Universitat autònoma de Barcelona

Trabajo de final de grado

Inferencia para datos funcionales: Estudio sobre la contaminación del aire en Barcelona

Autora:
María Hernández
NIU: 1334223

Tutora:
Alejandra Cabaña

19 de enero de 2016

Resumen

En este estudio introducimos el problema de comparación de dos muestras para datos funcionales, usando datos de niveles de dióxido de nitrógeno en diferentes puntos de la ciudad de Barcelona durante los años 2014 y 2015.

Contexto: La contaminación en las grandes ciudades es un grave problema. Analizar de manera adecuada los niveles de contaminación para poder detectar la necesidad de activación de protocolos, como ha pasado en Madrid, es muy importante.

Objectivos: A partir de datos de alta frecuencia tratados como datos funcionales ver si las medias de dos muestras son diferentes según tipos de días, zonas de Barcelona y años.

Métodos: Comparación de dos medias con datos funcionales, usando un estadístico de contraste basado en proyecciones en el espacio de las d primeras componentes principales del operador de covarianzas.

Resultados: Claras diferencias entre los días laborables y festivos (más contaminación en los laborables), también diferencias entre los niveles de NO_2 de 2014 y 2015; diferencias entre todas las zonas, excepto entre Sants i Palau Reial en días festivos. Además, l'Eixample es la zona más contaminada.

Conclusiones: En este estudio se ha intentado dar una visión general sobre los datos funcionales y una solución al problema de comparación de dos muestras con este tipo de datos. Para ello se han implementado funciones en R que han permitido afirmar los resultados anteriores, y éstos son bastante coherentes debido a que los niveles más altos de NO_2 corresponden a las horas de más tráfico y no hay resultados especialmente sorprendentes.

Palabras clave: análisis de datos funcionales (FDA), comparación de medias, dos muestras, análisis de componentes principales (PCA), dióxido de nitrógeno (NO2).

Abstract

In this research we introduce the problem about comparing two samples for functional data, using as data example the levels of nitrogen dioxide in diferents points of Barcelona between years 2014 and 2015.

Context: The pollution in big cities is a major problem. A properly analisis pollution levels to detect the need for protocols activation, as happened in Madrid, is very important.

Objectives: Use high frequency data treated as functional data to contrast if the means of two sample are different according to types of days, areas of Barcelona and years.

Methods: Comparison of two means with functional data, using a contrast statistic based on projections in the space of the d first principal components of the covariance operator.

Results: Clear differences between working days and non working days (most pollution in working), also differences between the levels of NO_2 in 2014 and 2015; differences between all areas except between Sants i Palau Reial at non working days. As more to say, l'Eixample is the most contaminated area.

Conclusions: This study has tried to give an overview to functional data and a way to solve the problem of comparing two samples with this type of data. For this we have implemented functions in R that empower we to claim the results previously commented, we trust R results because the highest levels of NO_2 correspond to rush hour of the working days, which leads us to say that there isn't specially surprising results.

Key words: functional data analysis (FDA), comparison of means, two samples, principal component analysis (PCA), nitrogen dioxide (NO2).

Resum

En aquest estudi introduirem el problema de comparació de dues mostres per a dades funcionals, utilitzant dades de nivells de diòxid de nitrogen en diferents punts de la ciutat de Barcelona durant els anys 2014 i 2015.

Context: La contaminació a les grans ciutats és un greu problema. Analitzar de manera adequada els nivells de contaminació per poder detectar la necessitat d'activació de protocols, com ha passat a Madrid, és molt important.

Objectius: A partir de dades d'alta freqüència tractades com a dades funcionals veure si les mitjanes de dues mostres són diferents segons els tipus de dies, zones de Barcelona i anys.

Mètodes: Comparació de dues mitjanes amb dades funcionals, utilitzant un estadístic de contrast basat en les projeccions a l'espai de les d primeres components principals de l'operador de covariàncies.

Resultats: Clares diferències entre els dies laborables i festius (més contaminació als laborables), també diferencies entre els nivells de NO_2 de 2014 i 2015; diferencies entre totes les zones, excepte entre Sants i Palau Reial en dies festius. A més, l'Eixample és la zona més contaminada.

Conclusions: En aquest estudi s'ha intentat donar una visió general sobre les dades funcionals i una solució al problema de comparació de dues mostres amb aquest tipus de dades. Per a això s'han implementat funcions en R que han permès afirmar els resultats anteriors, i aquests són bastant coherents ja que els nivells més alts de NO_2 corresponen a les hores de més tràfic i no hi han resultats especialment sorprenents.

Paraules clau: anàlisi de dades funcionals (FDA), comparació de mitjanes, dues mostres, anàlisi de components principals (PCA), diòxid de nitrogen (NO2).

Agradecimientos

Querría agradecer a Alejadra Cabaña su ayuda para poder realizar este trabajo. También agradecer a los profesores Manuel Oviedo y Manuel Febrero de la Universidad de Santiago de Compostela por ayudarnos a modificar código R de su paquete `fda.usc`.

Índice

1. Introducción

1.1. Sobre los datos funcionales

El *Análisis de Datos Funcionales* (FDA por sus siglas en inglés) se refiere a problemas estadísticos donde los datos son una muestra de n funciones $x_1(t), \ldots, x_n(t)$ definidas en un intérvalo $[a, b] \in \mathbb{R}$, donde t suele denotar el tiempo. Estos datos, por su naturaleza, suelen ser continuos o bien discretos con alta frecuencia, donde cada observación es una curva que evoluciona con el tiempo. En este estudio vamos a limitarnos a funciones $x : [a, b] \to \mathbb{R}$.

Las técnicas para hacer FDA incorporan herramientas estándar de la estadística multivariante puesto que, en el fondo, cada observación consiste en un vector de evaluaciones de la función en una discretización de $[a, b]$, generando una gran cantidad de datos.
A diferencia de la estadística clásica a veces destinguir si unos datos son funcionales o no es algo difícil. Hay dos razones principales para considerar que unos datos son funcionales:

1. La posibilidad (al menos teórica) de observar el fenómeno en un cuadrícula muy fina y, en el límite, observar $x(t)$ en cualquier instante t fijado.

2. La opción de escoger un modelo funcional para aproximar la representación de los datos.

Normalmente "las filas" de datos necesitan tratamiento preliminar antes de aplicar las técnicas de FDA. Esto usualmente está motivado por reducción de dimensión o por eliminación del ruido en las mediciones de los datos.
Una forma común de transformar los datos es mediante la representación de bases de splines, asumiendo $x \in L^2[a, b]$, donde $L^2[a, b]$ es el espacio de funciones en $\{f_{(a,b)} = \int_a^b f^2(t)dt < \infty\}$ y los splines son curvas diferenciables definidas en $[a, b]$ mediante polinomios.

En este estudio, la base de datos son niveles de NO_2 cada hora (medias horarias) que forman una curva que evoluciona con el tiempo para cada día (figura 1).

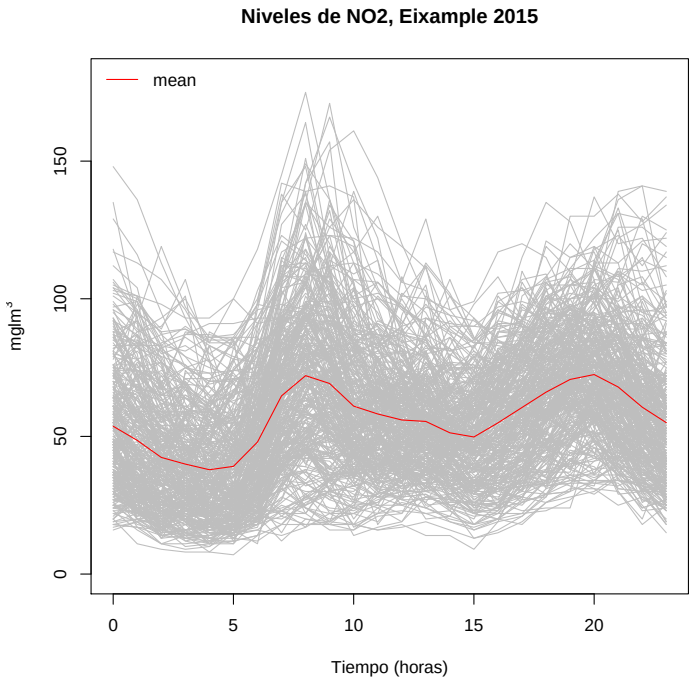

Figura 1: Gráfico datos Eixample 2015

1.2. Sobre el NO, NO_2 y NO_x

El óxido nítrico (NO) y el dióxido de nitrógeno (NO_2) son los únicos óxidos de nitrógeno en la atmósfera. Son introducidos por el hombre, y forman el óxido de nitrógeno en la fórmula $NO_x = NO + NO_2$.

Estos gases son contaminantes. En altas concentraciones producen problemas respiratorios, de crecimiento y además de pérdida de clorofila en la vegetación. También corroen tejidos vegetales y otros materiales.

La principal fuente de emisión de NO_x y NO_2 son las reacciones de combustión de los vehículos, por tanto al analizar los datos del NO_2 se podrá apreciar que a las horas de mayor tráfico en Bacerlona habrá una mayor concentración de NO_2. Otra fuente de emisión importante de estos gases son las instalaciones de combustión de las grandes industrias.

La concentración de estos gases está regulada por el El Real decreto 102/2011, relativo a la mejora de la calidad del aire, estableciendo los valores límites del dióxido de nitrógeno:

Media anual	40 $\mu g/m^3$
Media horaria	200 $\mu g/m^3$
3 horas consecutivas una zona	400 $\mu g/m^3$

En caso de superar estos umbrales es necesaria la activación de los protocolos destinados a disminuir los niveles de contaminantes en el aire.

1.3. Sobre el análisis de componentes principales

El *Análisis de Componentes Principales* (PCA por sus siglas en inglés) consiste en reducir las componentes que explican una variable cuando esta tiene muchas variables explicativas. En los datos funcionales se trata también de reducir las componentes, de manera que usando menos funciones podamos explicar un porcentage alto, sobre el 80 % - 90 %, de la varianza de éstos.
De esta manera, usando modelos más simples podemos analizar correctamente los datos que tenemos porque la varianza sigue informando de forma adecuada.

El análisis de componentes principales se basa en usar los vectores y valores propios (los valores propios indican la desviación estándar explicada) de los datos, cosa que usando datos funcionales en éstos se han de buscar estos valores y vectores propios sobre sus bases: los vectores propios serán, en este caso, funciones.

2. Material y métodos

Usando los datos que proporciona la Generalitat de Catalunya sobre la contaminación (disponibles en la web `http://dtes.gencat.cat/icqa/`), si cogemos las medias horarias (es decir, 24 datos para cada día) tendremos datos que se pueden considerar funcionales, con una curva para cada día, a los que podremos aplicar la teoría de datos funcionales.

En las estaciones de monitoreo se miden las siguentes sustancias contaminantes del aire:

- Arsénico (AR)
- Benceno (C6H6)
- Monóxido de nitrógeno (NO)
- Dióxido de nitrogeno (NO2)
- Plomo (Pb)
- Partículas en suspensión $< 10\mu g$ (PM10)
- Partículas en suspensión $< 2,5\mu g$ (PM2.5)
- Dióxido de azufre (SO2)

De las cuales únicamente NO y NO_2 se miden en diferentes estaciones de manera automática: el resto de contaminantes en muchas estaciones sólo se miden manualmente. Al medirse manualmente se obtienen únicamente medias diarias, lo que daría datos que no podríamos identificar con modelos funcionales porque únicamente tendríamos un valor por día. Los datos automáticos devuelven medias horarias y por tanto son 24 valores por día.

En defenitiva, los datos que usaremos de esta web para este trabajo son los niveles de NO_2 ya que son los que tienen una interpretación más secilla, obtenidos en l'Eixample, el Poblenou, Sants y Palau Reial en los años 2014 y 2015 (en medias horarias). Las ubicaciones concretas de las estaciones de monitoreo están marcadas en el mapa de la figura 2:

Figura 2: Situación de las estaciones de monitoreo

Los datos tienen la estructura de la tabla 1: cada fila es un día en el que se han medido los niveles de dióxido de nitrógeno y cada columna es la hora del día. De este modo los valores son la media de NO_2 en aquella hora, es decir, tenemos 24 valores que son realmente 24 medias para cada día.

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
2015-01-01	105.00	101.00	81.00	67.00	60.00	44.00	44.00	36.00	45.00	33.00	31.00	25.00
2015-01-03	42.00	56.00	33.00	19.00	14.00	15.00	16.00	14.00	17.00	31.00	24.00	50.00
2015-01-04	104.00	93.00	119.00	97.00	86.00	82.00	75.00	76.00	75.00	47.00	33.00	39.00
2015-01-05	87.00	83.00	61.00	46.00	18.00	17.00	57.00	90.00	118.00	90.00	107.00	106.00
2015-01-06	71.00	87.00	77.00	69.00	35.00	25.00	22.00	18.00	28.00	24.00	23.00	27.00
2015-01-07	56.00	45.00	43.00	36.00	42.00	38.00	53.00	97.00	122.00	111.00	93.00	75.00

	H12	H13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23
2015-01-01	22.00	32.00	26.00	26.00	25.00	35.00	68.00	88.00	69.00	53.00	77.00	64.00
2015-01-03	45.00	51.00	51.00	43.00	44.00	50.00	40.00	39.00	52.00	106.00	130.00	125.00
2015-01-04	42.00	31.00	38.00	32.00	33.00	48.00	91.00	107.00	102.00	110.00	95.00	97.00
2015-01-05	97.00	91.00	59.00	66.00	71.00	66.00	63.00	73.00	86.00	81.00	92.00	60.00
2015-01-06	30.00	31.00	58.00	47.00	59.00	78.00	83.00	60.00	70.00	74.00	66.00	62.00
2015-01-07	101.00	100.00	89.00	82.00	96.00	91.00	85.00	105.00	110.00	125.00	87.00	63.00

Cuadro 1: Datos de los primeros días de 2015 en l'Eixample

2.1. Componentes principales en datos funcionales

Sea X un elemento aleatorio que toma valores en el espacio de muestral $X = L^2[a,b]$, por analogía con el caso de dimensión finita el objetivo de los *Componentes Principales Funcionales* (FPC por sus siglas en inglés) es definir direcciones de proyecciones ortonormales $\varphi_1, \ldots, \varphi_k \in L^2[a,b]$ de manera que las proyecciones de X a lo largo de esas direcciones toman la máxima variabilidad posible (es la idea de PCA explicada en apartado 1.3, donde los vectores propios serían φ).
Por tanto, la primera componente principal es dada por la dirección de proyección φ_1 logrando alcanzar la máxima varianza: $V(\langle \varphi_1, X\rangle) = max\{V(\langle a, X\rangle) : ||a|| = 1\}$. Para la k ésima componente principal se define como $V(\langle \varphi_k, X\rangle) = max\{V(\langle a, X\rangle) : ||a|| = 1,\ \langle a, \varphi_j\rangle = 0, \text{ para } j = 1, \ldots, k-1\}$

La idea esencial sería sustituir el tratamiento estadístico de los datos originales X_i con el vector de k dimensiones correspondiente a las proyecciones $\langle \varphi_1, X_1\rangle, \ldots, \langle \varphi_k, X_i\rangle$. Como en el caso de dimensión finita, es demostrable que las direcciones de los FPC son una base ortonormal del operador de covarianzas $Z(s,t) = Cov(X(s), X(t))$. Además, se deduce que los correspondientes valores propios cumplen $\lambda_j = V\left(\langle \varphi_j, X\rangle\right)$.

La estimación mediante FPC las direcciones φ_j y las varianzas λ_j se pueden realizar usando un estimador apropiado del operador de covarianzas.
Pero hay un poblema de optimización para $V(\langle \varphi_k, X\rangle)$, que es equivalente a que para las funciones $\widehat{\varphi}_j$ con $||\varphi_j|| = 1$ que cumplen:

$$Z_n \varphi_j = \lambda_j \varphi_j$$

para algún $\widehat{\lambda}_j$, encontrar un estimador de Z_n. Éste estaría asociado con la función de covarianza empírica:

$$z(s,t) = \frac{1}{N} \sum_{i=1}^{n} ((X_i(s) - \overline{X}(s)) \cdot (X_i(t) - \overline{X}(t)))$$

El problema está en que, en la práctica, para encontrar una solución necesitamos o bien suavizar los datos o maximizar $\frac{V(\langle \varphi_k, X\rangle)}{||\varphi||^2 + \delta||\varphi''||^2}$, siendo $\delta > 0$ un coeficiente de penalización. Es este estudio usaremos la segunda opción.

2.2. Media de datos funcionales

Análogamente con caso de una sola variable aleatoria clásica, en cuanto a la media, la integral basada en la definición de esperanza funcional permite la motivación habitual en cuanto a las proyecciones: si X es una variable aleatoria que toma valores en el espacio χ y $\mathbb{E}||X||^2 < \infty$ entonces la media de X, $m = \mathbb{E}(X)$, cumple $\mathbb{E}||X - m||^2 = min_{a\in\chi}\mathbb{E}||X - a||^2$

La definición empírica de la media basada en una muestra $X_1, \ldots, X_n$ se define como:

$$\widehat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \quad \forall t \in \chi$$

La especial importancia de este parámetro ha llevado a considerar diferentes modelos en la práctica para estimarlo. Por ejemplo, una suposición habitual es que los datos que tenemos son escasos y afectados por ruido. Entonces observamos:

$$Y_{ij} = X(t_{ij}) + e_{ij} \text{ donde } j = 1, \ldots, m_i \text{ y } i = 1, \ldots, n$$

Donde $t_{ij} \in [0, 1]$ son "puntos de diseño" y e_{ij} son variables de ruido que, supuestamente, son idependientes, homocedásticas y con media 0. La estimación óptima de la función de la media bajo diferentes premisas del diseño de t_{ij} ha sido analizada por *Cai y Yuan (2011)*.

2.3. Comparación de medias de datos funcionales

Siguendo el libro `Inference for functional data with applications`, de Lajos Horváth y Piotr Kokoszka, en el capítulo 5 se muestra un test de igualdad de medias para dos muestras de datos funcionales.

Consideramos dos muestras $X_1, \ldots, X_N$ y $X_1^*, \ldots, X_M^*$ donde

$$X_i(t) = \mu(t) + \varepsilon_i(t), \;\; 1 \le i \le N$$

y

$$X_i^*(t) = \mu^*(t) + \varepsilon_i^*(t), \;\; 1 \le i \le M$$

Con estos datos queremos comprobar la hipótesis nula

$$H_0 : \mu = \mu^* \text{ en } L^2$$

contra la alternativa de que las dos muestras tienen distinta media. $\varepsilon_1, \ldots, \varepsilon_N$ son independientes idénticamente distribuidos (iid) con $\mathbb{E}\,\varepsilon_1(t) = 0$ y $\mathbb{E}\,||\varepsilon_1||^4 < \infty$ y, de manera similar, $\varepsilon_1^*, \ldots, \varepsilon_M^*$ son iid con $\mathbb{E}\,\varepsilon_1^*(t) = 0$ y $\mathbb{E}\,||\varepsilon_1^*||^4 < \infty$. Además ε_i^* y ε_i no tienen la misma distribución.

Los estimadores de μ y μ^* son insesgados, y son respectivamente:

$$\bar{X}_N(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t) \text{ y } \bar{X}_M^*(t) = \frac{1}{M}\sum_{i=1}^{M} X_i^*(t)$$

Por tanto, rechazaremos la hipótesis nula si el siguiente estadístico, $U_{N,M}$ es "demasiado alto".

$$U_{N,M} = \frac{NM}{N+M}\int_0^1 \left(\bar{X}_N(t) - \bar{X}_M^*(t)\right)^2 dt$$

Hay dos métodos para determinarlo. El primer método está basado en proyecciones y el segundo en el cálculo numérico de la integral de $U_{N,M}$ bajo H_0.

Los estadísticos se proyectan de la siguente manera y tienden asintóticamente a las distribuciones indicadas:

1. $T^{(1)}_{N,M} = \frac{NM}{N+M} \sum_{k=1}^{d} \frac{\widehat{a}_k^2}{\widehat{\tau}_k} \to \chi^2_d$

2. $T^{(2)}_{N,M} = \frac{NM}{N+M} \sum_{k=1}^{d} \widehat{a}_k^2 \to \sum_{k=1}^{d} \tau_k N_k^2$

Donde, en $T^{(2)}_{N,M}$ las $N_1, N_2, \ldots, N_d$ son variables aleatorias $\mathcal{N}(0,1)$ independientes.

En este estudio nos centraremos en calcular el estadístico de contraste con el primer método, usando proyecciones en el espacio determinado por las primeras componentes principales del operador de covarianzas Z.

A continuación detallamos el método para calcular $T^{(1)}_{N,M}$

Primero hay que calcular, a partir de los datos centrados, el operador de covarianzas:

$$Z = \theta_1 \cdot C + \theta_2 \cdot C^*$$

Donde $\theta_1 = \frac{M}{N\cdot(M+N)}$ y $\theta_2 = \frac{N}{M\cdot(M+N)}$, y dentro de estas dos fórmulas N es el número de funciones (número de días en este trabajo) de la muestra 1, y M es el número de funciones de la muestra 2.
C y C^* son las autocovarianzas de los datos y se calculan de igual manera, se diferencian en que C usa los datos centrados de la muestra 1 ($\bar{X}_N$) y C^* los datos centrados de la muestra 2 ($\bar{X}_M$).

El cálculo de C y C^* es el siguiente: $C = t(\bar{X}_{N_i}) \cdot \bar{X}_{N_i}$ para cada fila i, donde C son N matrices de nxn y n es la frecuencia de los datos, en este caso, $n = 24$ siempre porque son las horas del día. Por tanto, como cogeremos datos anuales, si por ejemplo si comparamos los datos de l'Eixample con los de Sants obtendríamos que C serían 365 matrices de 24x24 correspondientes a l'Eixample, y C^* tendría las mismas dimensiones, pero con los datos de Sants.
Una vez se ha calculado Z, para calcular el estadístico de contraste hay que extraer mediante la descomposición de la matriz singular los valores propios (λ) y vectores propios (φ), y con estos últimos calcular a:

$$\widehat{a}_i = \langle \bar{X}_N - \bar{X}^*_M, \widehat{\varphi}_i \rangle,\ 1 \leq i \leq d$$

Donde d son el número de primeras componentes principales escogidas. a será una matriz dx1 por ser el resultado de multiplicar una matriz dx24 (correspondiente a la matriz de $\widehat{\varphi}$) con una matriz 1x24 transpuesta (correspondiente a la diferencia de las medias de los datos de ambas muestras, para cada hora).

Una vez tenemos todos estos datos, el estadístico de contraste es:

$$T_{N,M} = \frac{M \cdot N}{M+N} \cdot \sum_{i=1}^{d} \frac{a_i^2}{\lambda_i}$$

Donde a^2 es la matriz a elevada al cuadrado elemento a elemento. Este estadístico tiende asintóticamente a una χ^2_d : los grados de libertad son iguales al número de componentes principales elegidas.

3. Resultados

Primero de todo, para aplicar la teoría del apartado 2.3 tenemos que escojer el número de componentes principales que expliquen correctamente la variabilidad de los datos (d).

La varianza explicada por las componentes principales de los datos funcionales, para las estaciones de monitoreo en 2014, son las siguentes (figura 3).

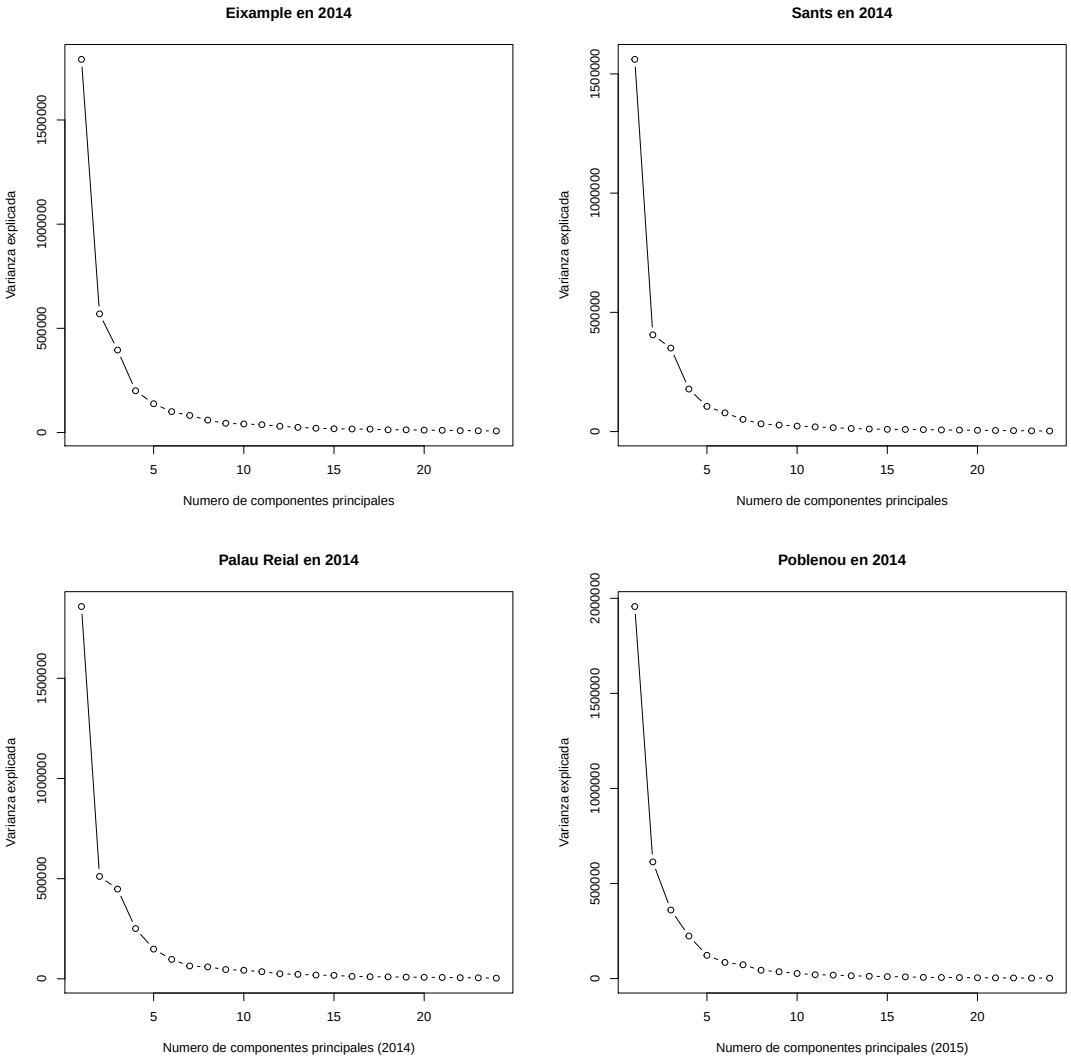

Figura 3: Scree graph de los componentes principales

Vemos que la zona plana se encuentra a partir de la 5^a o 6^a componente principal en cada gráfico, indicando que a partir de esa cantidad de componentes principales la varianza explicada "se estanca" y no explica demasido más.

Para los datos funcionales de 2015, la cantidad de varianza explicada se observa en la figura 4.

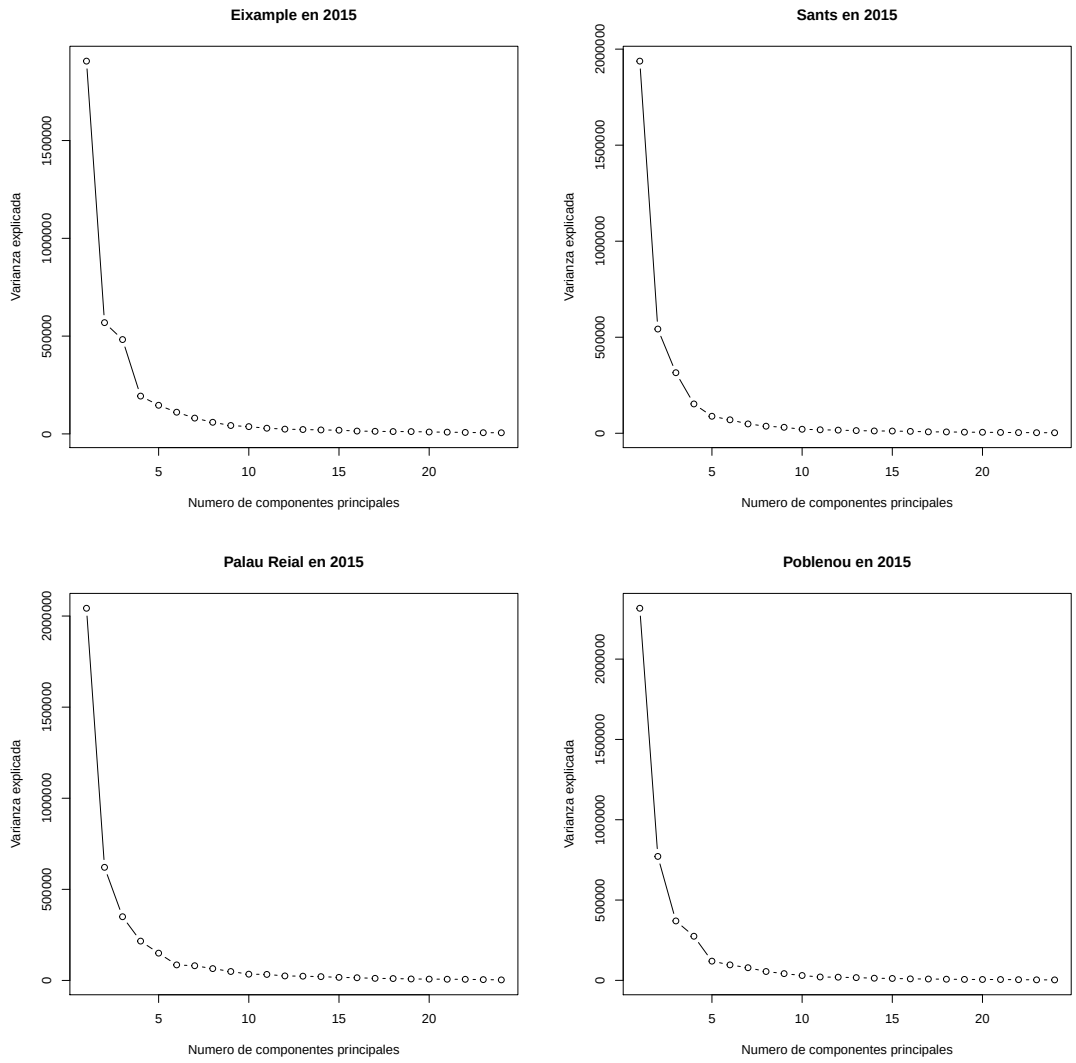

Figura 4: Scree graph de los componentes principales

Vemos que los resultados son prácticamente iguales que para 2014: la zona plana también se encuentra a partir de la 5^a o 6^a componente principal.

Por tanto, para la comparación de medias entre muestras de estos datos usaremos siempre las 5 primeras componentes principales. Con esta información y aplicando un algoritmo para calcular el estadístico de contraste obtenemos los resultados de los apartados siguentes, para las 4 estaciones de moniteo en 2014 y 2015.

3.1. Comparación de medias entre días laborables y festivos

Podemos ver como evolucionan con el tiempo (cada hora) los niveles de NO_2 gráficamente, y con ellos podríamos hacernos una idea de la diferencia entre días laborables los festivos de, por ejemplo, l'Eixample i Palau Reial en 2014:

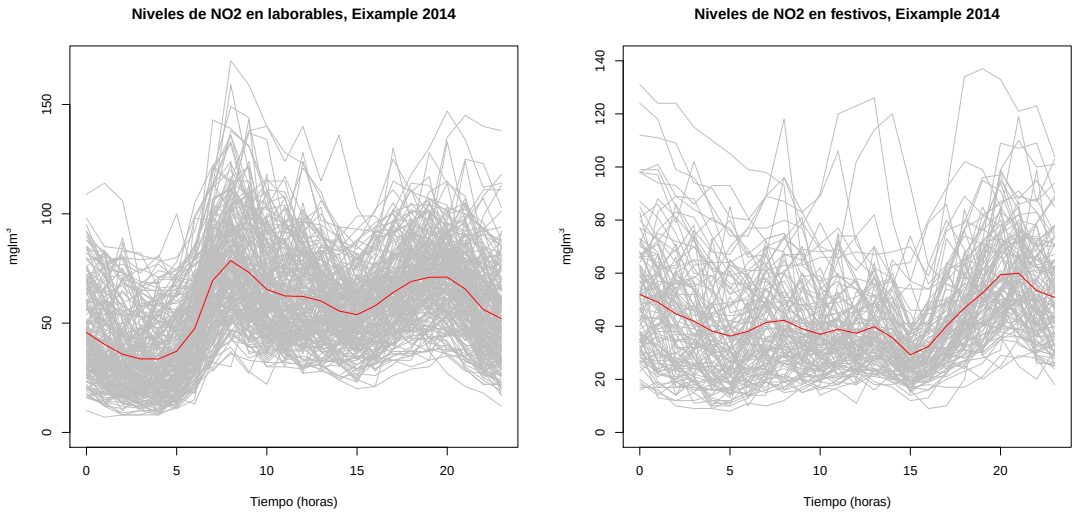

Figura 5: Niveles de NO_2 en laborables y festivos, Eixample 2014

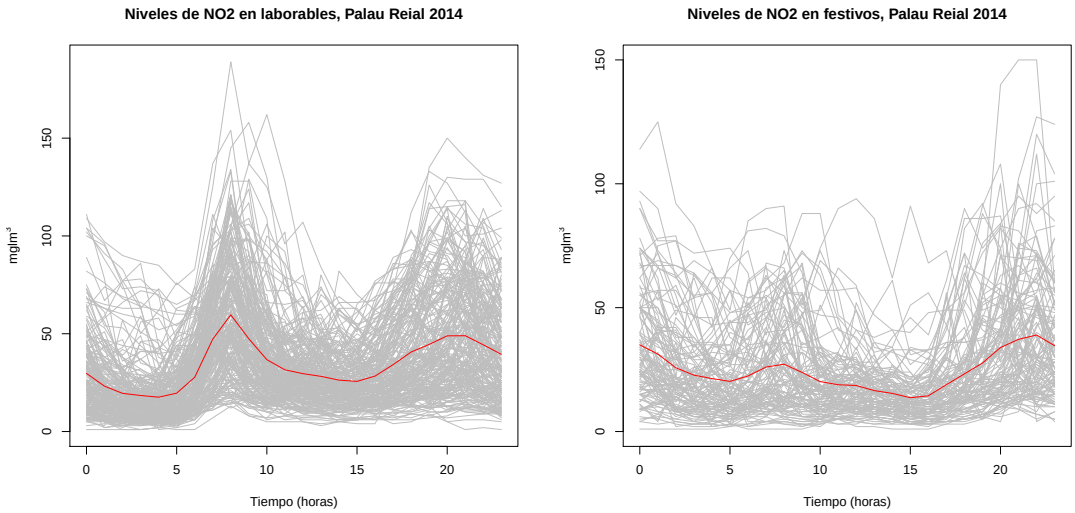

Figura 6: Niveles de NO_2 en laborables y festivos, Palau Reial 2014

Las líneas rojas de los gráficos anteriores (5 y 6) son para cada gráfico la curva media, indicando así un valor medio para cada hora que forma la curva. Las medias de las curvas medias para todas las zonas en 2014 y 2015 son:

	Días laborables y festivos	Días laborables	Días festivos
Sants 2014	31.59	35.61	33.98
Sants 2015	39.01	27.07	29.09
Eixample 2014	52.03	56.48	56.73
Eixample 2015	61.20	43.18	47.13
Palau Reial 2014	30.77	34.34	34.07
Palau Reial 2015	38.32	24.50	26.59
Poblenou 2014	38.96	44.98	42.26
Poblenou 2015	48.68	32.59	37.85

Cuadro 2: Medias anuales de NO_2 de días laborables y festivos para ambos años

Usando el estadístico de contraste del apartado 2.3 obtenemos los siguientes resultados cuantitativos (tabla 3), comparando entre días laborables y festivos de cada zona. El número de componentes principales usados en cada caso son 5 y por tanto el estadístico de contraste se compara con el valor teórico $\chi^2_5 = 11.07$ si usamos un 5 % de confianza.

	Valor del estadístico	P-valor
Sants 2014	259.94	<0.001
Sants 2015	193.69	<0.001
Eixample 2014	589.96	<0.001
Eixample 2015	352.42	<0.001
Palau Reial 2014	239.65	<0.001
Palau Reial 2015	193.42	<0.001
Poblenou 2014	291.11	<0.001
Poblenou 2015	229.29	<0.001

Cuadro 3: Comparación entre días laborables y festivos de 2014 y 2015

Se observan diferencias muy claras entre días laborables y festivos para todas las zonas de Barcelona, tanto en 2014 como en 2015.

3.2. Comparación de medias entre zonas de Barcelona, por años y tipo de días

Usando también las 5 primeras componentes principales los resultados de la comparación de la media de NO_2 de Sants contra la de las otras 3 zonas, en días laborables, es la tabla 4:

	Valor del estadístico	P-valor
Sants vs Eixample 2014	606.97	<0.001
Sants vs Eixample 2015	584.83	<0.001
Sants vs Palau Reial 2014	72.91	<0.001
Sants vs Palau Reial 2015	76.22	<0.001
Sants vs Poblenou 2014	51.37	<0.001
Sants vs Poblenou 2015	52.45	<0.001

Cuadro 4: Comparación entre días laborables de Sants con las otras 3 zonas en 2014 y 2015

Podemos observar que en los días laborables las zonas son diferentes: la media de NO_2 de Sants contra las otras 3 zonas siempre es diferente.

La comparación de la media de NO_2 de Sants contra las otras 3 zonas, en días laborables, da los resultados de la tabla 5:

	Valor del estadístico	P-valor
Sants vs Eixample 2014	184.47	<0.001
Sants vs Eixample 2015	187.1	<0.001
Sants vs Palau Reial 2014	10.57	0.06
Sants vs Palau Reial 2015	20.78	0.06
Sants vs Poblenou 2014	11.24	0.05
Sants vs Poblenou 2015	52.45	0.05

Cuadro 5: Comparación entre días festivos de Sants con las otras 3 zonas en 2014 y 2015

A diferencia que para los días laborables, entre Sants i Palau Reial se puede aceptar con un 5 % de confianza que las medias son iguales. Entre Sants i Poblenou vemos que hay evidencia de moderada a alta (no demasiado fuerte) de que las medias son diferentes.

3.3. Comparación de medias para los años 2014 y 2015, por zonas y tipo de días

Usando también las 5 primeras componentes principales el resultado de la comparación de la media de NO_2 entre años para cada zona se observa en la tabla 6:

	Valor del estadístico	P-valor
Sants	18.28	0.003
Eixample	16.98	0.005
Palau Reial	13.97	0.02
Poblenou	26.1	<0.001

Cuadro 6: Comparación de los datos entre años (2014 y 2015), por zonas

Vemos que sin distinción entre días laborables y festivos no podemos aceptar la hipótesis nula de igualdad de medias para ninguna zona al comparar las muetras de 2014 son las de 2015.

Si dividimos las muestras además de por zonas por días laborables o festivos obtenemos la tabla de resultados 7:

	Valor del estadístico	P-valor
Sants (laborables)	21.87	<0.001
Sants (festivos)	10.26	0.02
Eixample (laborables)	13.98	<0.001
Eixample (festivos)	11.85	0.02
Palau Reial (laborables)	17.53	0.004
Palau Reial (festivos)	5.57	0.004
Poblenou (laborables)	21.44	<0.001
Poblenou (festivos)	14.26	<0.001

Cuadro 7: Comparación de días laborables y festivos entre años (2014 y 2015) de las zonas

Donde observamos que aunque separemos también las muestras por días laborables y festivos existen diferencias entre la media de NO_2 de 2014 con la media de 2015, por tanto, los niveles de NO_2 son diferentes en 2014 y 2015.

4. Conclusión

En este estudio se ha intentado dar una visión general sobre los datos funcionales y una solución al problema de comparación de dos muestras con este tipo de datos.

Aplicando la teoría de análisis de datos funcionales mediante la implementación de funciones en R se han podido analizar unos datos de alta frecuencia, que por tanto tienen una gran dimensión, reduciéndolos y dando unos resultados con los que afirmar lo que indica el apartado anterior:

- Los niveles medios de NO_2 son diferentes entre los días laborales y festivos para cualquier zona y en ambos años: el nivel contaminación es más alto en los días laborables siempre.
- Claras diferencias también entre los niveles medios de NO_2 para los años 2014 y 2015, en todas las zonas, indistintamente entre días laborables y festivos. En 2015 bajaron los niveles de contaminación con respecto a 2014.
- Diferencias en los niveles medios de NO_2 entre las 4 zonas estudiadas de Barcelona, excepto entre Sants i Palau Reial en días festivos, donde se puede asumir que la media es la misma.
- L'Eixample es siempre la zona más contaminada indistintamente del tipo de día en ambos años (comparando los años respectivamente, es decir, los datos de 2014 con los de 2014 y los de 2015 con los de 2015).

Por otra parte, estos resultados podrían estar influenciados por datos extremos (*outliers*), los cuales no han sido retirados de la base de datos. Aún así, los resultados son bastante coherentes debido a que los niveles más altos de NO_2 corresponden a las horas de más tráfico y no hay resultados especialmente sorprendentes.

Referencias

[1] Lajos Horváth, Piotr Kokoszka; "Inference for functional data with applications"

[2] Antonio Cuevas; "A partial overview of the theory of statistics with functional data"

[3] James O. Ramsay, Bernard W. Silverman; "Applied Functional Data Analysis: Methods and Case Studies"

[4] Sobre el NO_2: `http://www.aspb.cat/quefem/docs/oxidos.pdf`

[5] Sobre el NO_2: `http://www.prtr-es.es/NOx-oxidos-de-nitrogeno,15595,11,2007.html`

[6] Base de datos: `http://dtes.gencat.cat/icqa/`

Anexo

Código de R

Para el cálculo del estadístico de contraste del apartado 2.3 se ha utilizado el software R, creando siguentes funciones (requieren los paquetes `fda` y `fda.usc`):

```
############################
## Funcion "fdata2pcnew": ##
############################
fdata2pcnew
function (fdataobj, Sigma = NULL, ncomp = 2, norm = TRUE, lambda = 0,
    P = c(0, 0, 1), ...)
{
    C <- match.call()
    if (!is.fdata(fdataobj))
        stop("No fdata class")
    nas1 <- apply(fdataobj$data, 1, function(x) sum(is.na(x)))
    if (any(nas1))
        stop("fdataobj contain ", sum(nas1), " curves with some NA value \n")
    X <- fdataobj[["data"]]
    tt <- fdataobj[["argvals"]]
    rtt <- fdataobj[["rangeval"]]
    nam <- fdataobj[["names"]]
    mm <- fdata.cen(fdataobj)
    xmean <- mm$meanX
    Xcen.fdata <- mm$Xcen
    dimx <- dim(X)
    n <- dimx[1]
    J <- dimx[2]
    Jmin <- min(c(J, n))
    if (lambda > 0) {
        if (is.vector(P)) {
            P <- P.penalty(tt, P)
        }
        dimp <- dim(P)
        if (!(dimp[1] == dimp[2] & dimp[1] == J))
            stop("Incorrect matrix dimension P")
        M <- solve(diag(J) + lambda * P)
        Xcen.fdata$data <- Xcen.fdata$data %*% t(M)
    }
    if (is.null(Sigma)) {
        eigenres <- svd(Xcen.fdata$data)
        v <- eigenres$v
        u <- eigenres$u
        d <- eigenres$d
        D <- diag(d)
    }
    else {
        eigenres = svd(Sigma)
        d = eigenres$d
        D = diag(d)
        v = eigenres$v
        u = eigenres$u
    }
    vs <- fdata(t(v), tt, rtt, list(main = "fdata2pc", xlab = "t",
        ylab = "rotation"))
```

```
    scores <- matrix(0, ncol = J, nrow = n)
    if (norm) {
        dtt <- diff(tt)
        drtt <- diff(rtt)
        eps <- as.double(.Machine[[1]] * 10)
        inf <- dtt - eps
        sup <- dtt + eps
        if (all(dtt > inf) & all(dtt < sup))
            delta <- sqrt(drtt/(J - 1))
        else delta <- 1/sqrt(mean(1/dtt))
        no <- norm.fdata(vs)
        vs <- vs/delta
        newd <- d * delta
        scores[, 1:Jmin] <- inprod.fdata(Xcen.fdata, vs, ...)
    }
    else {
        scores[, 1:Jmin] <- inprod.fdata(Xcen.fdata, vs, ...)
        newd <- d
    }
    colnames(scores) <- paste("PC", 1:J, sep = "")
    l <- 1:ncomp
    out <- list(call = C, d = newd, rotation = vs[1:ncomp], x = scores,
        lambda = lambda, P = P, fdataobj.cen = Xcen.fdata, norm = norm,
        type = "pc", mean = xmean, fdataobj = fdataobj, l = l,
        u = u[, 1:ncomp, drop = FALSE])
    class(out) = "fdata.comp"
    return(out)
}
##################################
## Funcion "ComparacionMedias": ##
##################################
ComparacionMedias
function (dades, working, nonworking, ncomponentes, confianza = 0.95)
{
    require(fda)
    require(fda.usc)
    if (class(working) == "fdata") {
        working <- working$data
    }
    if (class(nonworking) == "fdata") {
        nonworking <- nonworking$data
    }
    xmean <- apply(working, 2, mean)
    xcen.fdata <- apply(working, 2, function(y) y - mean(y))
    ymean <- apply(nonworking, 2, mean)
    ycen.fdata <- apply(nonworking, 2, function(y) y - mean(y))
    N <- dim(xcen.fdata)[1]
    M <- dim(ycen.fdata)[1]
    n <- dim(xcen.fdata)[2]
    c1 <- M/(N * (M + N))
    c2 <- N/(M * (M + N))
    Cx <- array(NA, dim = c(N, n, n))
    Cy <- array(NA, dim = c(M, n, n))
    for (k in 1:N) {
        Cx[k, , ] <- outer(xcen.fdata[k, ], xcen.fdata[k, ])
    }
```

```
    for (k in 1:M) {
        Cy[k, , ] <- outer(ycen.fdata[k, ], ycen.fdata[k, ])
    }
    Z <- c1 * (apply(Cx, c(2, 3), sum)) + c2 * (apply(Cy, c(2,
        3), sum))
    pcR <- fdata2pcnew(dades, Sigma = Z, ncomp = ncomponentes)
    autovectores <- pcR$rotation$data
    difmedias <- xmean - ymean
    a <- autovectores %*% difmedias
    estad <- (M * N/(M + N)) * sum(a^2/pcR$d[1:ncomponentes])
    res <- list()
    res$Estad <- estad
    res$chi <- qchisq(confianza, df = ncomponentes)
    res$p.val <- format.pval(pchisq(estad, df = ncomponentes,
        lower.tail = FALSE), eps = 0.001, digits = 1)
    res
}
```

Gráficos de los datos

Los gráficos de las curvas de todos los datos de este estudio son los siguentes.

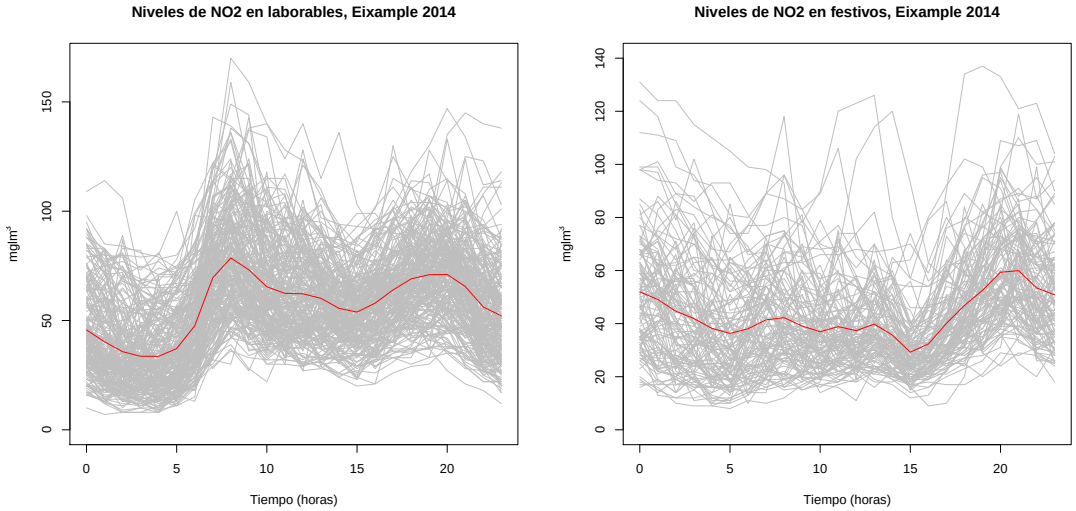

Figura 7: Niveles de NO_2 en laborables y festivos, Eixample 2014

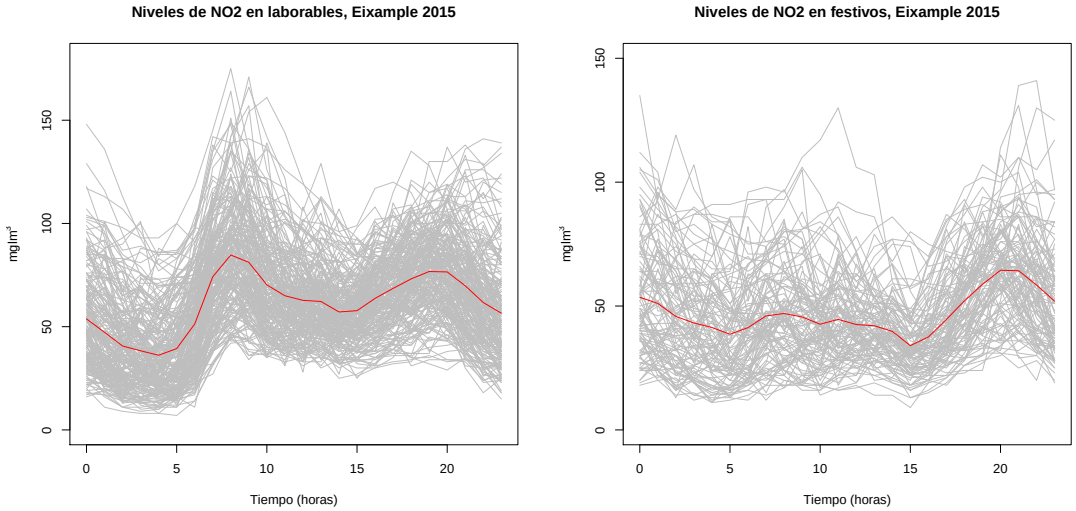

Figura 8: Niveles de NO_2 en laborables y festivos, Eixample 2015

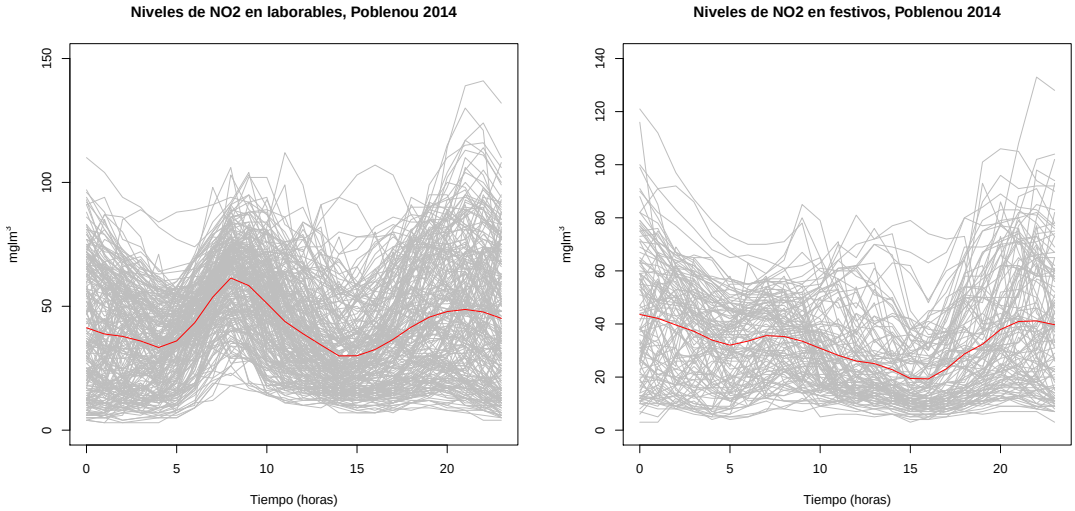

Figura 9: Niveles de NO_2 en laborables y festivos, Poblenou 2014

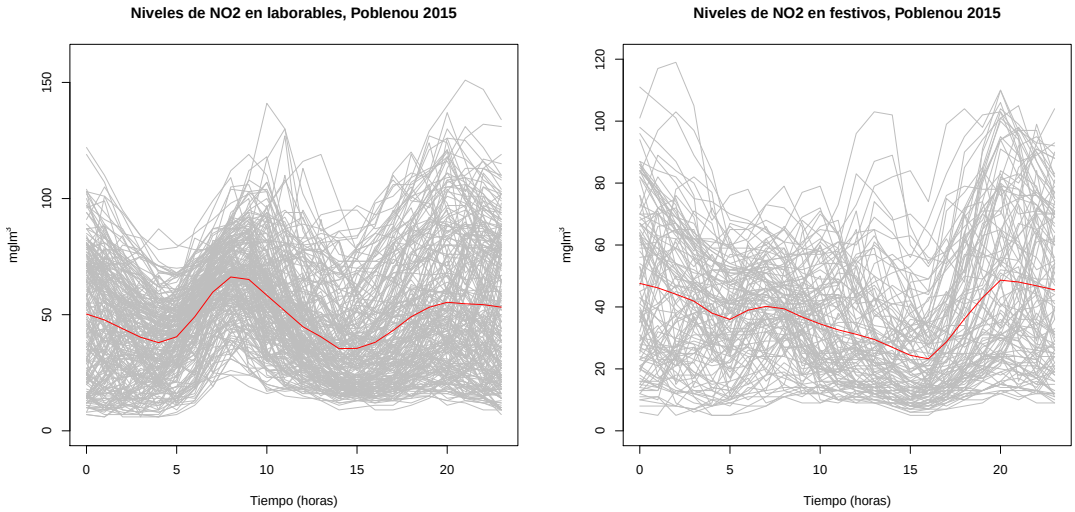

Figura 10: Niveles de NO_2 en laborables y festivos, Poblenou 2015

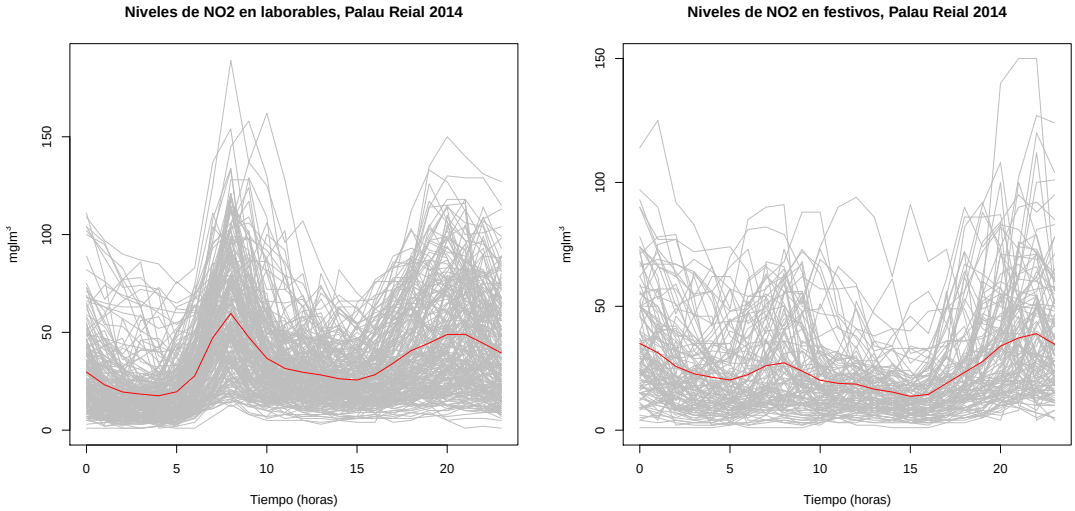

Figura 11: Niveles de NO_2 en laborables y festivos, Palau Reial 2014

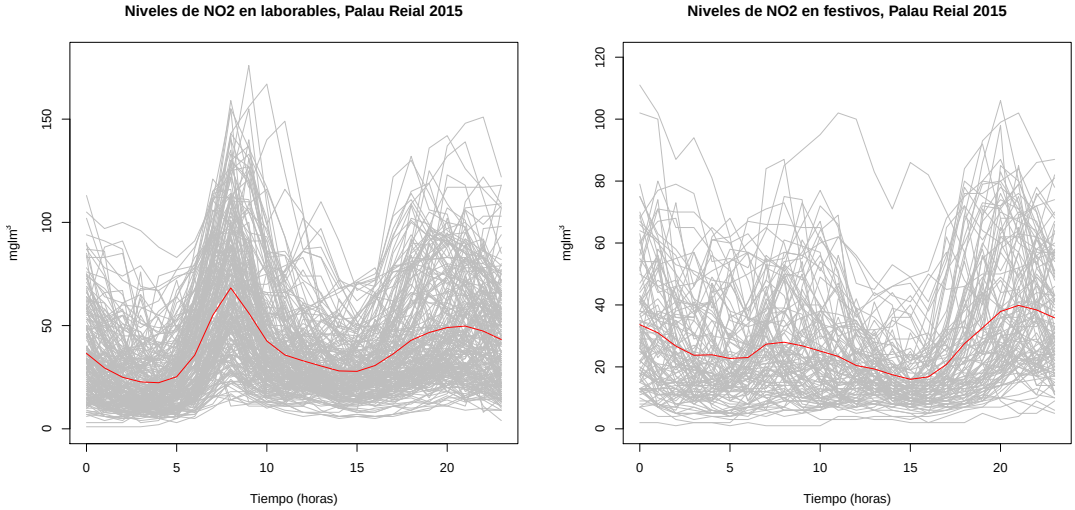

Figura 12: Niveles de NO_2 en laborables y festivos, Palau Reial 2015

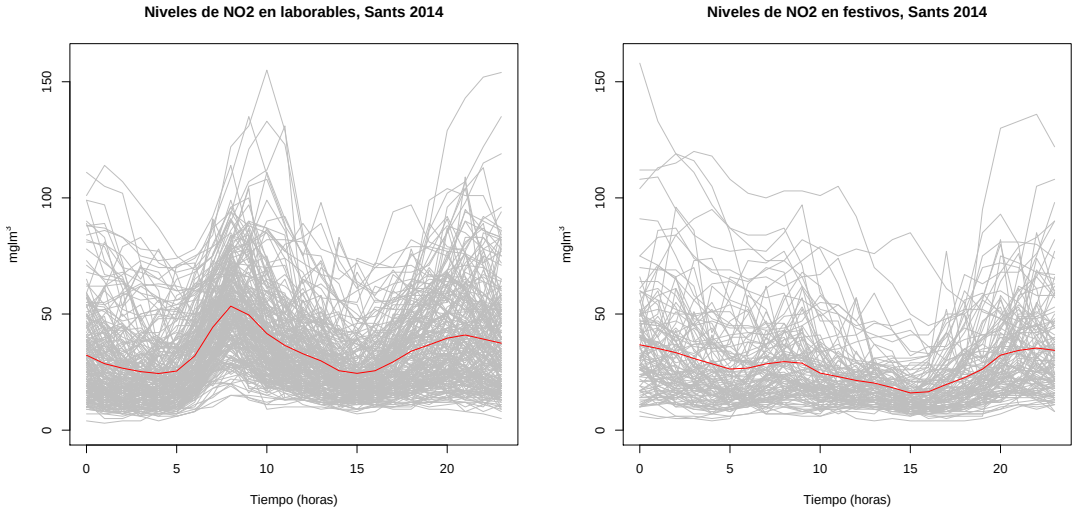

Figura 13: Niveles de NO_2 en laborables y festivos, Sants 2014

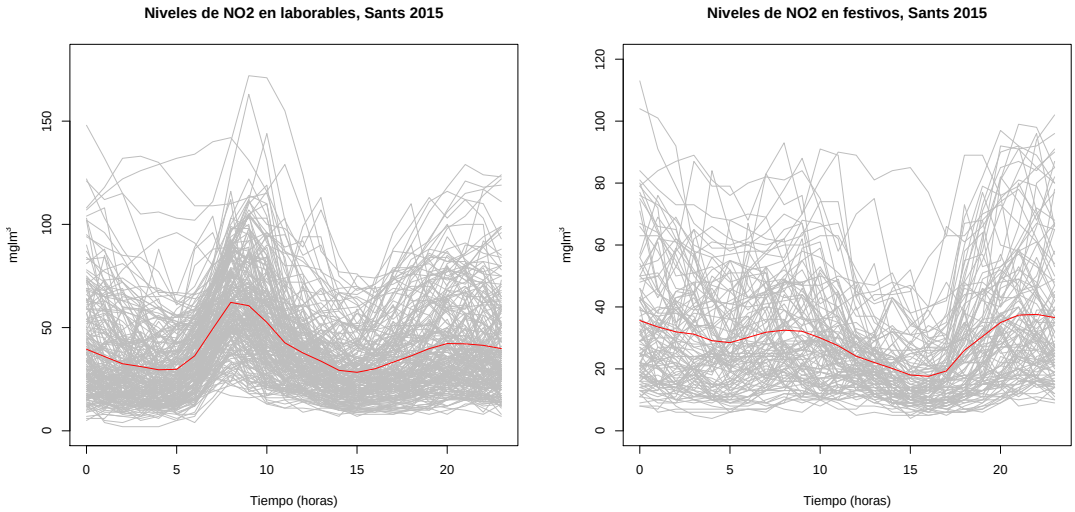

Figura 14: Niveles de NO_2 en laborables y festivos, Sants 2015