Trabajo Fin de Máster: Smart Bike - Sevilla.

Máster en Data Science y Big Data - Univ. de Sevilla. 2016/17.

Jerónimo Carranza Carranza

. .

Directores: Gonzalo A. Aranda Corral y Joaquín Borrego Díaz

22 de marzo de 2018

Introducción

Situación

Los ***Sistemas de Bicicletas Compartidas*** (Bycicle Sharing System) ponen a disposición de un grupo de usuarios una serie de bicicletas para que sean utilizadas temporalmente como medio de transporte. El usuario sólo necesita tener la bicicleta en su posesión durante el desplazamiento.

Los sistemas de bicicletas compartidas son un modo de movilidad urbana que se ha extendido de forma muy notable en ciudades de todo el mundo y con una gran aceptación de público. Existen más de 1400 sistemas activos en el mundo, con una flota operativa de más de 14 millones de bicicletas.

Esquemas de uso compartido de bicicletas:

1. *No regulado*
2. *Depósito*
3. *Afiliación*
4. *Sin estaciones*

Justificación

En los sistemas de afiliación con estaciones, los más extendidos a nivel mundial, cada estación está dotada con una serie de **sensores** que indican en tiempo real el número de bicicletas que se pueden retirar así como el total de plazas libres en las que se pueden devolver.

Con el objetivo de disponer de **datos históricos de múltiples sistemas** de bicicletas compartidas para el desarrollo de estudios diversos sobre los mismos, investigadores de las **universidades de Sevilla y Huelva** llevan algún tiempo almacenando datos sobre el uso de bicicletas públicas de 27 ciudades europeas. En concreto de los datos instantáneos publicados por la empresa **JCDecaux**.

Objetivos

Se plantea como objetivo de este trabajo, el análisis y modelado de los datos históricos recopilados del sistema de bicicletas compartidas de la ciudad de Sevilla (Sevici), y más concretamente:

1. Identificar patrones espacio-temporales del uso de bicicletas de Sevici.
2. Predecir índices de ocupación de las estaciones de Sevici.

Metodología

Descripción del conjunto de datos

Sevici. Datos dinámicos.

Campo:	Descripción:
id	Id registro autonumérico
status	Estado de la estación; OPEN o CLOSED
contract	Contrato, en nuestro caso; Seville
num	Número de la estación
last_update	Momento de última actualización
add_date	Fecha-Hora en fracciones de 5 minutos
stands	Número de estacionamientos operativos en la estación
availablestands	Número de estacionamientos disponibles
availablebikes	Número de bicicletas operativas y disponibles

Sevici. Datos estáticos.

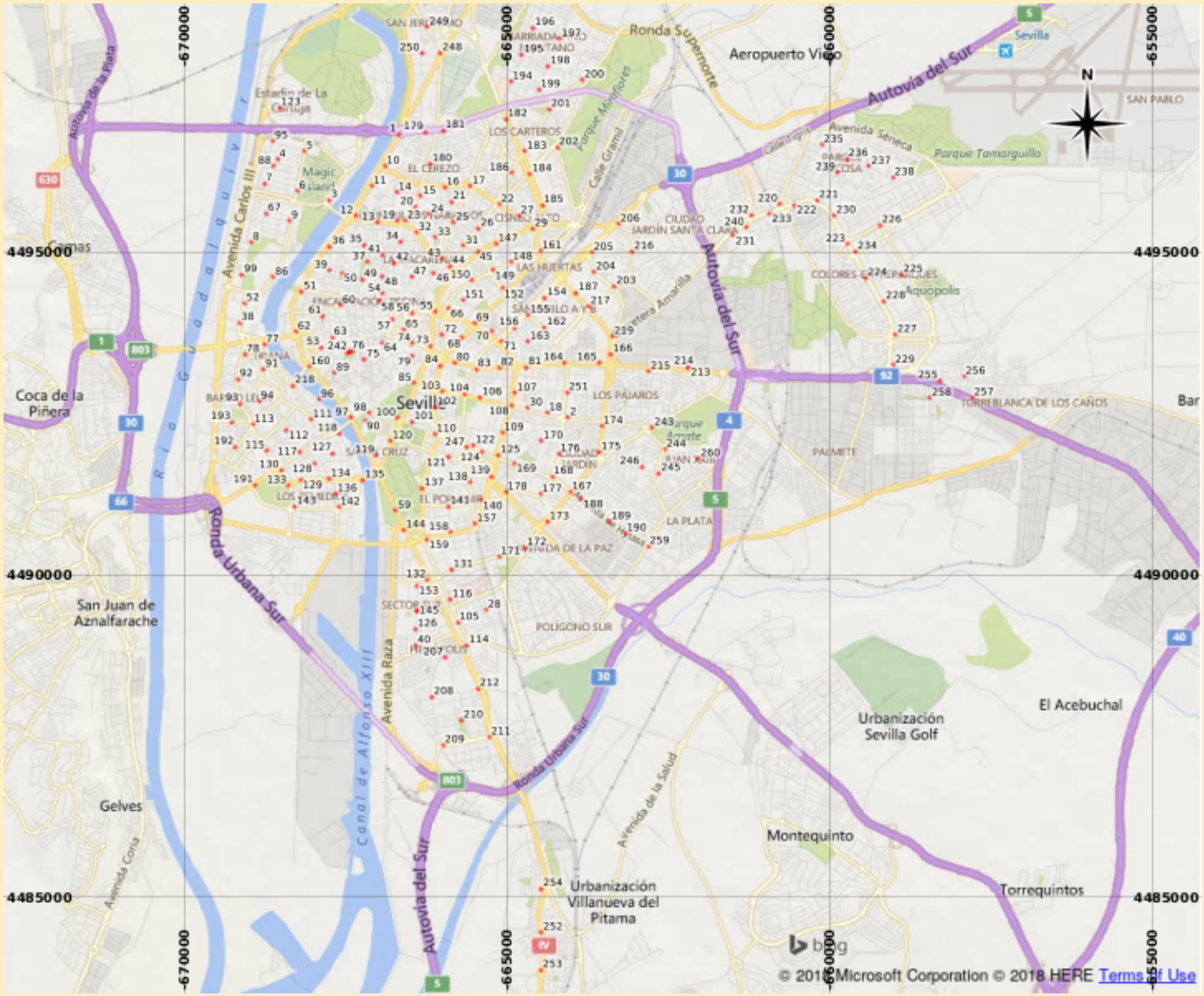
Campo:	Descripción:
Number	Número de la estación
Name	Nombre de la estación
Address	Dirección
Latitude	Latitud (grados WGS84)
Longitude	Longitud (grados WGS84)

Datos Meteorológicos, Calendario de festivos.

Campo:	Descripción:	Campo:	Descripción:
fecha	Fecha	fecha	Fecha del día festivo
p	Precipitación total	festivo	Festivo
tmax	Temperatura máxima		
tmin	Temperatura mínima		

Pretratamiento y depuración

Localización de estaciones



Datos replicados

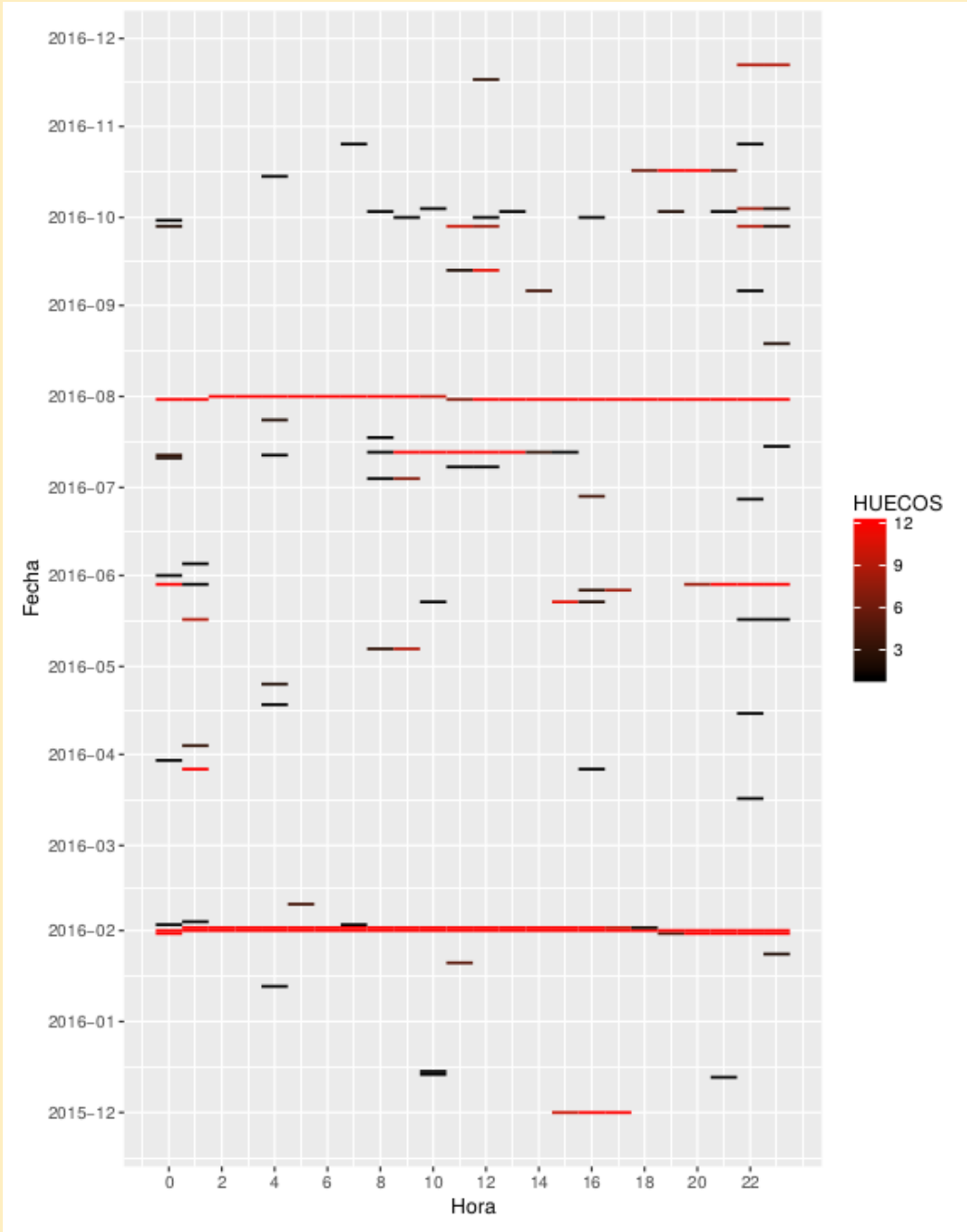
Todas las estaciones, salvo la estación 109, presentan 12 datos duplicados que se concentran en un día, entre 2016-10-30 02:00:01 y 2016-10-30 02:55:01. (6216).

Datos faltantes

Para explorar los huecos de datos faltantes se construye una serie entre inicio y fin con paso de 5min y se vincula al minuto con la secuencia real de *add_date* allí donde exista.

Variable	Valor
Número de huecos	936353
Número de huecos en estación 109	81533
Número de huecos sin estación 109	854820
Número de p5min con huecos globales	962
Número de huecos globales (sin e109)	249158
Número de huecos específicos (sin e109)	605662

Huecos globales por fecha y hora



Datos anómalos

Cuadro5: Resumen de incidencias por datos duplicados o anómalos

ok	Descripción	N
1	Sin incidencia aparente	22094898
2	Dato duplicado	3108
3	Estacionamientos disponibles $>$ Est. operativos	954
4	Bicis disponibles $>$ Estac. operativos	965
5	Estac. $+$ Bicis disponibles $>$ Estac. operativos	5728
6	Estac. $+$ Bicis disponibles $<$ Estac. operativos	4367165

Tanto si responde a retrasos puntuales como al no correcto registro de los momentos de inoperatividad, la consideración del número efectivo de estacionamientos operativos como la suma de estacionamientos y bicicletas disponibles siempre que ésta sea menor o igual al número registrado de estacionamientos operativos (nominal) permite tener en consideración estas situaciones (ok=6) como registros válidos y tratarlos de forma conjunta a los de la situación sin incidencia aparente (ok=1).

Análisis exploratorio

Se consideran finalmente como **datos válidos**, entre los datos dinámicos Sevici, todos aquellos provenientes de registros no duplicados en los que la suma de estacionamientos y bicis disponibles es menor o igual que el número (nominal) de estacionamientos operativos. (ok=1 o ok=6). Son datos válidos globales la agregación de datos válidos entre todas las estaciones en un momento determinado.

Se ha realizado el análisis exploratorio de los datos válidos, tanto datos globales como por estaciones. Básicamente **análisis gráfico y resúmenes estadísticos** con respecto a agregados temporales y variables meteorológicas.

Identificación de patrones espacio-temporales

Para el análisis de clasificación de estaciones e identificación de patrones se parte de *sevicip5m*. Este dataframe tiene las 105132 filas correspondientes a los periodos de cinco minutos entre inicio y fin y las 522 columnas correspondientes a:

Cuadro6: Estructura *sevicip5m*

Variable	Descripción
p5min	Periodo de 5min (Datetime)
hueco	Hueco global (Boolean)
si	Estacionamientos disponibles estación i
bi	Bicicletas disponibles estación i
:	para i en 1:260

A partir de estos datos se ha calculado la **matriz de correlación (Pearson)** entre estaciones para la variable número de bicicletas disponibles. Dada la presencia de datos faltantes en gran número se ha optado por utilizar para cada celdilla de la matriz los casos en que existen datos disponibles para el par de variables (pairwise).

La matriz de correlación obtenida se ha segregado en un dataframe con todos los pares y el valor de correlación, para su tratamiento posterior como **grafo con nodos geoposicionados**.

Se utiliza la matriz de correlación como base para la clasificación de las estaciones. Para ello en primer lugar convertimos los coeficientes de correlación en disimilaridades y éstas son tratadas como distancias. Se realiza un **análisis cluster jerárquico** con la matriz de distancias así obtenida. En dicho análisis se han comparado los resultados obtenidos para los distintos métodos de agregación del paquete ***hclus***. Se opta finalmente por el método '*complete*', que visualmente muestra mayor coherencia espacial. En este método de agregación, la distancia entre dos clusters se define como la máxima distancia entre sus componentes individuales.

Unas vez clasificadas las estaciones según el método descrito, se ha obtenido la estadística básica del número de bicis disponibles por clase de estación, día de la semana y hora del día y su representación gráfica.

Finalmente se contrasta la validez de los **patrones espacio-temporales** identificados mediante la construcción de un modelo lineal general con las variables independientes clase de estación, día de la semana y hora del día y variable dependiente el número de bicicletas disponibles.

Los datos para construir el modelo no son los datos completamente desagregados sino que se utilizan las medias del número de bicicletas disponibles por estación, fecha y hora. Este conjunto de datos tiene más de 2 millones de registros.

Modelos predictivos

Se considera en los modelos que para cada estación (i) en un momento determinado (t), el número de bicicletas disponibles $(Y(i,t))$ es una función lineal de:

- los valores de dicha variable en esa estación en momentos anteriores:

$$Y(i, t - \delta_t), \delta_t \in \{15min, 30min, 1h, 4h, 8h, 24h\}$$

- los valores de dicha variable en la estación más cercana a ella (j) en momentos anteriores:

$$Y(j, t - \delta_t), \delta_t \in \{15min, 30min, 1h, 4h, 8h, 24h\}$$

- el día de la semana que es t: $DSEM$,
- la hora del día del momento t: $HORA$,
- si es día festivo: $FEST$,
- si es fin de semana o festivo: $FSOF$,
- temperatura máxima del día: $TMAX$,
- temperatura mínima del día: $TMIN$,
- precipitación total del día: P

Los modelos predictivos desarrollados son **modelos de regresión con regularización Elasticnet**.

Se toman para el modelado sólo los **casos completos** existentes en el conjunto de datos, lo que supone **52543 casos para 1834 variables** (originales y retardadas). No se utilizan los datos de la estación 109, que sólo dispone de datos durante los tres primeros meses de registro.

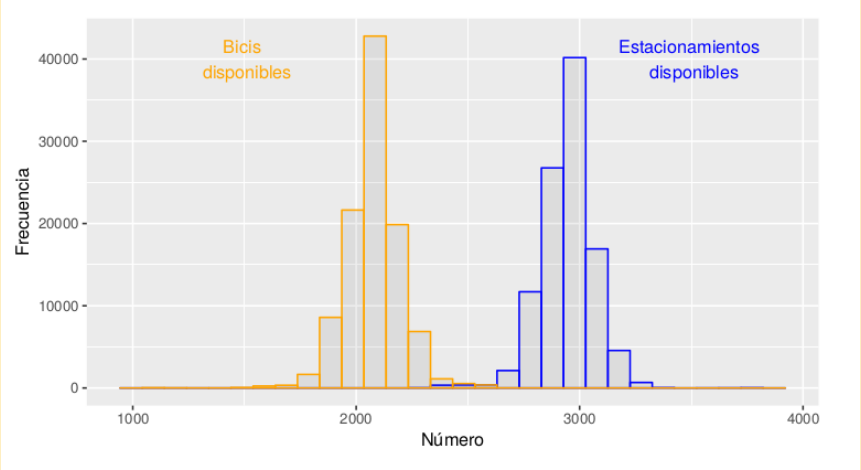
Se obtienen muestras relativamente pequeñas para **training (14 %)** y **test (6 %)** a partir de una división del conjunto de datos con **fecha de corte 2016-09-15**, que deja aproximadamente el 70 % de las observaciones a su izquierda (anteriores) y aproximadamente el 30 % a su derecha (posteriores). Se garantiza así que toda la muestra test sea posterior a la muestra de entrenamiento.

Se utilizan modelos con optimización de parámetros por **validación cruzada** según implementa el paquete de R ***glmnet***.

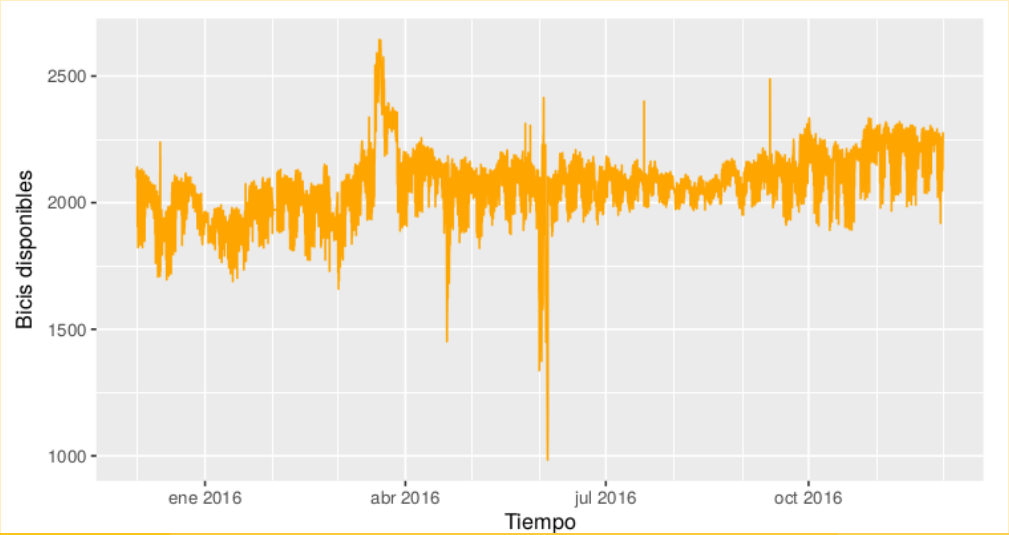
Principales resultados y conclusiones

Análisis exploratorio

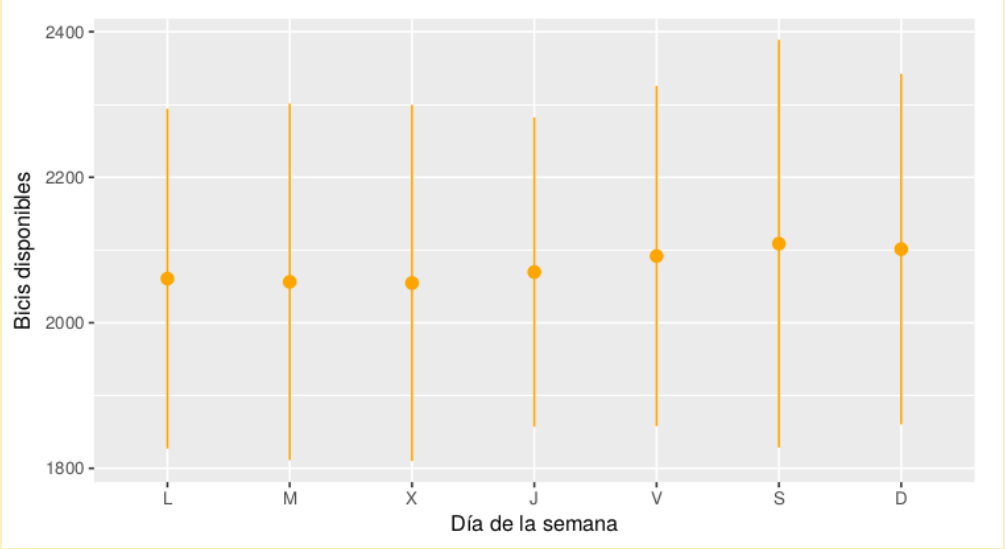
Distribución de estacionamientos y bicis totales disponibles



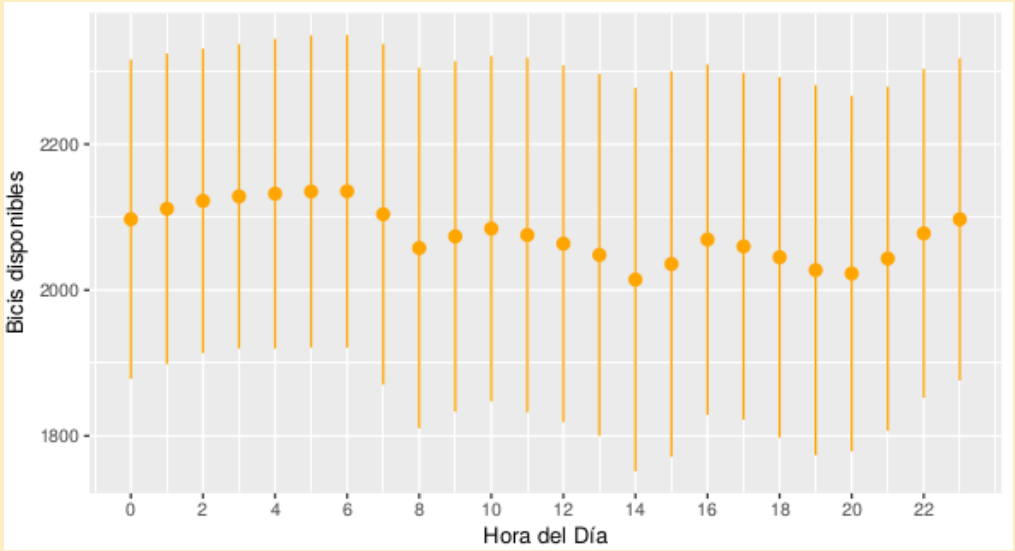
Serie de bicis totales disponibles



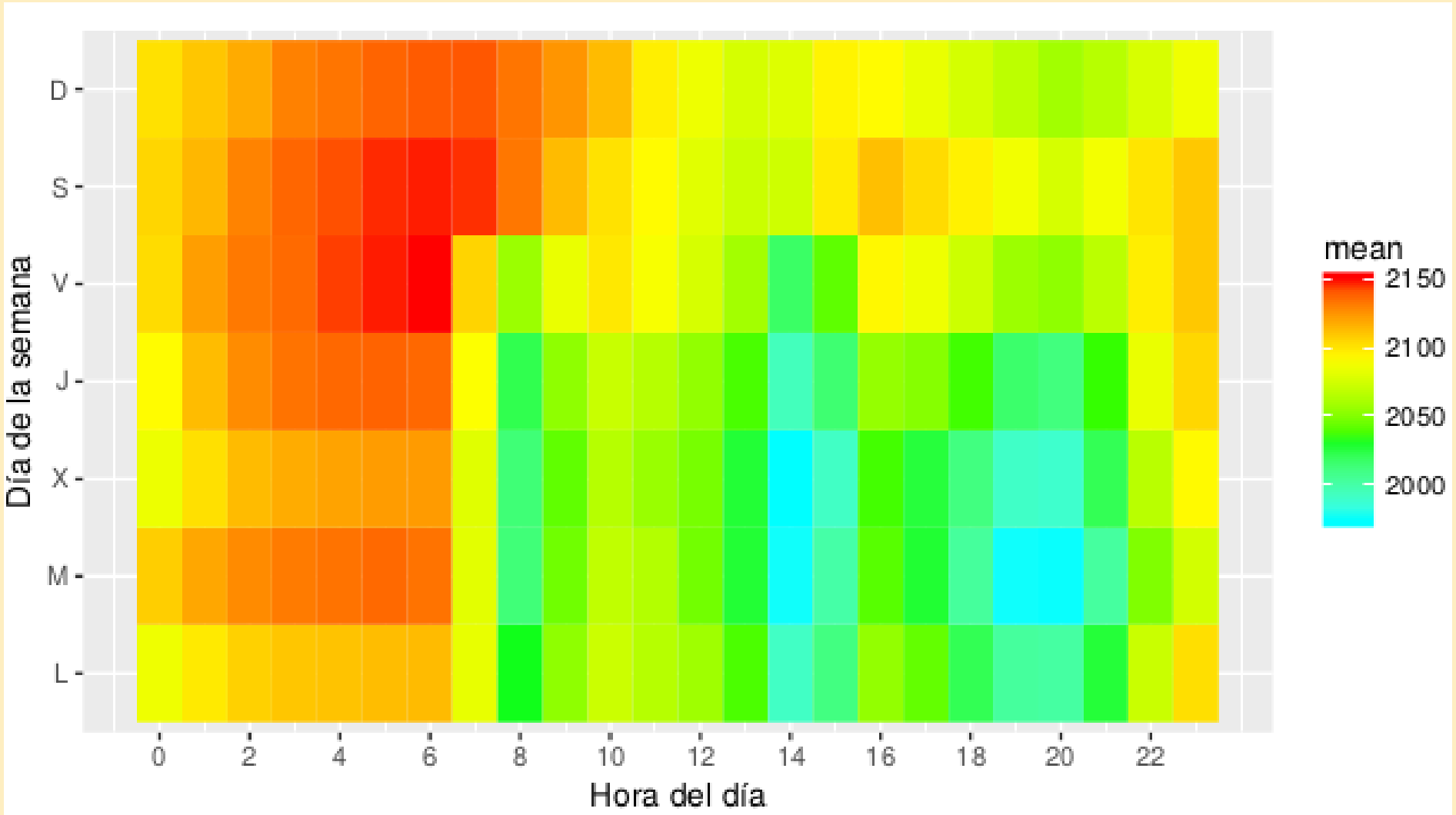
Análisis según días de la semana



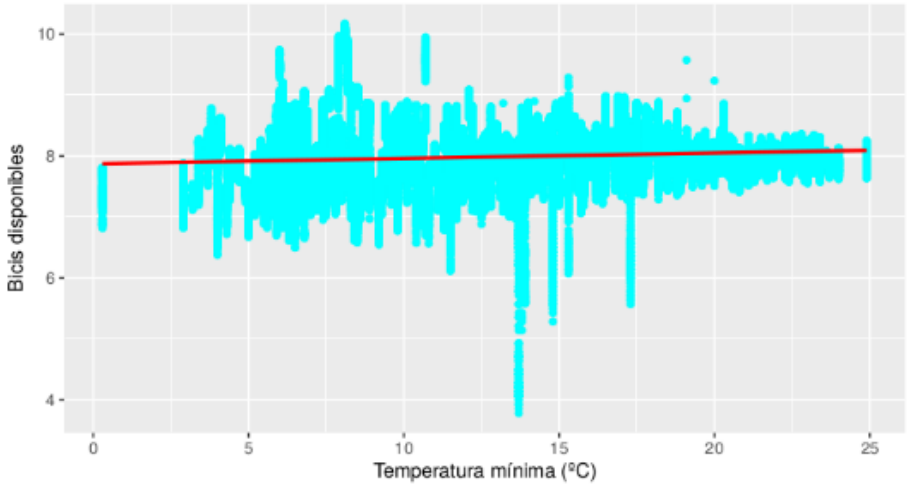
Análisis según hora del día

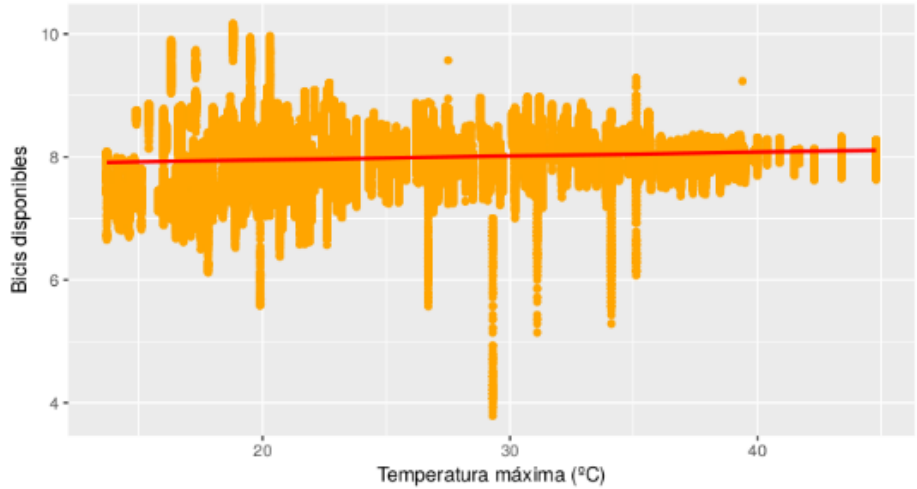


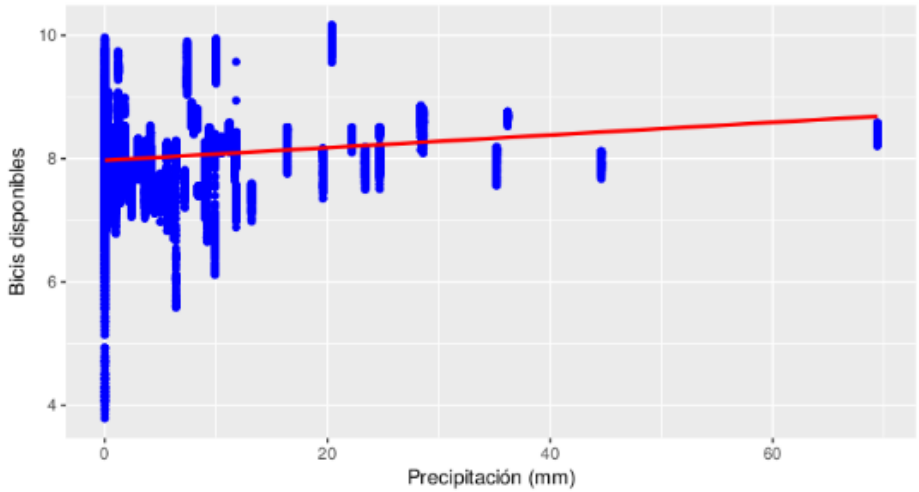
Análisis según hora del día y día de la semana



Análisis según condiciones meteorológicas

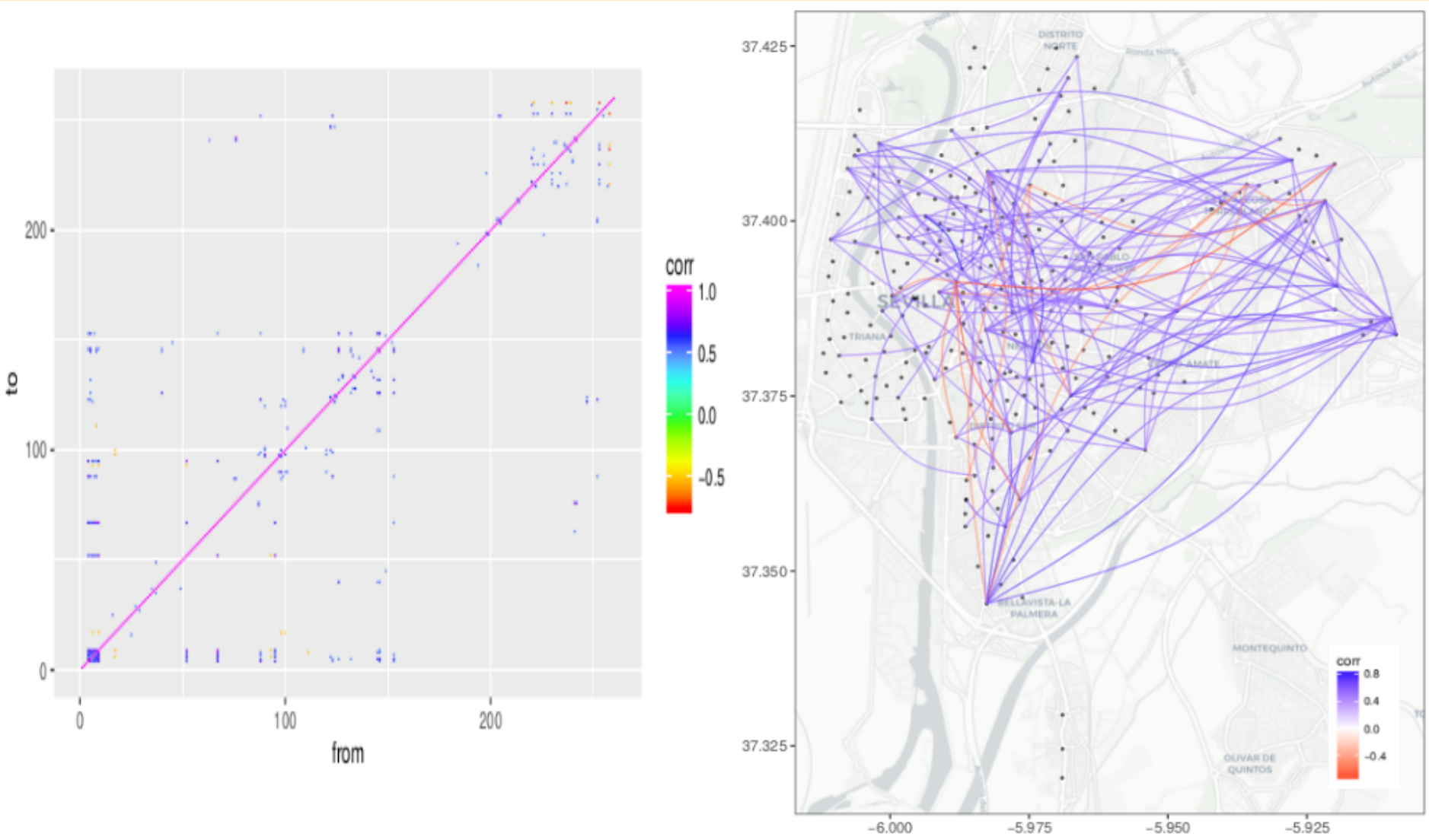




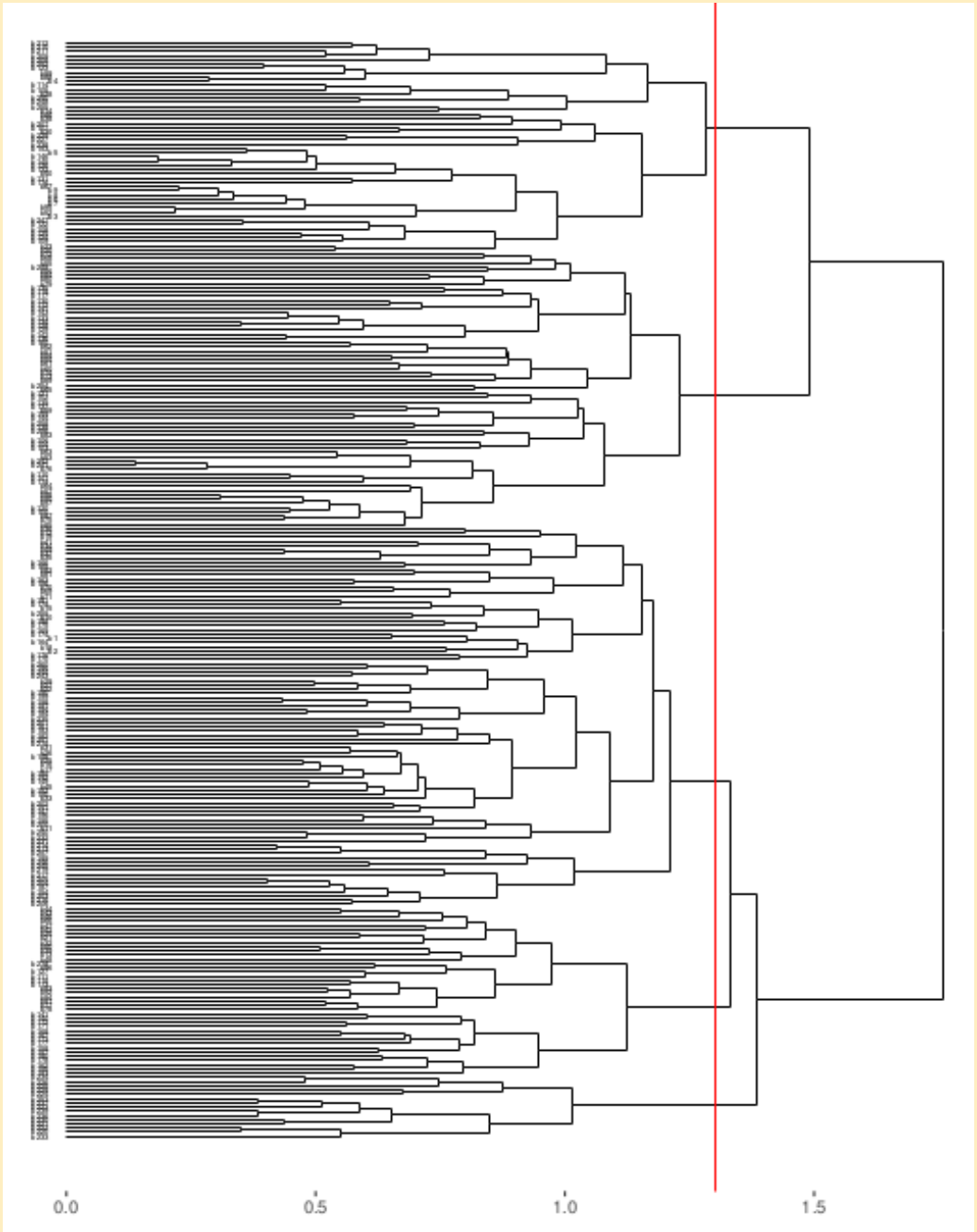


Identificación de patrones espacio-temporales

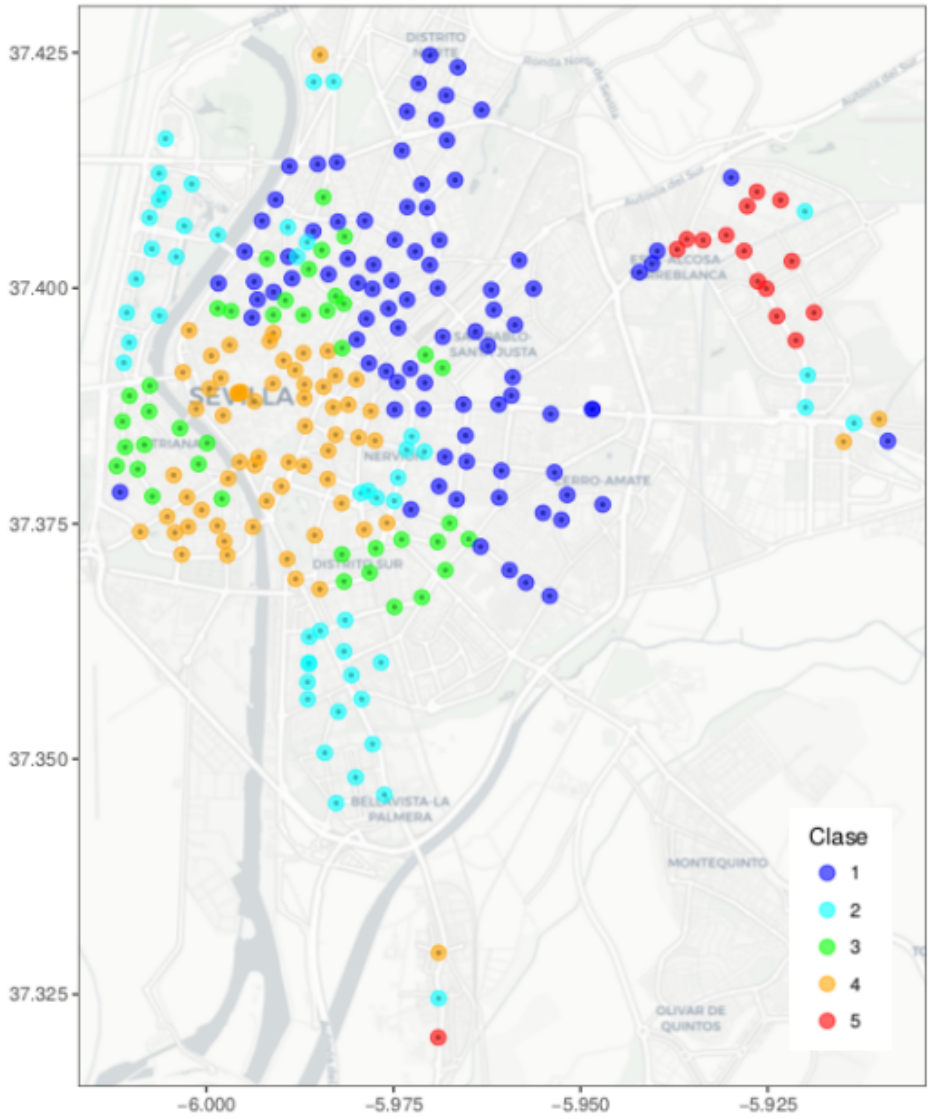
Matriz y grafo de correlaciones entre estaciones ($|corr| > 0.5$).

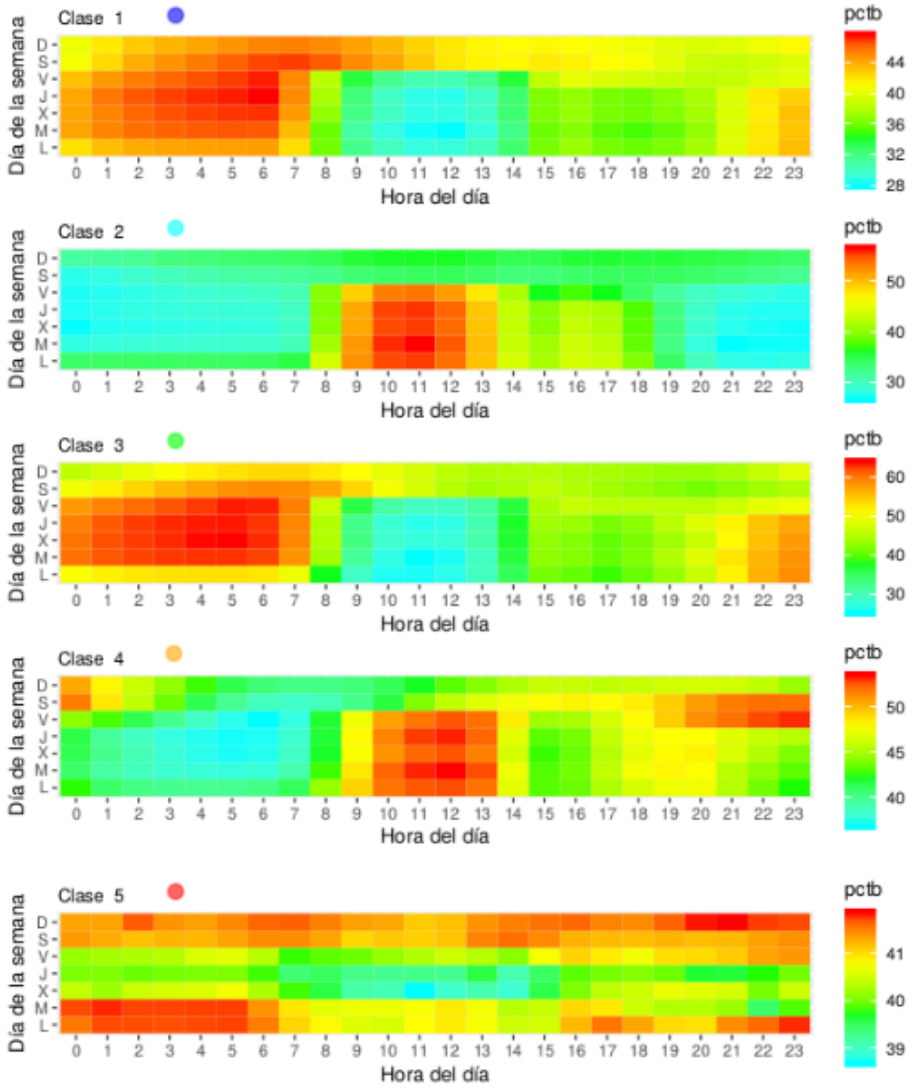


Dendrograma de clasificación de las estaciones



Patrones espacio-temporales





La distribución por día de la semana y hora del día del porcentaje de bicis disponibles entre las distintas clases de estaciones muestra:

1. Los patrones para las clases 1 y 4 son claramente complementarios, correspondiendo la clase 1 a estaciones con concentración de bicicletas disponibles todos los días de noche y madrugada y la clase 4 a estaciones con máxima presencia de bicis disponibles entre las 9:00 y las 13:00 horas de Lunes a Viernes. Lo que se correspondería a desplazamientos entre residencia (1) y trabajo o estudio (4).
2. Los patrones para las clases 2 y 3 son igualmente complementarios y muy parecidos a los indicados para 4 y 1, respectivamente.
3. La clase 5 presenta un comportamiento temporal bien distinto al de las otras clases, con máximos en las madrugadas de lunes y martes, niveles relativamente altos durante todo el fin de semana, y mínimos en la parte central del día de los días centrales de la semana (X,J,V).

4. Las clases 2 y 4 aunque presentan patrones generales muy parecidos, según lo dicho, se diferencian sobre todo por su comportamiento los viernes y sábados a partir de las 20:00, con niveles altos en la clase 4, posiblemente inducida por desplazamientos a actividades de tipo lúdico.
5. La distinción entre los patrones 1 y 3 está vinculada al comportamiento en fin de semana relacionada con la extensión de niveles altos hasta horas más tardías en la clase 1.

Modelos predictivos

Persistencia de modelos

```
> head(modelos)
  id  modelo vi   vj     lambda          a0 df        r2      RMSE    R2test
1  2 GLMNET0 b1 b179 0.2433850   0.5907011  4 0.9778934  5.190999 0.9682191
2  3 GLMNET1 b1 b179 0.5555811   1.2567319  5 0.9540914  7.321472 0.9368288
3  4 GLMNET2 b1 b179 1.3720174   2.9205337  4 0.9043455 10.114414 0.8804563
4  5 GLMNET3 b1 b179 1.7024080   5.6792678  7 0.6711944 18.023582 0.6169176
5  6 GLMNET4 b1 b179 2.5113937   9.7058241  6 0.4913999 21.980414 0.4312904
6  7 GLMNET5 b1 b179 1.8018208  12.4920250  9 0.3386743 25.230830 0.2498336

> head(betas)
  id  modelo vi   vj  nom beta
1  2 GLMNET0 b1 b179 hora    0
2  2 GLMNET0 b1 b179  lun    0
3  2 GLMNET0 b1 b179  mar    0
4  2 GLMNET0 b1 b179  mie    0
5  2 GLMNET0 b1 b179  jue    0
6  2 GLMNET0 b1 b179  vie    0

> head(residuos)
  id  modelo vi    residuo
1  2 GLMNET0 b1   3.602684
2  2 GLMNET0 b1  0.7217409
3  2 GLMNET0 b1 -0.7253595
4  2 GLMNET0 b1  -2.075419
5  2 GLMNET0 b1 -0.4769004
6  2 GLMNET0 b1  0.4898732
```

Cuadro7: Campos en la estructura de persistencia de modelos

Campo	Descripción
id	Número de modelo
modelo	Nombre del modelo GLMNET0..6
vi	Identificador de estación objetivo
vj	Identificador de estación más cercana
lambda	Parámetro ajustado equilibrio regularización L1 - L2
a0	Intersect del modelo
df	Número de coeficientes no nulos
r2	R cuadrado ajuste
RMSE	Raíz cuadrada del error cuadrático medio (test)
R2test	R cuadrado (test)
nom	Identificador del regresor
beta	Coeficiente del regresor en el modelo
residuo	Residuo (test): Y-predY

Bondad de ajuste de los modelos

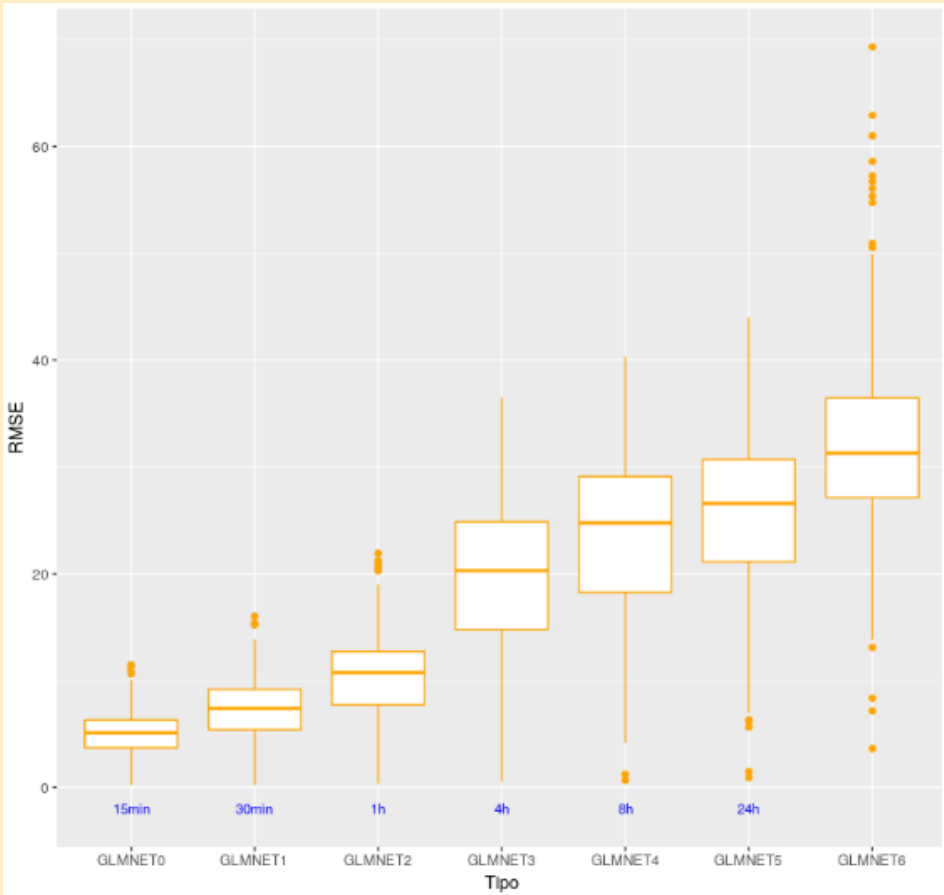
Bondad de ajuste global de los modelos

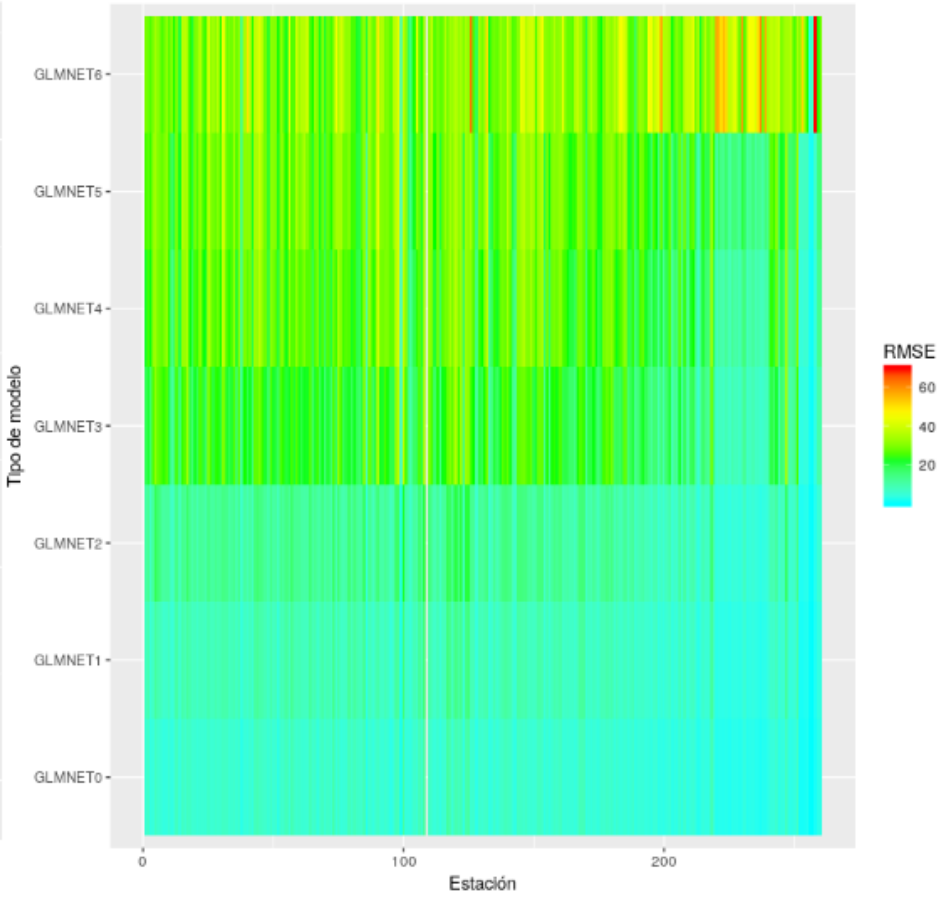
r2_median	R2test_median	RMSE_median	r2_mean	R2test_mean	RMSE_mean
0.7843	0.6775	15.511	0.6482	0.5486	17.5913

Bondad de ajuste por tipo de modelo

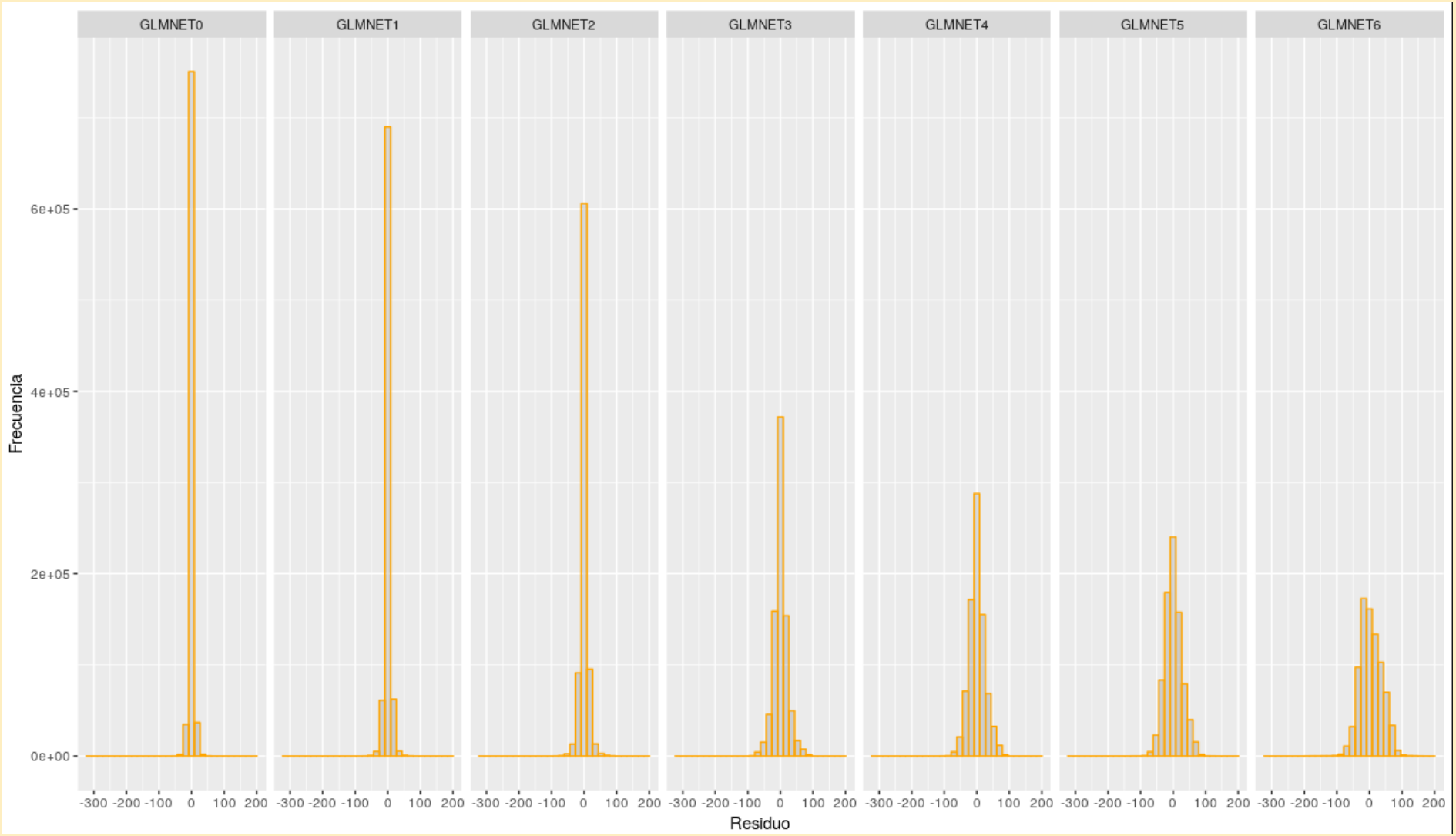
modelo	r2_median	R2test_median	RMSE_median	r2_mean	R2test_mean	RMSE_mean
GLMNET0	0.9774	0.9677	5.1183	0.9741	0.9623	5.0735
GLMNET1	0.9546	0.9320	7.3979	0.9490	0.9234	7.2812
GLMNET2	0.9066	0.8551	10.7559	0.8992	0.8463	10.4202
GLMNET3	0.6584	0.4512	20.2921	0.6613	0.4976	19.3752
GLMNET4	0.4778	0.2180	24.7588	0.5115	0.3212	23.1190
GLMNET5	0.3419	0.1132	26.5877	0.4090	0.2304	25.3272
GLMNET6	0.0971	0.0262	31.2985	0.1329	0.0472	32.5427

Bondad de ajuste de los modelos





Residuos


GLMNET0
GLMNET1
GLMNET2
GLMNET3
GLMNET4
GLMNET5
GLMNET6
Frecuencia
6e+05
4e+05
2e+05
0e+00
-300 -200 -100 0 100 200
Residuo

Regresores significativos