AQI_pm25),
                      .SDcols = 6:13]

d<- dunn.test(df6$co2, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$co2)) - 1)
d$Z/sqrt(sum(!is.na(df6$co2)))

d<- dunn.test(df6$humidity, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$humidity)) - 1)
d$Z/sqrt(sum(!is.na(df6$humidity)))

d<- dunn.test(df6$pm25, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$pm25)) - 1)
d$Z/sqrt(sum(!is.na(df6$pm25)))

d<- dunn.test(df6$temperature, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$temperature)) - 1)
d$Z/sqrt(sum(!is.na(df6$temperature)))

d<- dunn.test(df6$tvocs, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$tvocs)) - 1)
d$Z/sqrt(sum(!is.na(df6$tvocs)))

d<- dunn.test(df6$formaldehyde, df6$AQI_pm25)
(d$chi2 - 3 - 1)/(sum(!is.na(df6$formaldehyde)) - 1)
d$Z/sqrt(sum(!is.na(df6$formaldehyde)))

##### AQI_formaldehyde
df7 <- HOM_in_tr[, lapply(.SD, mean, na.rm = TRUE),
                  by = .(format(X_time, "%Y-%m-%d %H"),
AQI_formaldehyde), .SDcols = 6:13]

d<- dunn.test(df7$co2, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$co2)) - 1)
d$Z/sqrt(sum(!is.na(df7$co2)))

d<- dunn.test(df7$humidity, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$humidity)) - 1)
d$Z/sqrt(sum(!is.na(df7$humidity)))

d<- dunn.test(df7$pm25, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$pm25)) - 1)
d$Z/sqrt(sum(!is.na(df7$pm25)))

d<- dunn.test(df7$temperature, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$temperature)) - 1)
d$Z/sqrt(sum(!is.na(df7$temperature)))

```

```
d<- dunn.test(df7$tvocs, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$tvocs)) - 1)
d$Z/sqrt(sum(!is.na(df7$tvocs)))

d<- dunn.test(df7$formaldehyde, df7$AQI_formaldehyde)
(d$chi2 - 3 - 1)/(sum(!is.na(df7$formaldehyde)) - 1)
d$Z/sqrt(sum(!is.na(df7$formaldehyde)))
```