

# Trabajo Fin de Máster: Smart Bike - Sevilla.

## Memoria.

Máster en Data Science y Big Data - Universidad de Sevilla, 2016/2017.

*Jerónimo Carranza Carranza*

*1 de marzo de 2018*

## Índice

<b>1. Introducción</b>	<b>3</b>
1.1. Situación . . . . .	3
1.2. Justificación . . . . .	3
1.3. Objetivos . . . . .	4
<b>2. Resumen ejecutivo</b>	<b>5</b>
<b>3. Metodología</b>	<b>12</b>
3.1. Descripción del conjunto de datos. Obtención y carga de datos . . . . .	12
3.1.1. Datos dinámicos Sevici . . . . .	12
3.1.2. Datos estáticos Sevici . . . . .	12
3.1.3. Datos meteorológicos . . . . .	13
3.1.4. Calendario de festivos . . . . .	13
3.2. Pretratamiento y depuración . . . . .	13
3.2.1. Localización de estaciones . . . . .	13
3.2.2. Datos replicados . . . . .	17
3.2.3. Datos faltantes . . . . .	17
3.2.4. Datos anómalos . . . . .	19
3.3. Análisis exploratorio . . . . .	19
3.4. Clasificación de estaciones e identificación de patrones espacio-temporales . . . . .	20
3.5. Modelos predictivos . . . . .	20
<b>4. Principales resultados y conclusiones</b>	<b>22</b>
4.1. Análisis exploratorio . . . . .	22
4.1.1. Datos meteorológicos . . . . .	22
4.1.2. Análisis de datos válidos globales . . . . .	23
4.1.2.1. Análisis según días de la semana y festivos . . . . .	25
4.1.2.2. Análisis según hora del día . . . . .	27
4.1.2.3. Análisis según hora del día y día de la semana . . . . .	28
4.1.2.4. Análisis según condiciones meteorológicas . . . . .	29
4.2. Clasificación de estaciones e identificación de patrones espacio-temporales . . . . .	31
4.2.1. Análisis de correlación entre estaciones . . . . .	31
4.2.2. Clasificación de las estaciones . . . . .	32
4.2.3. Patrones espacio-temporales . . . . .	36
4.3. Modelo predictivo . . . . .	37
4.3.1. Persistencia de modelos . . . . .	37
4.3.2. Bondad de ajuste de los modelos . . . . .	38
4.3.3. Residuos . . . . .	40
4.3.4. Regresores significativos . . . . .	41

## Índice de cuadros

1.	Sevici. Datos dinámicos . . . . .	12
2.	Sevici. Datos estáticos . . . . .	12
3.	Estructura de <i>Meteo</i> . . . . .	13
4.	Estructura de <i>Festivos</i> . . . . .	13
5.	Resumen de huecos en datos dinámicos . . . . .	17
6.	Resumen de incidencias por datos duplicados o anómalos . . . . .	19
7.	Estructura <i>sevicip5m</i> . . . . .	20
8.	Resumen estadístico de variables meteorológicas . . . . .	22
9.	Bicis disponibles por día de la semana. Estadística básica. . . . .	25
10.	Bicis disponibles según sea fin de semana - festivo o no. Estadística básica. . . . .	26
11.	Bicis disponibles por hora del día. Estadística básica. . . . .	27
12.	Campos en la estructura de persistencia de modelos . . . . .	38
13.	Bondad de ajuste global de los modelos. . . . .	38
14.	Bondad de ajuste por tipo de modelo. . . . .	38

## Índice de figuras

1.	Total de Bicis disponibles en el conjunto de estaciones por hora del día. . . . .	5
2.	Total de Bicis disponibles en el conjunto de estaciones por hora del día y día de la semana. . . . .	6
3.	Mapa de estaciones clasificadas. . . . .	7
4.	% Bicis disponibles por hora del día y día de la semana. Patrones según clase de estación. . . . .	8
5.	Bondad de ajuste. Raíz del error cuadrático medio (RMSE) por tipo de modelo. . . . .	10
6.	Distribución de residuos por tipo de modelo. . . . .	11
7.	Mapa de Estaciones Sevici sobre BingMap. . . . .	14
8.	Localización de estaciones SEVICI sobre OSM - 'hikebike' . . . . .	15
9.	Localización de estaciones SEVICI sobre OSM - 'cartolight' . . . . .	16
10.	Datos faltantes. Huecos Globales por Fecha y Hora . . . . .	18
11.	Datos meteorológicos. Serie de precipitación total diaria. . . . .	22
12.	Datos meteorológicos. Series de temperaturas máximas y mínimas diarias. . . . .	23
13.	Datos válidos globales. Bicis disponibles. . . . .	24
14.	Datos válidos globales. Distribución de Estacionamientos y Bicis disponibles. . . . .	24
15.	Datos válidos globales. Diferencia Estacionamientos y Bicis disponibles. . . . .	25
16.	Datos válidos globales. Bicis disponibles por día de la semana. Media +/- 2 · Desviación . . . . .	26
17.	Datos válidos globales. Bicis disponibles según sea fin de semana - festivo o no. . . . .	27
18.	Datos válidos globales. Bicis disponibles según hora del día. . . . .	28
19.	Datos válidos globales. Bicis disponibles según hora del día y día de la semana. . . . .	29
20.	Datos válidos globales. Bicis disponibles según precipitación total diaria. . . . .	29
21.	Datos válidos globales. Bicis disponibles según temperatura mínima diaria. . . . .	30
22.	Datos válidos globales. Bicis disponibles según temperatura máxima diaria. . . . .	30
23.	Matriz de correlación ( $ corr  > 0.5$ ) entre estaciones. . . . .	31
24.	Grafo espacial de correlaciones $ corr  > 0.5$ . . . . .	32
25.	Dendrograma de estaciones basado en correlación. . . . .	33
26.	Mapa de estaciones clasificadas . . . . .	34
27.	% de Bicis disponibles por hora del día y día de la semana. Patrones por clase de estación. . . . .	36
28.	Bondad de ajuste. Raíz del error cuadrático medio (RMSE) por tipo de modelo. . . . .	39
29.	Errores (RMSE) por estación y tipo de modelo . . . . .	40
30.	Distribución de residuos por tipo de modelo. . . . .	41
31.	Frecuencia de modelos con regresor significativo por tipo de modelo. . . . .	42

## 1. Introducción

### 1.1. Situación

Los **Sistemas de Bicicletas Compartidas** (Bicycle Sharing System), también conocidos como sistemas de bicicletas públicas, ponen a disposición de un grupo de usuarios una serie de bicicletas para que sean utilizadas temporalmente como medio de transporte. Normalmente estos sistemas son gestionados por un estamento público y permiten recoger una bicicleta y devolverla en un punto diferente, para que el usuario sólo necesite tener la bicicleta en su posesión durante el desplazamiento.<sup>1</sup>

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.<sup>2</sup>

Los sistemas para compartir bicicletas han sufrido cambios que pueden clasificarse en tres fases clave o generaciones. Estos incluyen la primera generación, llamada *bicicletas blancas* o bicicletas gratuitas (Amsterdam, 1965); la segunda generación de sistemas de *depósito de monedas* (Copenhagen, 1995); y la tercera generación, o sistemas basados en tecnología de la información o *smart bikes* (Rennes, 1998). Las recientes mejoras tecnológicas y operacionales también están allanando el camino para una cuarta generación, conocida como sistema multimodal sensible a la demanda.<sup>3 4</sup>

Todos los esquemas de uso compartido de bicicletas que se han desarrollado a lo largo de los años están basados en uno o más de los siguientes sistemas:

*No regulado*: En este tipo de programa, las bicicletas se distribuyen simplemente en una ciudad o área determinada para que las use cualquier persona.

*Depósito*: Un pequeño depósito en efectivo libera la bicicleta de una terminal bloqueada y solo se puede recuperar devolviéndola a otra.

*Afiliación*: En esta versión del sistema, las bicicletas se guardan en centros operados por personal o en terminales de autoservicio distribuidos en toda la ciudad. Las personas registradas en el servicio se identifican con su tarjeta de usuario (o con una tarjeta inteligente, por teléfono celular u otros métodos) en cualquiera de los centros para hacer uso de una bicicleta por un período corto de tiempo, por lo general de tres horas o menos. En muchos esquemas, la primera media hora es gratis. El individuo es responsable de cualquier daño o pérdida hasta que la bicicleta se devuelva a otro centro y se registre. Muchos de estos sistemas de afiliación se operan a través de asociaciones de instituciones públicas con empresas privadas (JCDecaux, Clear Channel, Smoove, etc.).

*Sin estaciones*: Los sistemas de bicicletas sin estaciones están diseñados para que un usuario no necesite devolver la bicicleta a un centro o un punto de anclaje en una estación; sino que, el próximo usuario puede encontrarlo por GPS. Este tipo de sistemas se han desarrollado de forma muy rápida, sobre todo en China desde 2010, impulsadas por la iniciativa privada (Mobike, Ofo, etc.).<sup>5</sup>

### 1.2. 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. El estado actual del sistema, es decir el estado de todas sus estaciones, es publicado por la entidad que lo gestiona a través de servicios web, por lo que pueden ser

<sup>1</sup>Wikipedia. [https://es.wikipedia.org/wiki/Sistema\\_de\\_bicicletas\\_compartidas](https://es.wikipedia.org/wiki/Sistema_de_bicicletas_compartidas). Consultado 21 de febrero de 2018.

<sup>2</sup><http://www.iteworld.com/wp-content/uploads/2017/11/bikeshare-draft-v7.2lr.mp4>. Consultado 21 de febrero de 2018.

<sup>3</sup>Wikipedia. [https://en.wikipedia.org/wiki/Bicycle-sharing\\_system](https://en.wikipedia.org/wiki/Bicycle-sharing_system). Consultado 21 de febrero de 2018.

<sup>4</sup>Susan Shaheen & Stacey Guzman (Fall 2011). "Worldwide Bikesharing". Access Magazine No. 39. University of California Transportation Center.

<sup>5</sup>Wikipedia. [https://en.wikipedia.org/wiki/Bicycle-sharing\\_system](https://en.wikipedia.org/wiki/Bicycle-sharing_system). Consultado 21 de febrero de 2018.

recolectados, tratados y analizados con el objetivo de mejorar el propio sistema o adaptar aquellos que se vean influenciados por él.

Aunque hay diversos estándares para la publicación de los datos como por ejemplo General Bikeshare Feed Specification (GBFS), empleado por algunos sistemas en Estados Unidos, la mayoría de las compañías no siguen ninguno, proporcionando diferentes datos y estructurándolos, así mismo, de distinta forma. Por lo tanto, la unificación de los mismos es imprescindible si se pretende realizar un estudio con datos de más de un sistema al mismo tiempo. Por otro lado, la mayoría de las entidades no publican un histórico de los datos, únicamente se puede acceder al estado actual del sistema, por lo que, a priori, no es posible realizar un estudio que abarque períodos de tiempo.

Existen aplicaciones como Citybikes que, además de almacenar el estado actual de más de 400 sistemas, permite acceder a ellos mediante servicios web. O, la aplicación creada por el investigador de University College London (UCL) Bike Share Map que publica a través de la web los datos de las últimas 24 horas de más de 100 sistemas.

Si bien se dispone de servicios que unifican múltiples sistemas en tiempo real, no existe ninguno que, además, facilite los datos durante un periodo temporal de más de 24 horas por lo que resulta imposible realizar un estudio histórico<sup>6</sup>.

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, que opera las ciudades de: Amiens, Besancon, Bruxelles-Capitale, Cergy-Pontoise, Creteil, Goteborg, Kazan, Lillestrom, Ljubljana, Luxembourg, Lyon, Marseille, Mulhouse, Namur, Nancy, Nantes, Paris, Rouen, Santander, Seville, Stockholm, Toulouse, Toyama, Valence, Vilnius.

### 1.3. 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.

---

<sup>6</sup><http://opendatalab.uhu.es/index.php/TFG>. Consultado 21 de febrero de 2018.

## 2. Resumen ejecutivo

En base a la información histórica recopilada por las Universidades de Huelva y Sevilla relativa al número de bicicletas disponibles en cada estación en el periodo comprendido entre el 1 de diciembre de 2015 y el 30 de noviembre de 2016, se ha podido caracterizar el patrón espacio-temporal del uso de bicicletas de Sevici en la ciudad de Sevilla y construir modelos predictivos de su uso para horizontes temporales de corto alcance (15min, 30min, 1h, 4h, 8h y 24h).

El patrón temporal de uso de bicicletas de Sevici pone de manifiesto que este servicio funciona preferentemente como medio de transporte público (complementario) para los desplazamientos relacionados con la actividad laboral o académica en la ciudad. Así lo atestiguan los resultados siguientes:

- El número total de bicicletas disponibles en fin de semana o festivo es mayor que en los días laborables en general.
- El número total de bicicletas disponibles a lo largo del día tiene dos mesetas; una de madrugada y otra en el tramo horario de actividad laboral más frecuente.
- Los descensos del número de bicicletas disponibles más acusados a lo largo del día tienen lugar entorno a las 8:00, las 14:00 y las 20:00.

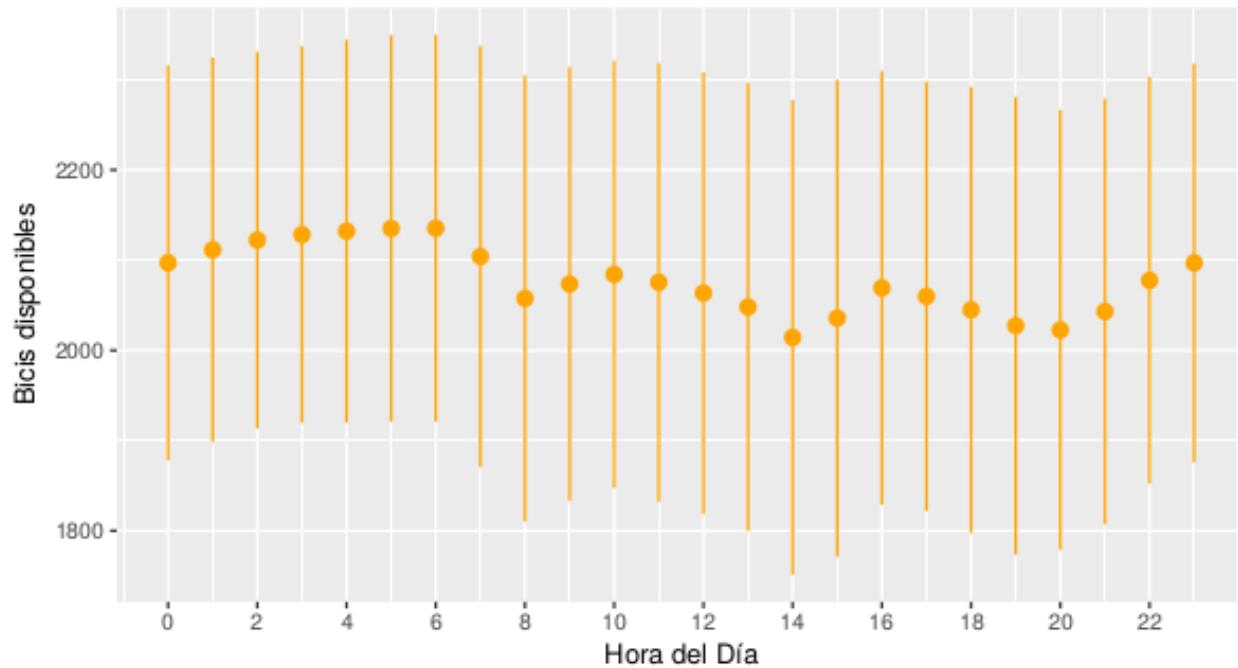


Figura 1: Total de Bicis disponibles en el conjunto de estaciones por hora del día.

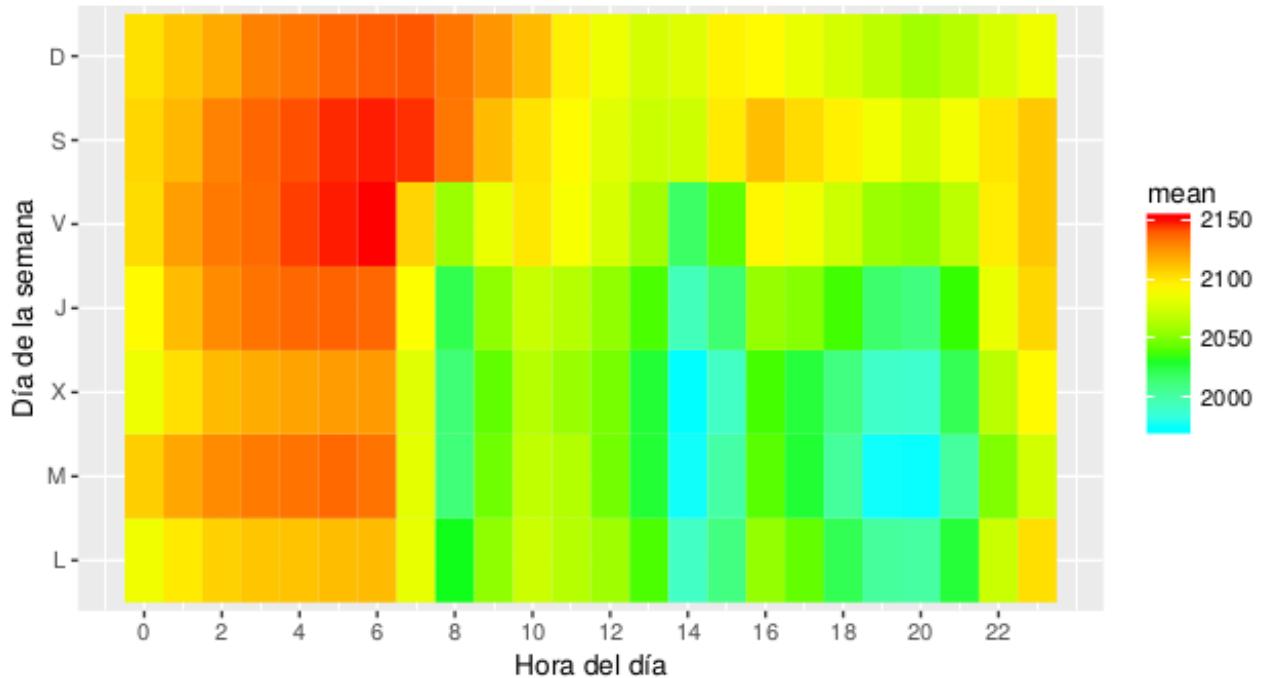


Figura 2: Total de Bicis disponibles en el conjunto de estaciones por hora del día y día de la semana.

La clasificación de las estaciones en base al parecido de su comportamiento temporal nos ha permitido obtener un patrón espacio-temporal en el que se refuerza la conclusión señalada anteriormente y en el que se puede apreciar cómo del análisis del uso de las bicicletas podría deducirse el papel funcional preferente en la trama urbana de las áreas en que se ubican las estaciones. El mapa de estaciones clasificadas y los patrones temporales de cada una de las clases de estación ponen lo dicho claramente de manifiesto.

Puede apreciarse una distribución espacial de las cinco clases identificadas muy concentrada o compacta, esto es, sería posible establecer una zonificación con un número de zonas casi homogéneas relativamente bajo (entre los vecinos de cada estación son en general mayoría los de su misma clase). Se aprecia así mismo una disposición con cierto carácter concéntrico para las clases.

- La clase 4 ocupa una posición central extendiéndose por parte del casco histórico de la ciudad, los barrios de Nervión, Los Remedios, Felipe II.
- La clase 3 ocupa la primera corona entorno a la clase 4, en Triana, centro Norte - Macarena, Santa Justa, Provenir, Tiro de Linea - La Paz.
- La clase 2 ocupa toda la Isla de la Cartuja, todo el entorno de La Palmera (Sur) y núcleos menores en Nervión, Macarena, Sevilla-Este, Alcosa-Torreblanca y Bellavista.
- La clase 1 ocupa la periferia Norte y Este adentrándose hacia el centro sobre todo por el norte (Macarena, Alameda).
- La clase 5 está en exclusiva en Parque Alcosa-Torreblanca y una estación en Bellavista.

Las zonas de Sevilla-Este, Parque Alcosa-Torreblanca (al Este), Bellavista (al Sur) y posiblemente también San Jerónimo (al Norte) están suficientemente distanciadas del resto como para generar dinámicas propias con patrones de centralidad distintos, lo que podría explicar la distribución de clases que se observan en las mismas.

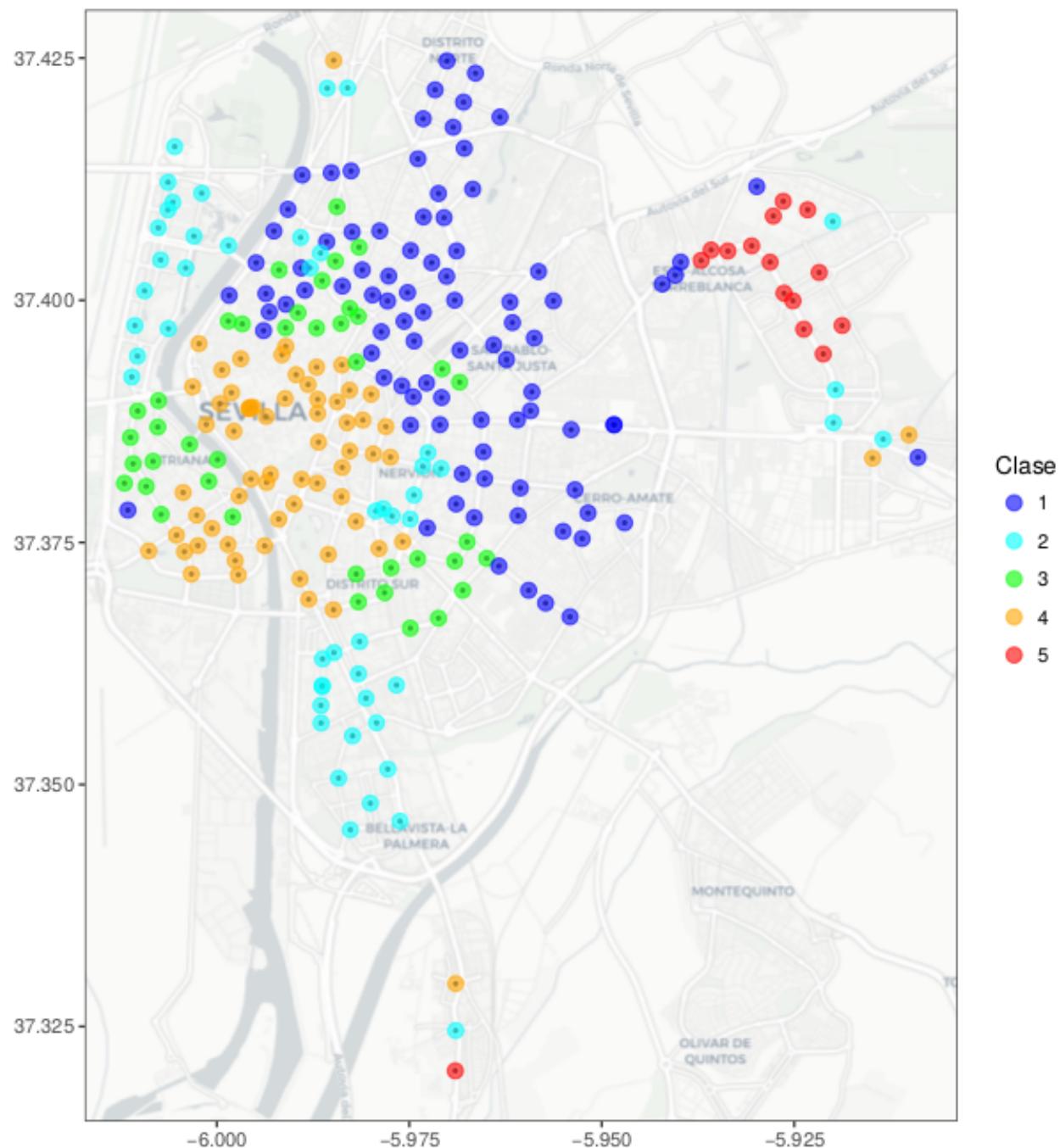


Figura 3: Mapa de estaciones clasificadas.

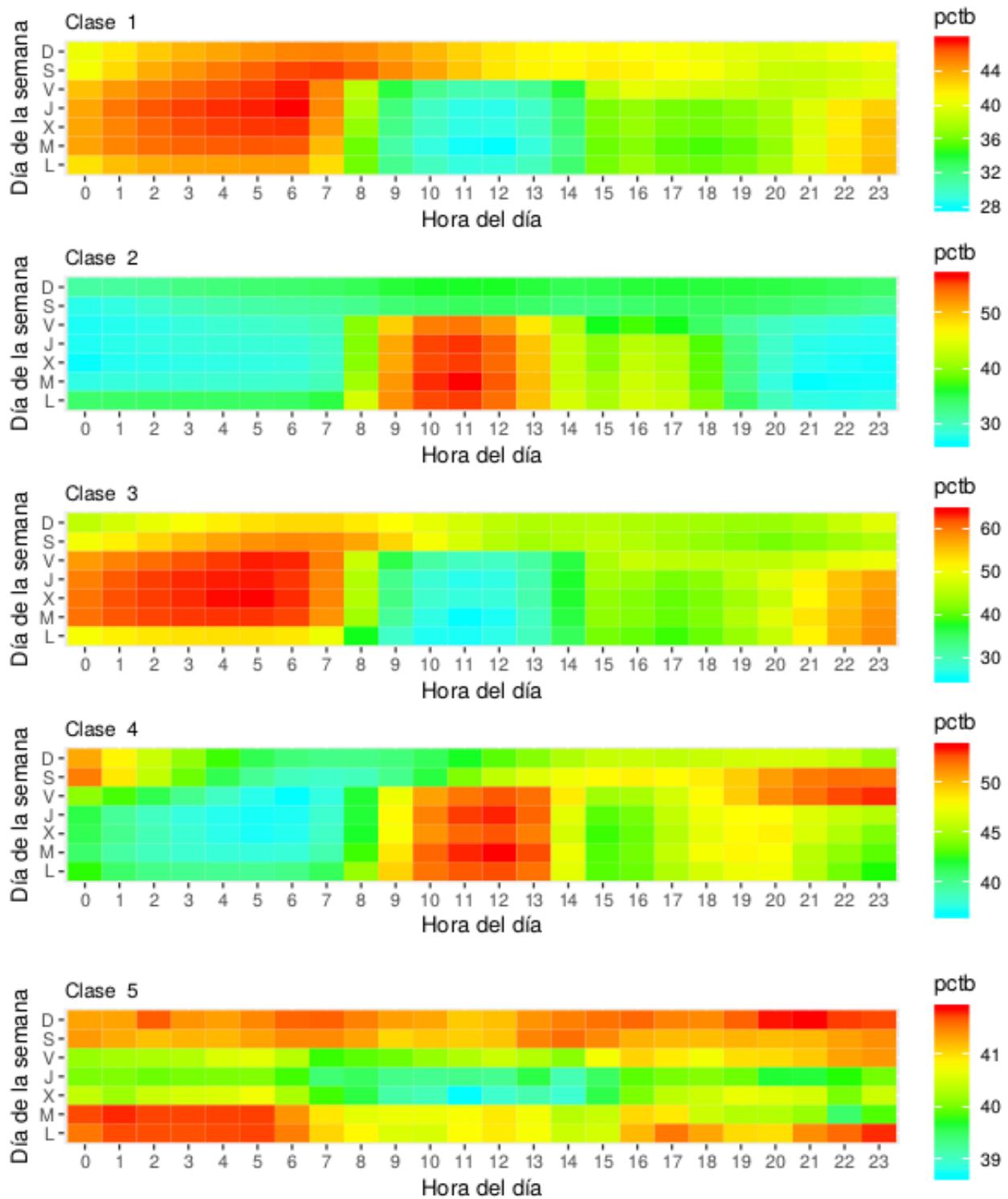


Figura 4: % Bicis disponibles por hora del día y día de la semana. Patrones según clase de estación.

Los datos recopilados por las Universidades de Huelva y Sevilla han permitido también la construcción de modelos predictivos del número de bicicletas disponibles en las estaciones de Sevici.

Se considera en los modelos desarrollados 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:

$$Y(i, t - 15min), Y(i, t - 30min), Y(i, t - 1h), Y(i, t - 4h), Y(i, t - 8h), Y(i, t - 24h)$$

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

$$Y(j, t - 15min), Y(j, t - 30min), Y(j, t - 1h), Y(j, t - 4h), Y(j, t - 8h), Y(j, t - 24h)$$

- 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. En total se han estimado 1813 modelos, siete por cada una de las estaciones.

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).

Los resultados del testeo de los modelos han puesto de manifiesto una muy buena capacidad predictiva, con errores (RMSE, raíz del error medio cuadrático), bajos, 17.59 en media y mediana de 15.51. Hay que tener en cuenta que RMSE se mide en las mismas unidades que la variable objetivo de predicción, en nuestro caso porcentaje de bicicletas disponibles.

La bondad de los modelos por tipo es también bastante buena, si bien, como era esperable con un notable incremento del error a medida que se dispone de menor información para la predicción. En los modelos con disponibilidad de información muy reciente (15min), la bondad del ajuste, se dispara con R2 muy próximo a 1 y RMSE entorno a 5. Para un horizonte de predicción de 4h, RMSE se sitúa entorno a 20, con 24h en 26 y para horizontes más lejanos, el RMSE está entorno a 32.

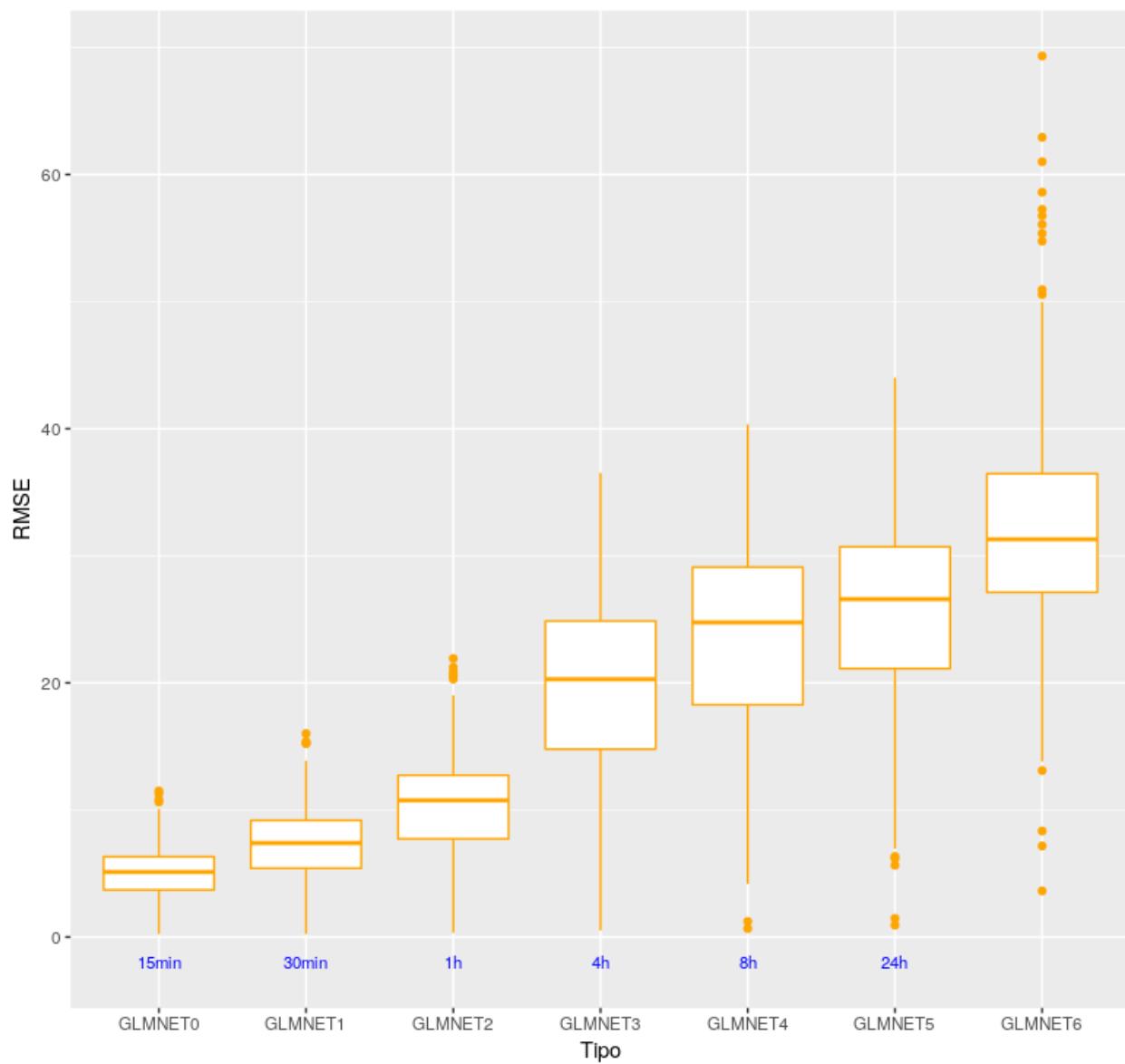


Figura 5: Bondad de ajuste. Raíz del error cuadrático medio (RMSE) por tipo de modelo.

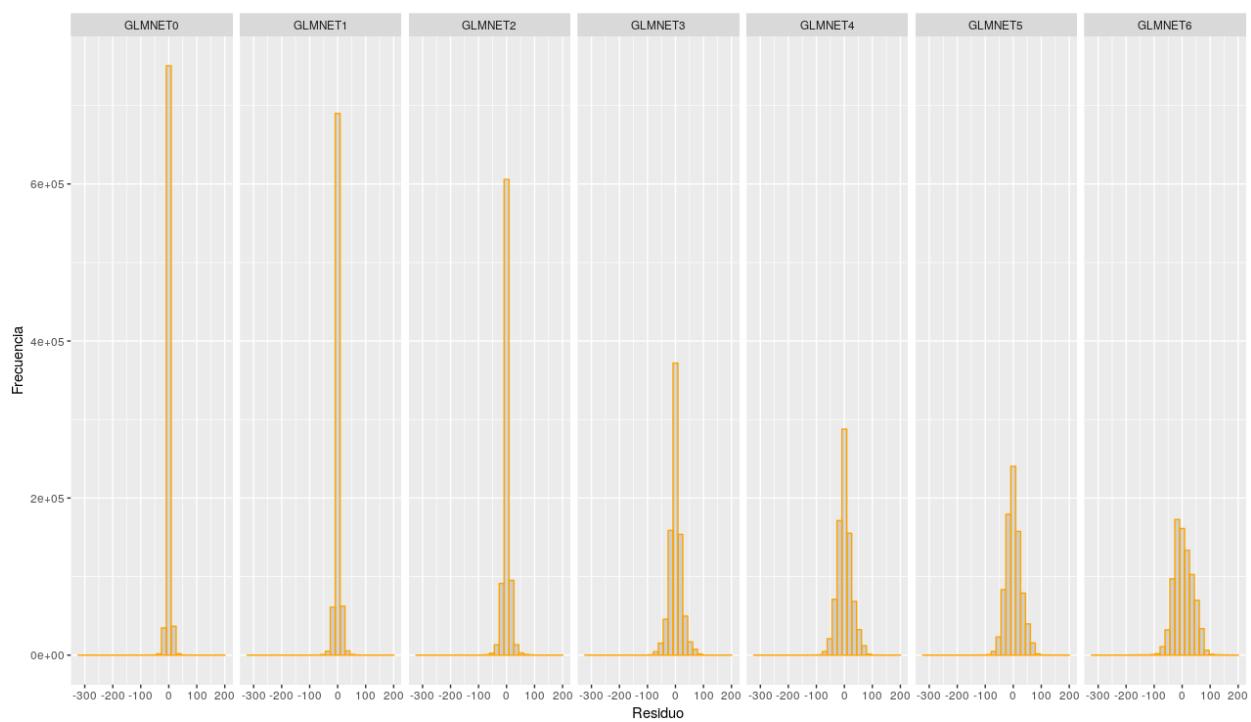


Figura 6: Distribución de residuos por tipo de modelo.

### 3. Metodología

#### 3.1. Descripción del conjunto de datos. Obtención y carga de datos

##### 3.1.1. Datos dinámicos Sevici

Como se ha comentado en la introducción los datos provienen de una recopilación realizada por la Universidades de Huelva y Sevilla, que captura los datos instantáneos ofrecidos a través de un servicio web por JCDecaux en 27 ciudades en las que opera los servicios de bicicletas compartidas.

El punto de partida ha sido un fichero comprimido que contiene para cada ciudad un conjunto de backups (mysql) en formato sql correspondiente cada uno de ellos a los datos registrados en las distintas estaciones de la ciudad en un día y que en la base de datos se corresponde cada uno con una tabla de igual nombre al fichero sql (salvo extensión).

Se ha creado un base de datos (MariaDB) con igual nombre a la original, *pfcbicis*, y se ha realizado la importación de los datos con script bash.

Se han creado así 365 tablas correspondientes a cada uno de los días entre 2015-12-01 y 2016-11-30. Todas ellas con el mismo esquema.

Cuadro 1: 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

##### 3.1.2. Datos estáticos Sevici

Al margen de los datos anteriormente descritos, que corresponde a los denominados datos dinámicos, en la página web del operador (<https://developer.jcdecaux.com/#/opendata/vls?page=static>) están disponibles los denominados datos estáticos que hacen referencia a las características de las estaciones. Esta información se ha descargado en formato csv y contiene los siguientes datos para un total de 260 estaciones:

Cuadro 2: 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)

Los datos originales, descritos hasta aquí, se han reorganizado de diversas formas para los diversos pretratamientos y tratamientos realizados.

### 3.1.3. Datos meteorológicos

Se han obtenido otros datos de interés para el estudio que incluye datos meteorológicos; precipitación y temperaturas, descargados de <https://datosclima.es>.

Los datos meteorológicos pertenecen a las estaciones de Aeropuerto de San Pablo y Tablada. Se ha obtenido una combinación de los registros de ambas estaciones que incluye para cada día la precipitación total (máxima de las dos estaciones), temperatura máxima (máxima) y temperatura mínima (mínima).

Cuadro 3: Estructura de *Meteo*.

Campo:	Descripción:
fecha	Fecha
p	Precipitación total
tmax	Temperatura máxima
tmin	Temperatura mínima

### 3.1.4. Calendario de festivos

Un factor de interés para el estudio es obviamente el calendario laboral en el periodo considerado. Dicha información se ha incorporado manualmente a la base de datos en forma de tabla con la siguiente estructura:

Cuadro 4: Estructura de *Festivos*.

Campo:	Descripción:
fecha	Fecha del día festivo
festivo	Festivo

## 3.2. Pretratamiento y depuración

Para facilitar los tratamientos posteriores se ha unificado la información de las 365 tablas diarias en una sola tabla (*sevadata*) y se han creado tablas con los datos estáticos (*seviesta*), datos meteorológicos (*meteo*) y festivos (*festivos*).

Sobre los datos ya unificados se ha realizado un primer análisis encaminado a la identificación de posibles deficiencias en los mismos.

### 3.2.1. Localización de estaciones

Los datos estáticos de Sevici relativos a la identificación y ubicación de estaciones no muestran indicios de errores relevantes, ubicándose por sus coordenadas en las direcciones que le dan nombre. Las figuras siguientes muestran su localización sobre distintos mapas base con indicación de su codificación.

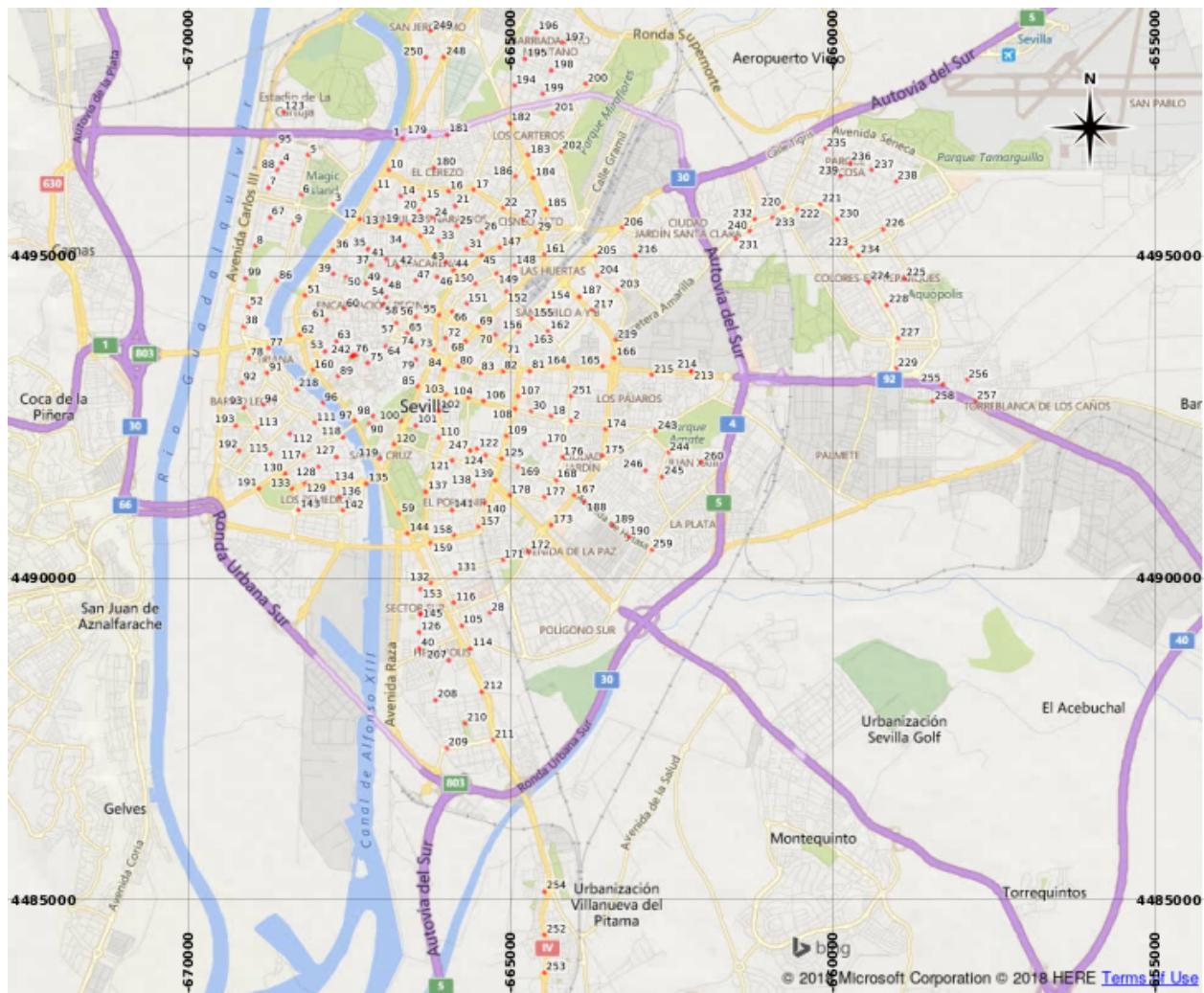


Figura 7: Mapa de Estaciones Sevici sobre BingMap.

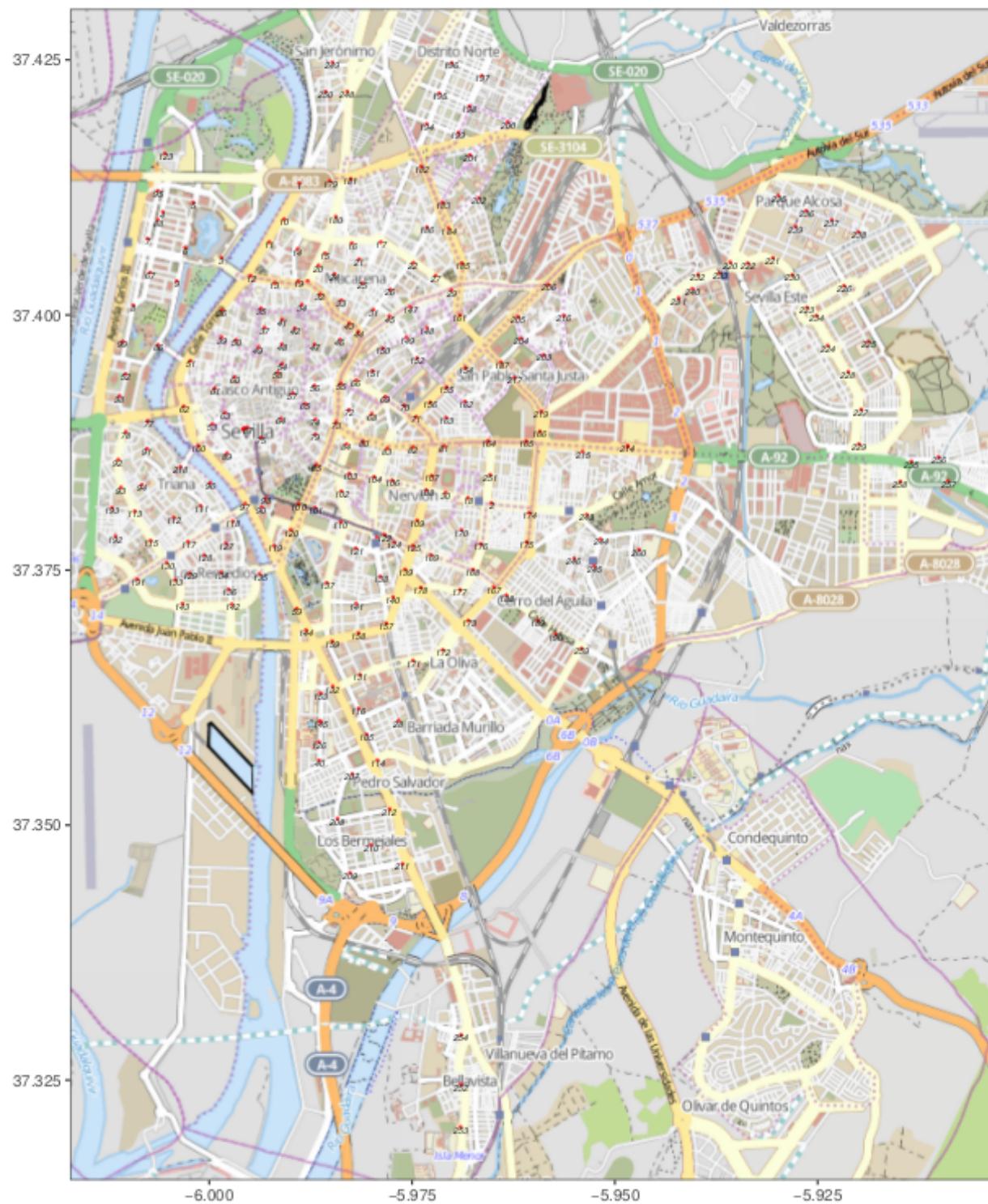


Figura 8: Localización de estaciones SEVICI sobre OSM - 'hikebike'

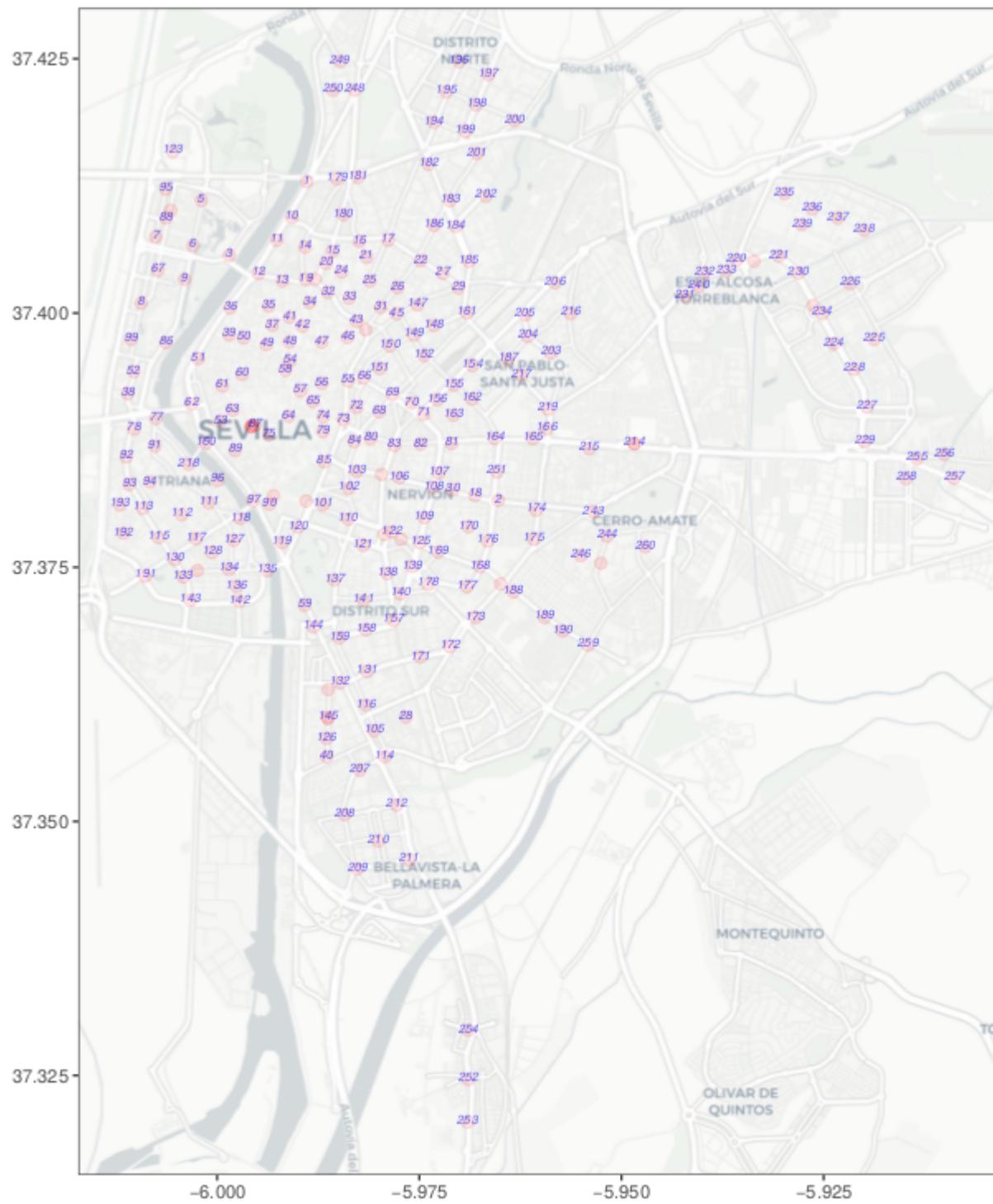


Figura 9: Localización de estaciones SEVICI sobre OSM - 'cartolight'

### 3.2.2. Datos replicados

En teoría la combinación (num - add\_date) debe ser única, esto es, para cada estación y periodo (de 5min) debe haber un único registro.

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.

$259 \times 12 \times 2 = 6216$  que es el número total de réplicas.

Volviendo a los datos originales (backups) se comprueba que todos se encuentran en la tabla *z\_Seville\_2016\_10\_30*.

Para facilitar el manejo de duplicados y otras incidencias en los datos, antes en su caso de eliminación de los registros implicados, se modifica la estructura de *seidata* añadiendo un campo indicador, *ok*, para recoger las distintas incidencias. Se crea también un índice *ok\_idx* para acelerar filtrados.

- Sin incidencia: ok = 1.
- Duplicado: ok = 2.

### 3.2.3. Datos faltantes

Para identificar los posibles huecos en las series temporales de cada estación, vamos en primer lugar a obtener la secuencia temporal de 5 min de paso, para el conjunto de las estaciones, esto es, el listado ordenado de valores únicos de la columna *add\_date*.

El listado lo forman 104170 registros. Como ya se ha señalado anteriormente el número teórico de registros entre inicio y fin para cada estación es de 105120 (365 días x 24 horas x 12 p5min), a lo que hay que añadir 12 registros adicionales hasta las 00:55:01 del día de fin 2016-11-30 (105132), lo que supone que existen 962 huecos de 5min sin datos que afectan a la totalidad de las estaciones. El número total de huecos entre todas las estaciones será obviamente muy superior y al menos de ese tamaño para cada estación.

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.

La variable *hueco* indica si es un hueco global o no, es decir, si no existe ningún dato para ninguna estación en ese momento (TRUE) o existe al menos una con datos.

La distribución temporal de los huecos globales por fecha y hora se muestra en la figura siguiente.

El número total de huecos en el conjunto de datos es 936353 de los cuales 81533 corresponden a la estación 109 que sólo estuvo en funcionamiento durante aproximadamente los tres primeros meses de estudio. Excluyendo la estación 109 se tendrán 854820 huecos. Recordar que el número de periodos de 5min que afectan a la totalidad de estaciones es de 962, lo que supone un total de  $962 \times 259 = 249158$  huecos globales, excluida la estación 109, y por tanto, el número de huecos en los que existe al menos una estación con datos es de 605662.

Cuadro 5: Resumen de huecos en datos dinámicos

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

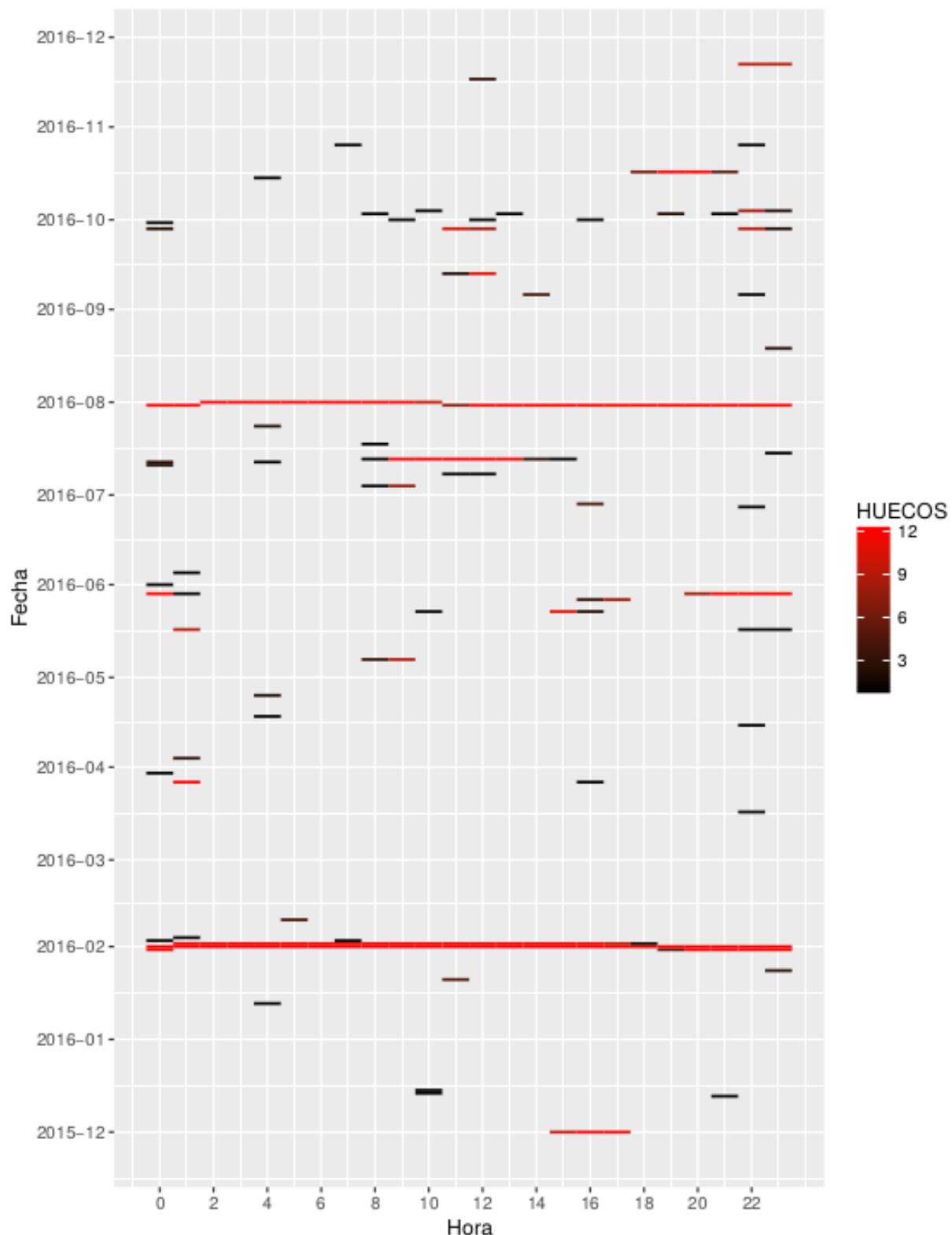


Figura 10: Datos faltantes. Huecos Globales por Fecha y Hora

### 3.2.4. Datos anómalos

Entre los datos anómalos se consideran las siguientes situaciones:

- a) Número de estacionamientos disponibles mayor que operativos
- b) Número de bicicletas disponibles mayor que estacionamientos operativos
- c) Suma de estacionamientos disponibles y bicicletas disponibles mayor que el número de estacionamientos operativos.
- d) Suma de estacionamientos disponibles y bicicletas disponibles menor que el número de estacionamientos operativos.

Se codifican dichas situaciones en la tabla *seidata* en el campo *ok* con los siguientes valores: a)  $\rightarrow ok = 3$  b)  $\rightarrow ok = 4$  c)  $\rightarrow ok = 5$  d)  $\rightarrow ok = 6$

Se muestra seguidamente el resumen incidencias relativas a datos duplicados y anómalos identificados

Cuadro 6: 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	Bicicletas disponibles > Est. operativos	965
5	Estacionamientos + Bicicletas disponibles > Est. operativos	5728
6	Estacionamientos + Bicicletas disponibles < Est. operativos	4367165

El número de datos anómalos representa el 16.54 % del total de datos registrados. De todos ellos la situación más frecuente, con diferencia, es aquélla en la que la suma de estacionamientos disponibles y bicicletas disponibles es menor que el número de estacionamientos operativos (*ok=6*), con un 16.50 % del total de registros.

La situación anómala descrita, *ok=6*, podría en teoría responder a retrasos puntuales en la transmisión de datos.

Es necesario tener en cuenta que el número de estacionamientos operativos se ha comprobado que aparece como constante durante todo el periodo estudiado para todas las estaciones, que es lo realmente sorprendente, si dicho dato se obtiene de modo similar al de disponibilidad de estacionamientos y bicicletas.

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*).

### 3.3. Análisis exploratorio

Se consideran 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. 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.

### 3.4. Clasificación de estaciones e 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:

Cuadro 7: 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), lo que ha permitido realizar representaciones en forma de mapa de las correlaciones temporales entre estaciones.

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 *hclust*. 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.

### 3.5. 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:

$$Y(i, t - 15min), Y(i, t - 30min), Y(i, t - 1h), Y(i, t - 4h), Y(i, t - 8h), Y(i, t - 24h)$$

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

$$Y(j, t - 15min), Y(j, t - 30min), Y(j, t - 1h), Y(j, t - 4h), Y(j, t - 8h), Y(j, t - 24h)$$

- 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  $P$

Dieciocho variables regresoras que, al convertir en binaria  $DSEM$ , con siete niveles, pasan a ser 24.

Para cada momento (t) se realizará la predicción para t, t+15min, t+30min, t+1h, t+4h, t+8h y t+24h.

El modelo predictivo desarrollado es un modelo 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*.

## 4. Principales resultados y conclusiones

### 4.1. Análisis exploratorio

Se presentan aquí datos y resultados relevantes del análisis exploratorio.

#### 4.1.1. Datos meteorológicos

Cuadro 8: Resumen estadístico de variables meteorológicas

fecha	p	tmax	tmin
Min. :2015-11-30	Min. : 0.00	Min. :13.70	Min. : 0.30
1st Qu.:2016-02-29	1st Qu.: 0.00	1st Qu.:19.50	1st Qu.: 8.40
Median :2016-05-31	Median : 0.00	Median :23.80	Median :13.20
Mean :2016-05-31	Mean : 1.77	Mean :26.27	Mean :13.34
3rd Qu.:2016-08-30	3rd Qu.: 0.00	3rd Qu.:33.10	3rd Qu.:18.00
Max. :2016-11-30	Max. :69.40	Max. :44.80	Max. :24.90

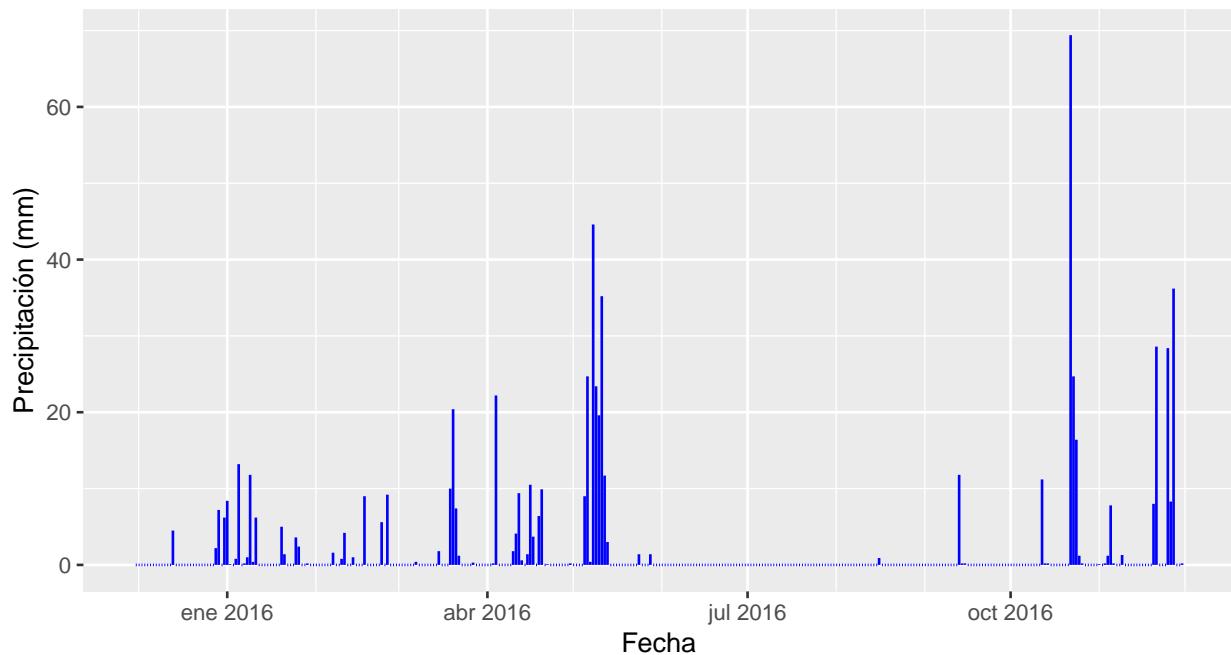


Figura 11: Datos meteorológicos. Serie de precipitación total diaria.

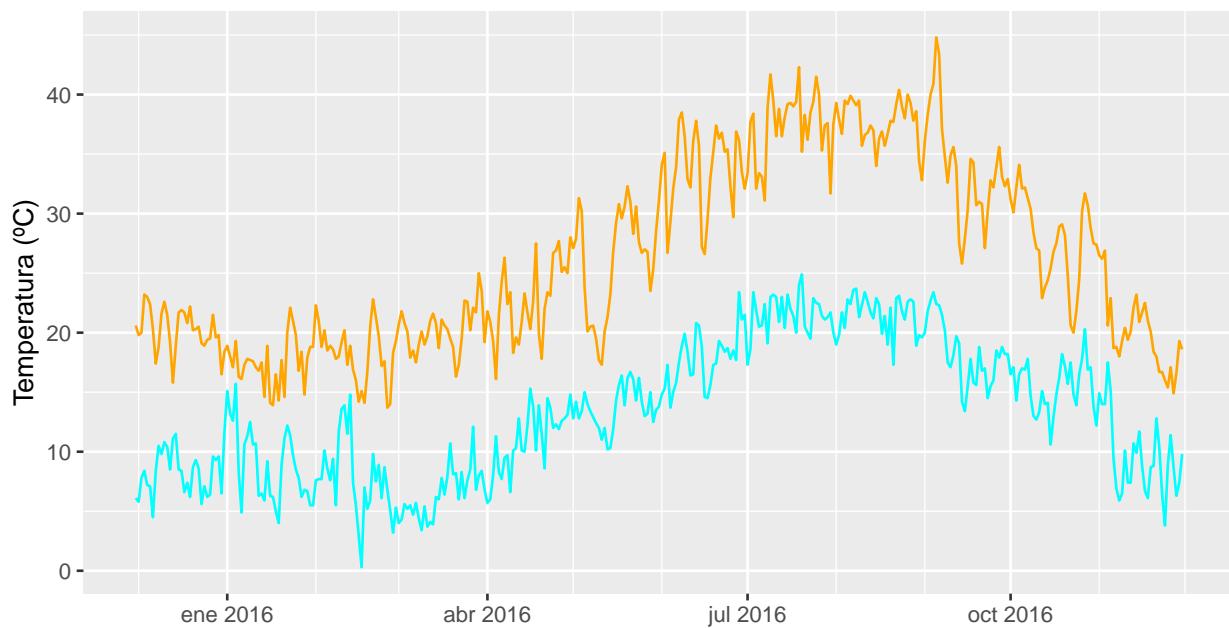


Figura 12: Datos meteorológicos. Series de temperaturas máximas y mínimas diarias.

#### 4.1.2. Análisis de datos válidos globales

La figura siguiente muestra el número bicis disponibles a lo largo del periodo de estudio estimada a partir de la media y expandida al conjunto de estaciones. Se aprecia una ligera tendencia ascendente en el periodo.

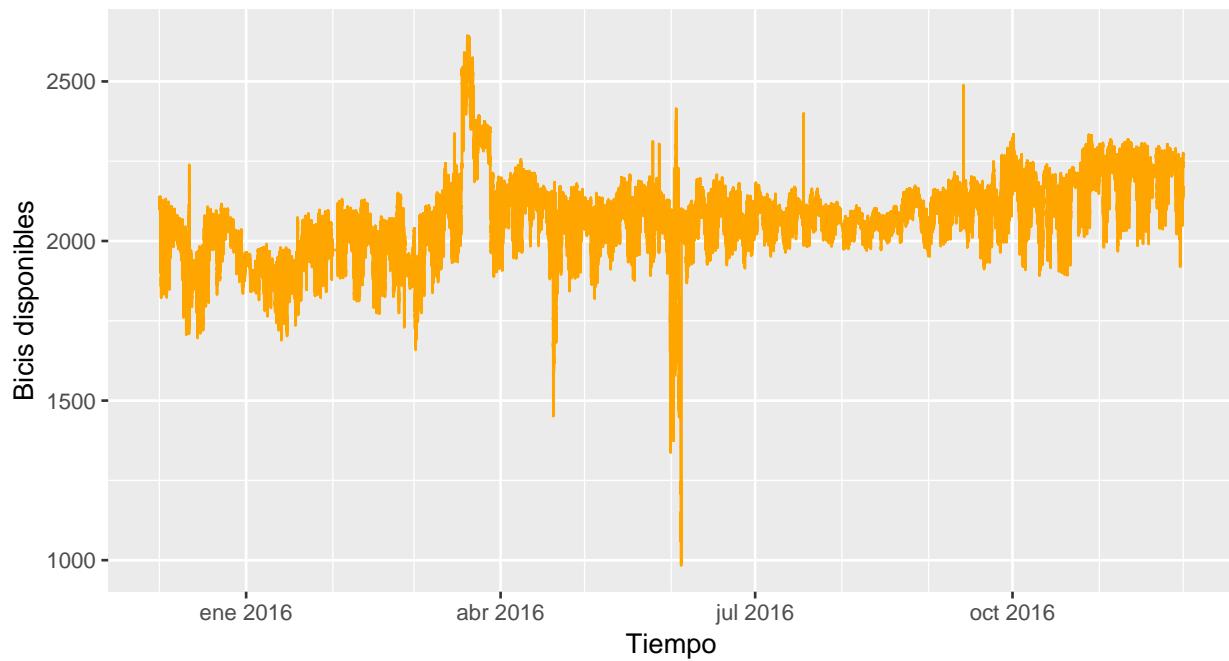


Figura 13: Datos válidos globales. Bicis disponibles.

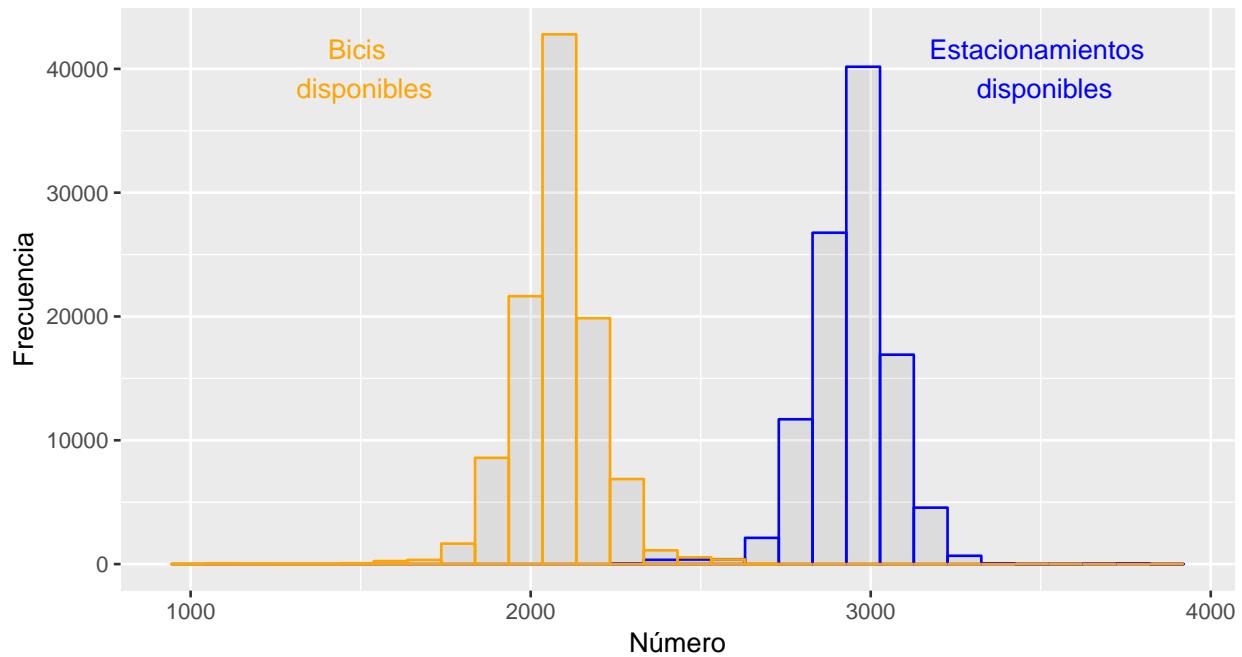


Figura 14: Datos válidos globales. Distribución de Estacionamientos y Bicis disponibles.

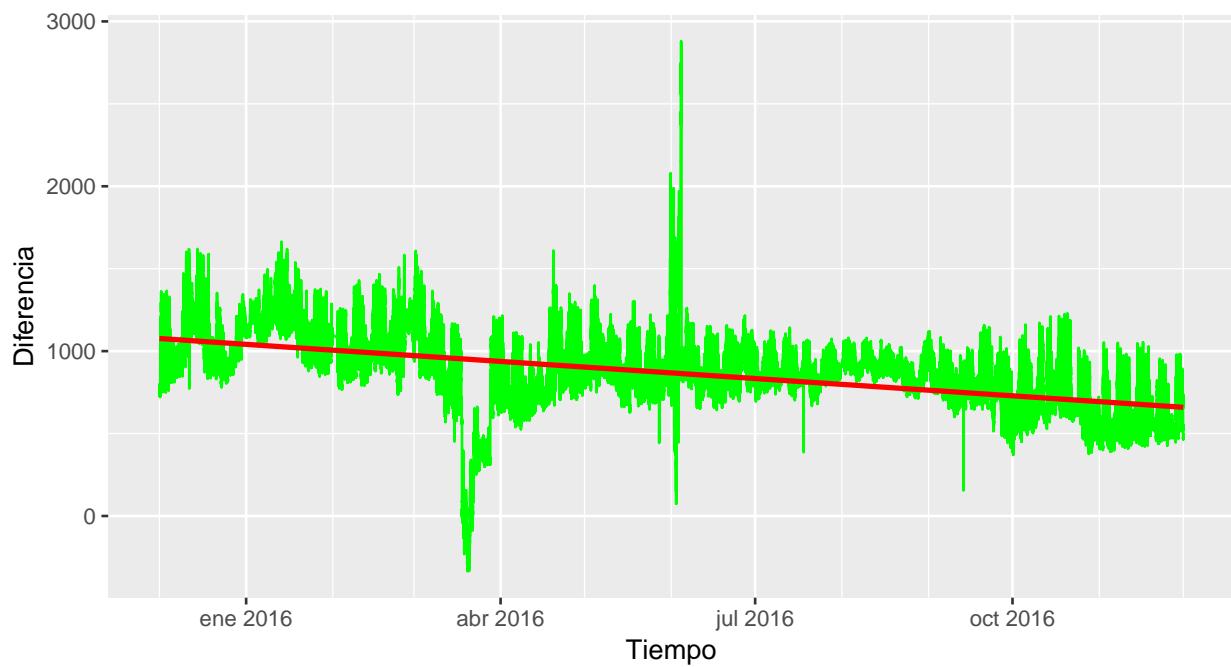


Figura 15: Datos válidos globales. Diferencia Estacionamientos y Bicis disponibles.

El número de estacionamientos disponibles, asimilable en cierta forma a bicicletas circulantes o en uso, es superior al de bicis disponibles a lo largo de todo el periodo analizado (salvo a finales de marzo, semana santa). Fenómeno con una aparente tendencia a su reducción.

#### 4.1.2.1. Análisis según días de la semana y festivos

Cuadro 9: Bicis disponibles por día de la semana. Estadística básica.

dsem	mean	median	min	max	sd
X	2054.6	2057.9	1374.3	2378.8	122.3
M	2056.3	2056.9	1337.1	2532.2	122.5
L	2060.7	2059.9	1696.0	2575.1	116.8
J	2069.7	2071.0	1578.6	2414.3	106.3
V	2091.7	2089.1	1448.6	2591.2	116.8
D	2101.2	2094.1	1834.0	2644.1	120.5
S	2108.6	2113.3	984.3	2586.8	140.1

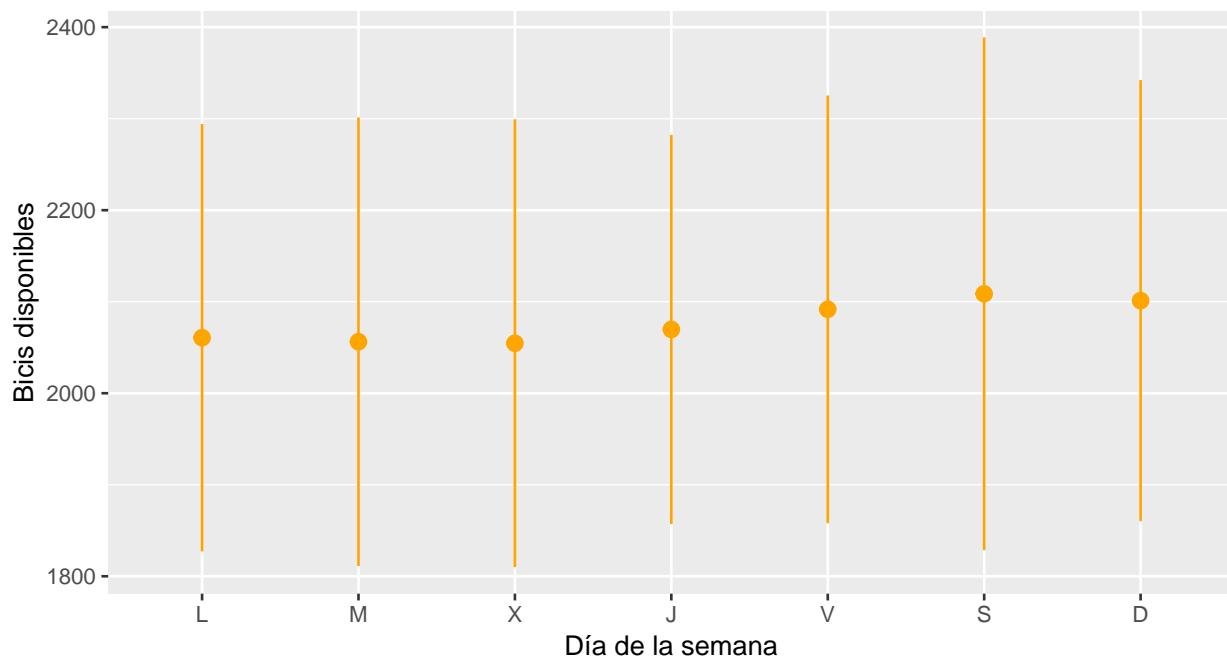


Figura 16: Datos válidos globales. Bicis disponibles por día de la semana. Media +/- 2 · Desviación

Cuadro 10: Bicis disponibles según sea fin de semana - festivo o no.  
Estadística básica.

fsof	mean	median	min	max	sd
FALSE	2064.3	2068.0	1337.1	2591.2	116.4
TRUE	2105.3	2101.1	984.3	2644.1	131.2

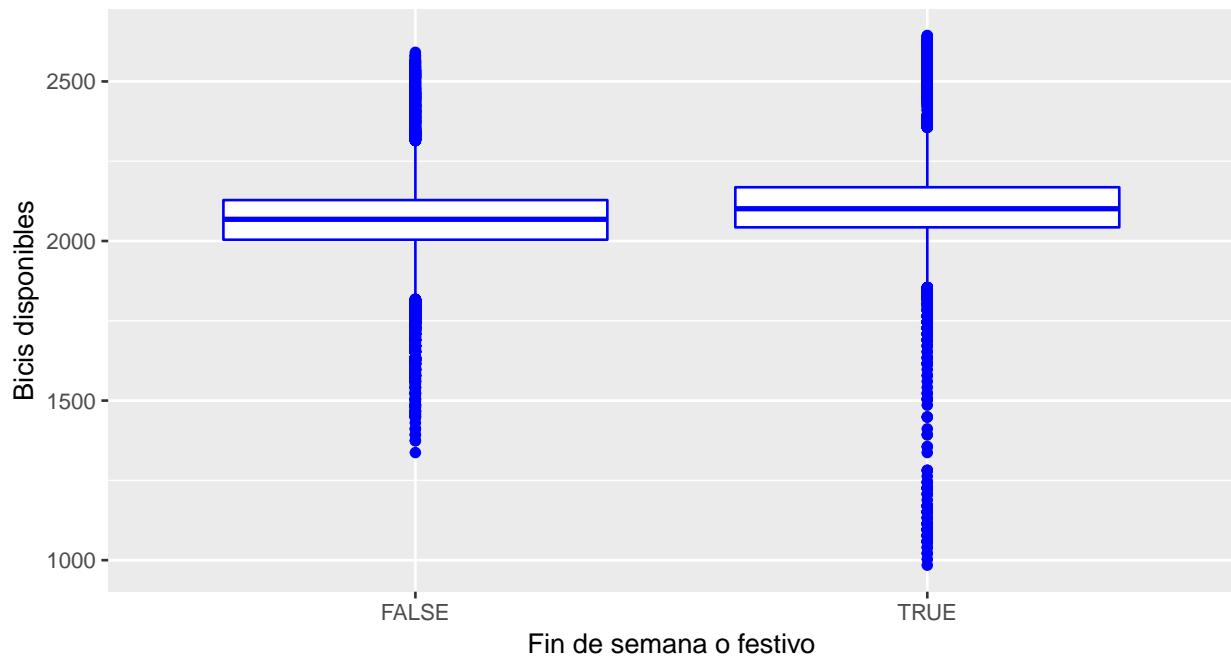


Figura 17: Datos válidos globales. Bicis disponibles según sea fin de semana - festivo o no.

#### 4.1.2.2. Análisis según hora del día

Cuadro 11: Bicis disponibles por hora del día. Estadística básica.

hora	mean	median	min	max	sd
14	2014.3	2019.8	1058.6	2602.2	131.5
20	2022.7	2023.0	1448.6	2580.2	122.0
19	2027.3	2035.8	1337.1	2604.4	126.9
15	2035.7	2043.9	1095.7	2619.8	132.1
21	2043.1	2036.8	1455.2	2566.9	118.0
18	2044.9	2053.8	1504.3	2608.8	123.7
13	2048.1	2054.8	984.3	2591.2	124.1
8	2057.6	2049.9	1392.9	2617.6	123.6
17	2059.9	2068.0	1411.4	2633.1	118.9
12	2063.5	2066.9	1058.6	2593.4	122.3
16	2069.2	2077.0	1170.0	2641.9	120.2
9	2073.5	2066.9	1225.7	2619.8	120.2
11	2075.5	2075.0	1058.6	2591.2	121.6
22	2077.6	2068.0	1549.9	2575.1	112.8
10	2084.2	2080.0	1170.0	2608.8	118.4
23	2097.0	2088.0	1567.1	2591.2	110.5
0	2097.1	2088.0	1504.3	2602.2	109.5
7	2104.0	2099.0	1615.7	2615.4	116.8
1	2111.5	2104.1	1612.4	2630.8	106.6
2	2122.3	2117.3	1618.4	2644.1	104.4
3	2128.4	2123.3	1626.5	2608.8	104.4
4	2132.1	2128.2	1624.5	2622.0	106.2
5	2135.2	2129.4	1624.5	2617.6	107.1

hora	mean	median	min	max	sd
6	2135.4	2131.2	1624.5	2608.8	107.3

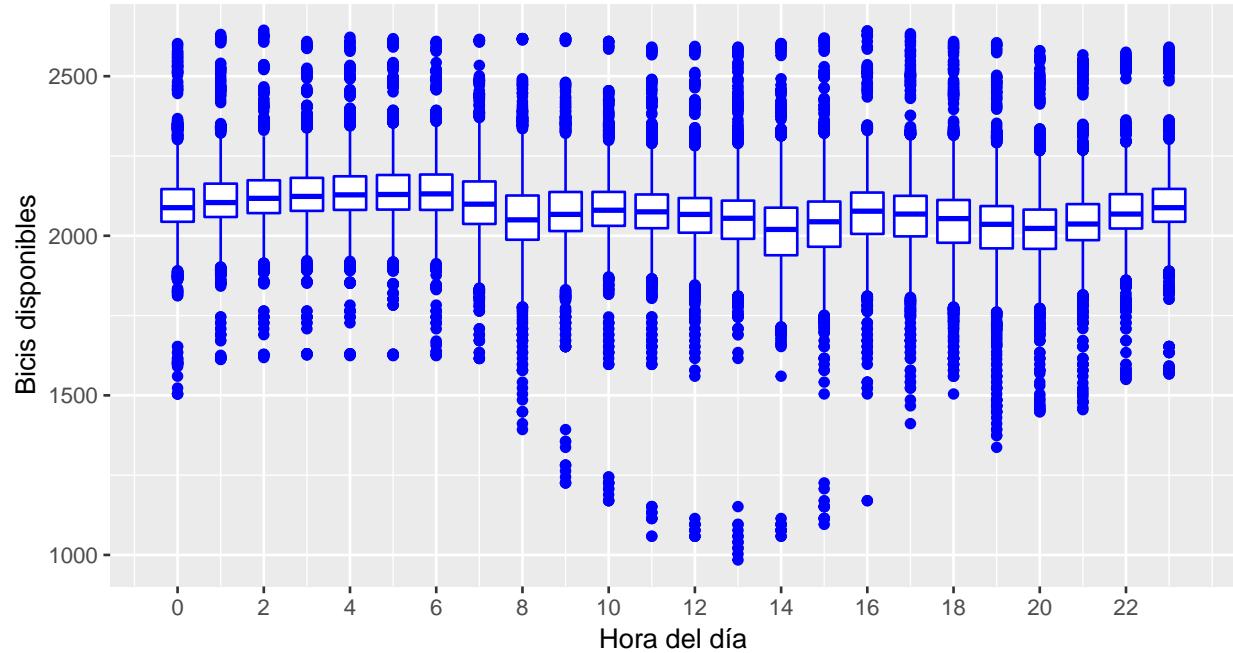


Figura 18: Datos válidos globales. Bicis disponibles según hora del día.

#### 4.1.2.3. Análisis según hora del día y día de la semana

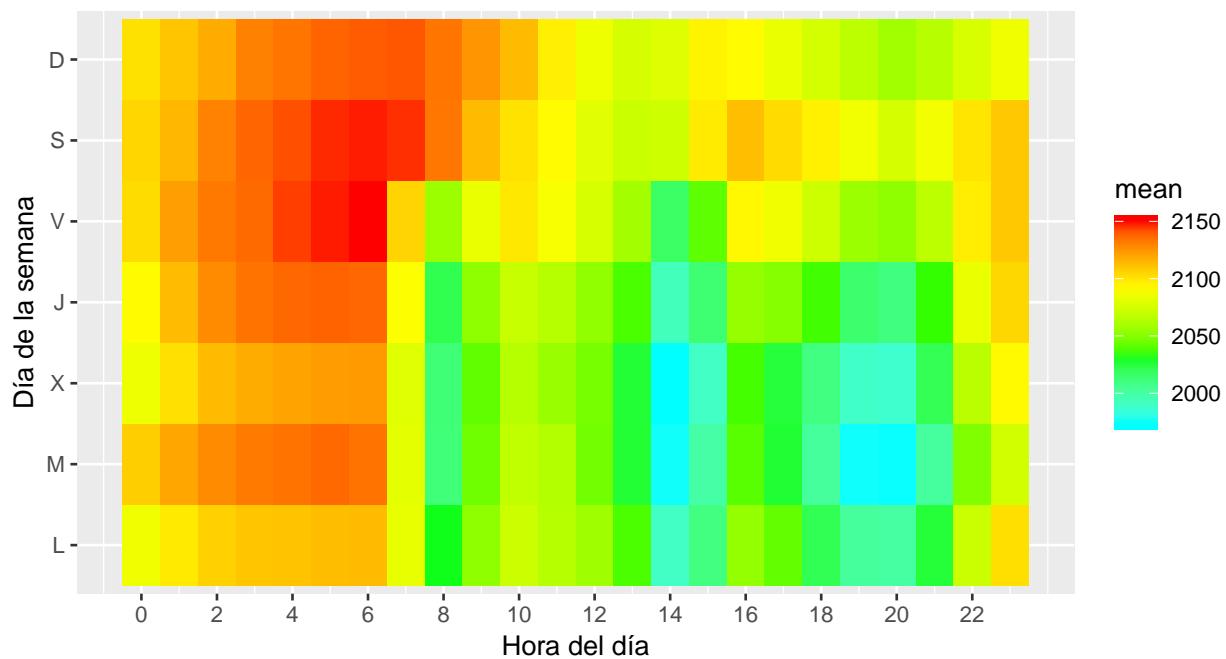


Figura 19: Datos válidos globales. Bicis disponibles según hora del día y día de la semana.

#### 4.1.2.4. Análisis según condiciones meteorológicas

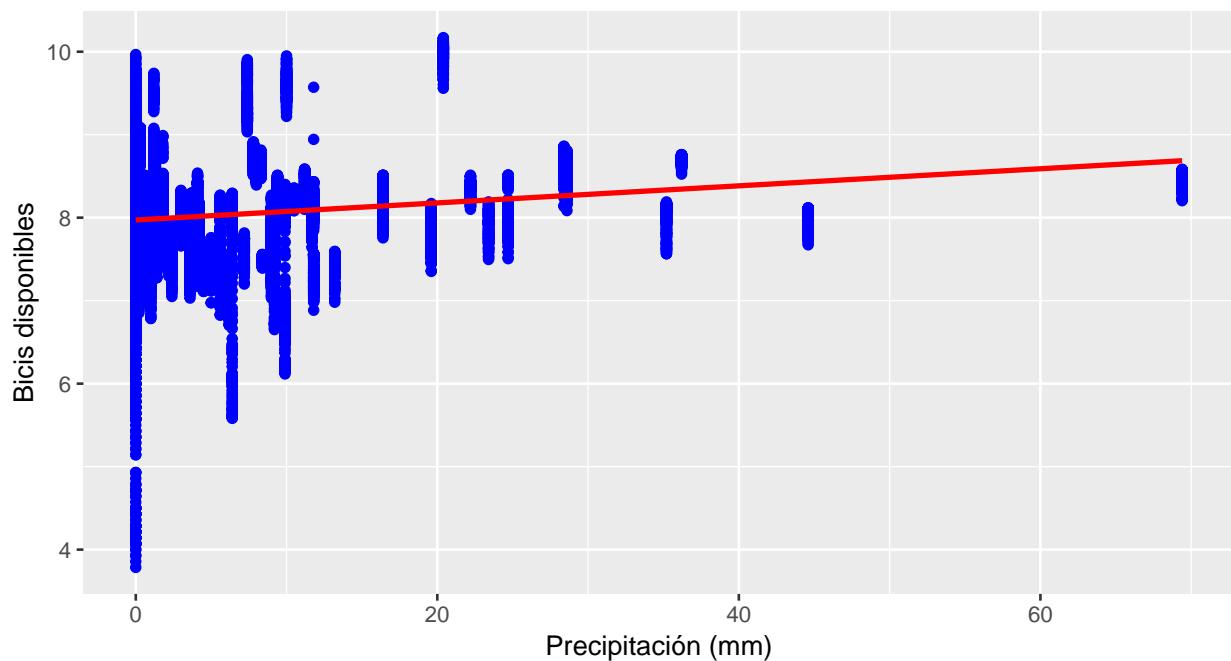


Figura 20: Datos válidos globales. Bicis disponibles según precipitación total diaria.

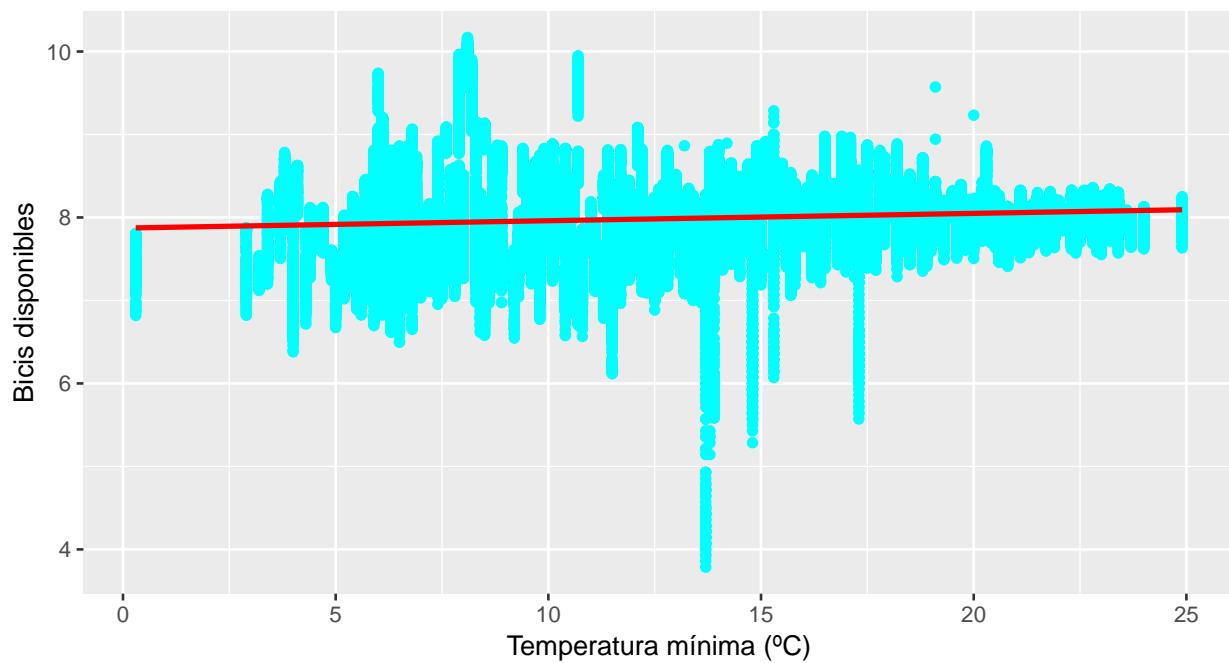


Figura 21: Datos válidos globales. Bicis disponibles según temperatura mínima diaria.

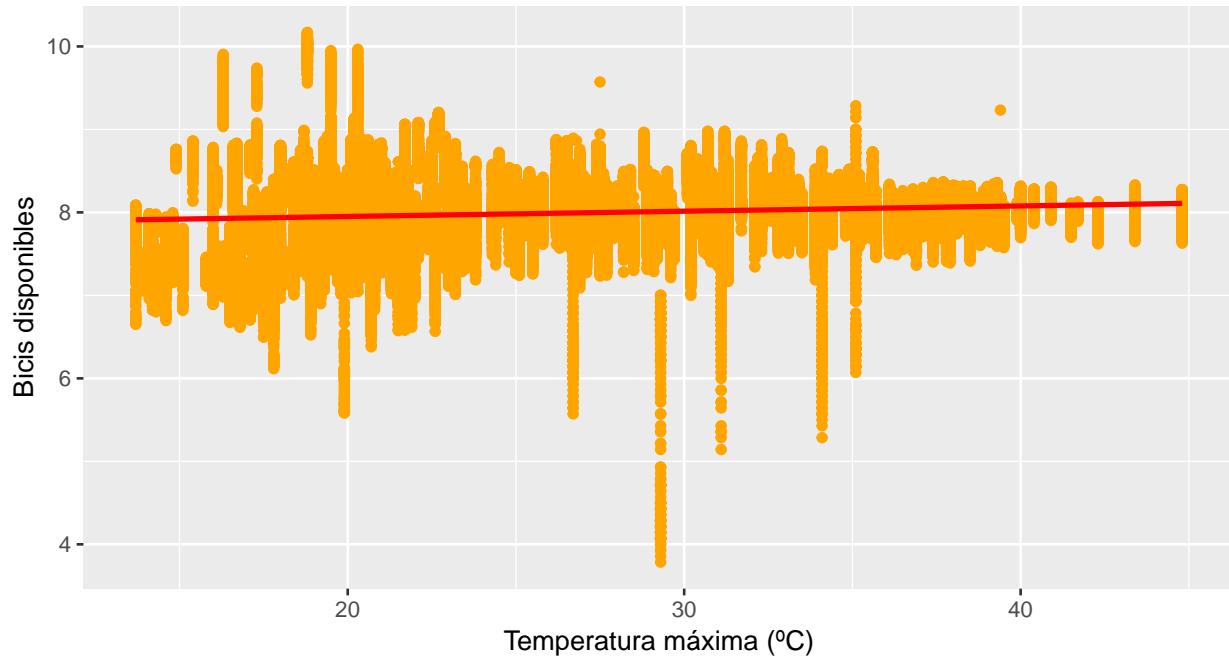


Figura 22: Datos válidos globales. Bicis disponibles según temperatura máxima diaria.

## 4.2. Clasificación de estaciones e identificación de patrones espacio-temporales

### 4.2.1. Análisis de correlación entre estaciones

La matriz de correlación entre estaciones se muestra de forma gráfica en la siguiente figura. En ella se representan las celdas con un coeficiente de correlación en valor absoluto mayor de 0.5.

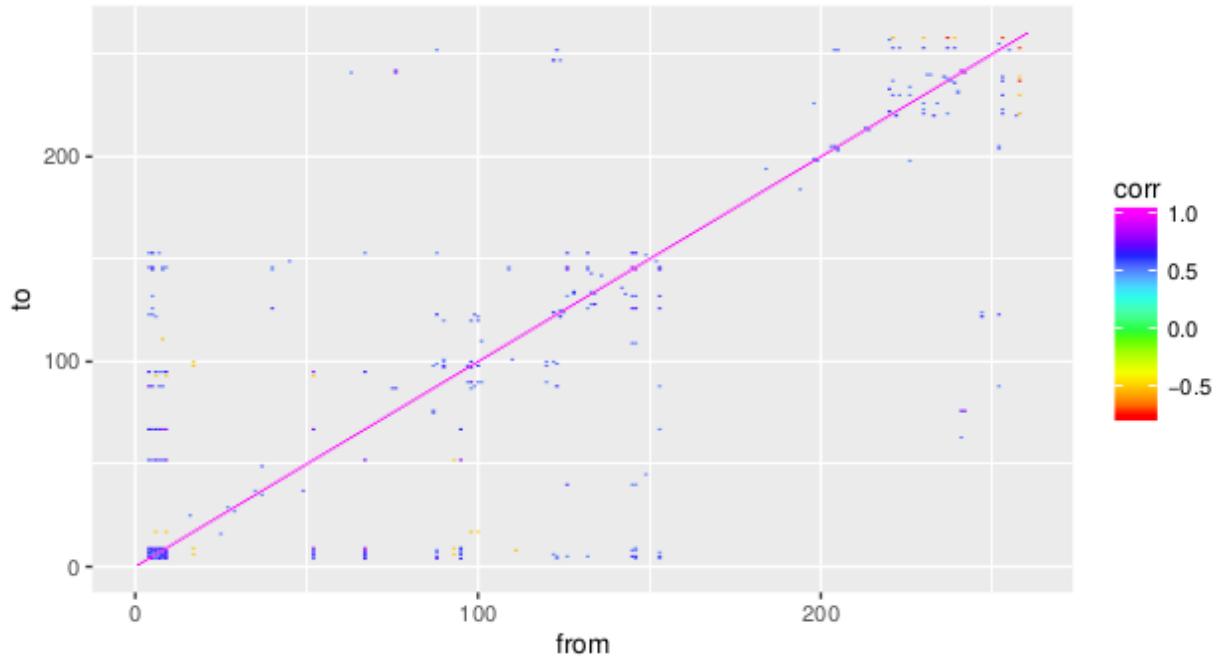


Figura 23: Matriz de correlación ( $|\text{corr}|>0.5$ ) entre estaciones.

La figura siguiente muestra una representación de la matriz de correlación en forma de grafo con nodos georreferenciados. Se muestran los arcos que conectan estaciones correladas en valor absoluto mayor de 0.5 distinguiéndose por color las correlaciones positivas y negativas.

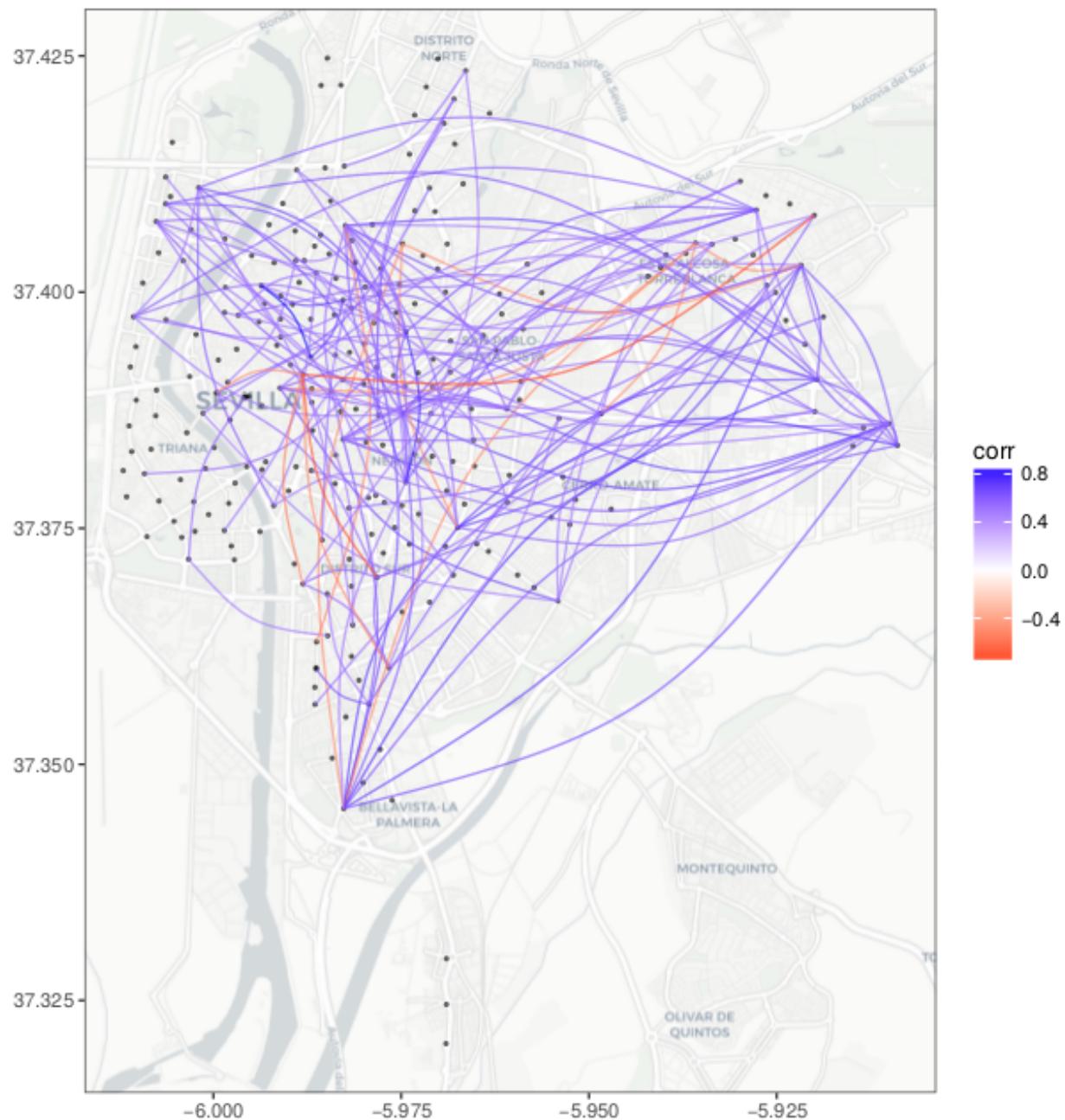


Figura 24: Grafo espacial de correlaciones  $|\text{corr}| > 0.5$ .

#### 4.2.2. Clasificación de las estaciones

La figura siguiente muestra el dendrograma de clasificación de estaciones en base a los datos de la matriz de correlación.

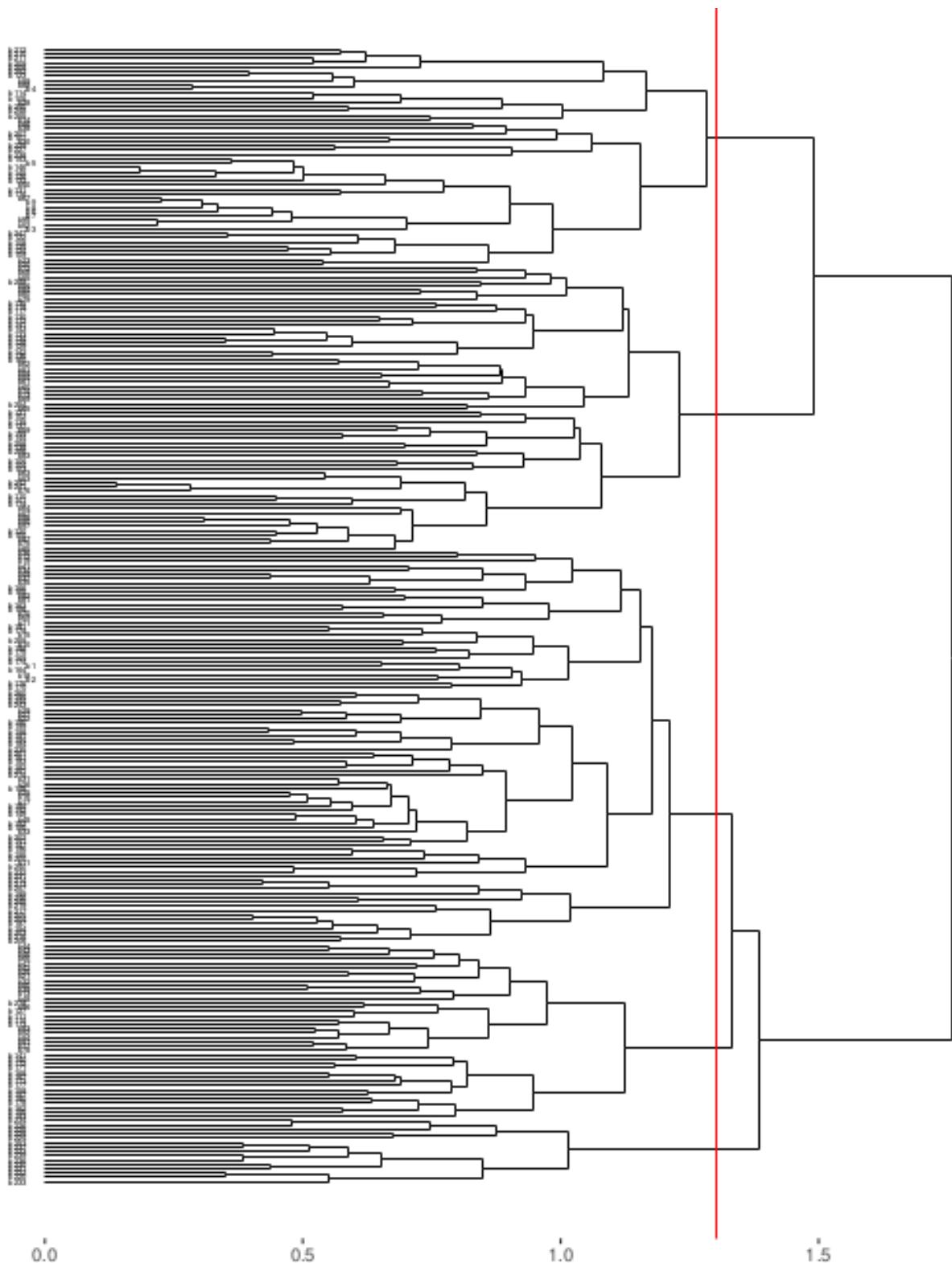


Figura 25: Dendrograma de estaciones basado en correlación.

La clasificación de estaciones obtenida del proceso se muestra en forma de mapa en la siguiente figura.

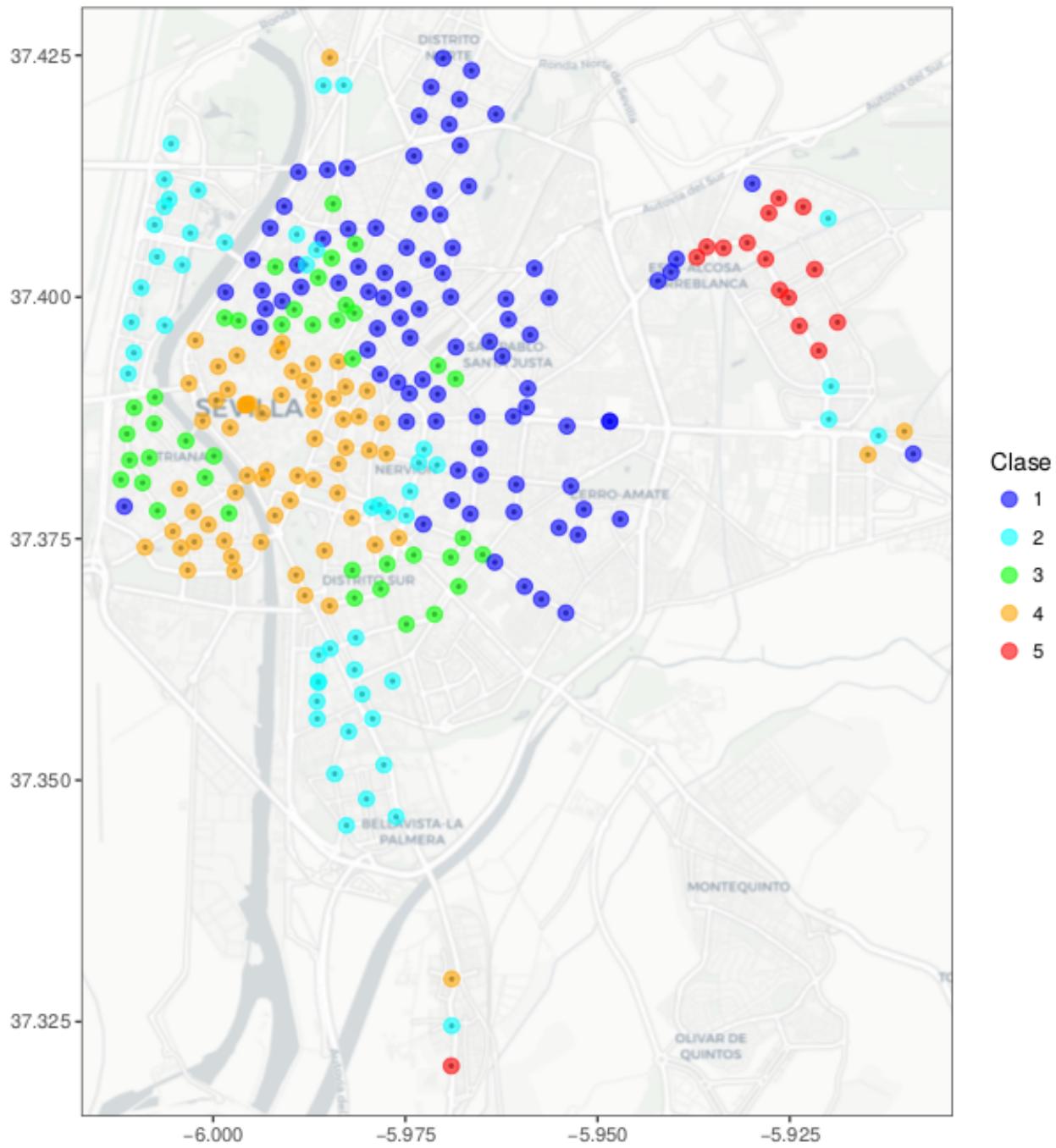


Figura 26: Mapa de estaciones clasificadas

Puede apreciarse una distribución espacial de las cinco clases identificadas muy concentrada o compacta, esto es, sería posible establecer una zonificación con un número de zonas casi homogéneas relativamente bajo (entre los vecinos de cada estación son en general mayoría los de su misma clase). Se aprecia así mismo una disposición con cierto carácter concéntrico para las clases.

- La clase 4 ocupa una posición central extendiéndose por parte del casco histórico de la ciudad, los barrios de Nervión, Los Remedios, Felipe II.
- La clase 3 ocupa la primera corona entorno a la clase 4, en Triana, centro Norte - Macarena, Santa Justa, Provenir, Tiro de Linea - La Paz.
- La clase 2 ocupa toda la Isla de la Cartuja, todo el entorno de La Palmera (Sur) y núcleos menores en Nervión, Macarena, Sevilla-Este, Alcosa-Torreblanca y Bellavista.
- La clase 1 ocupa la periferia Norte y Este adentrándose hacia el centro sobre todo por el norte (Macarena, Alameda).
- La clase 5 está en exclusiva en Parque Alcosa-Torreblanca y una estación en Bellavista.

Las zonas de Sevilla-Este, Parque Alcosa-Torreblanca (al Este), Bellavista (al Sur) y posiblemente también San Jerónimo (al Norte) están suficientemente distanciadas del resto como para generar dinámicas propias con patrones de centralidad distintos, lo que podría explicar la distribución de clases que se observan en las mismas.

#### 4.2.3. Patrones espacio-temporales

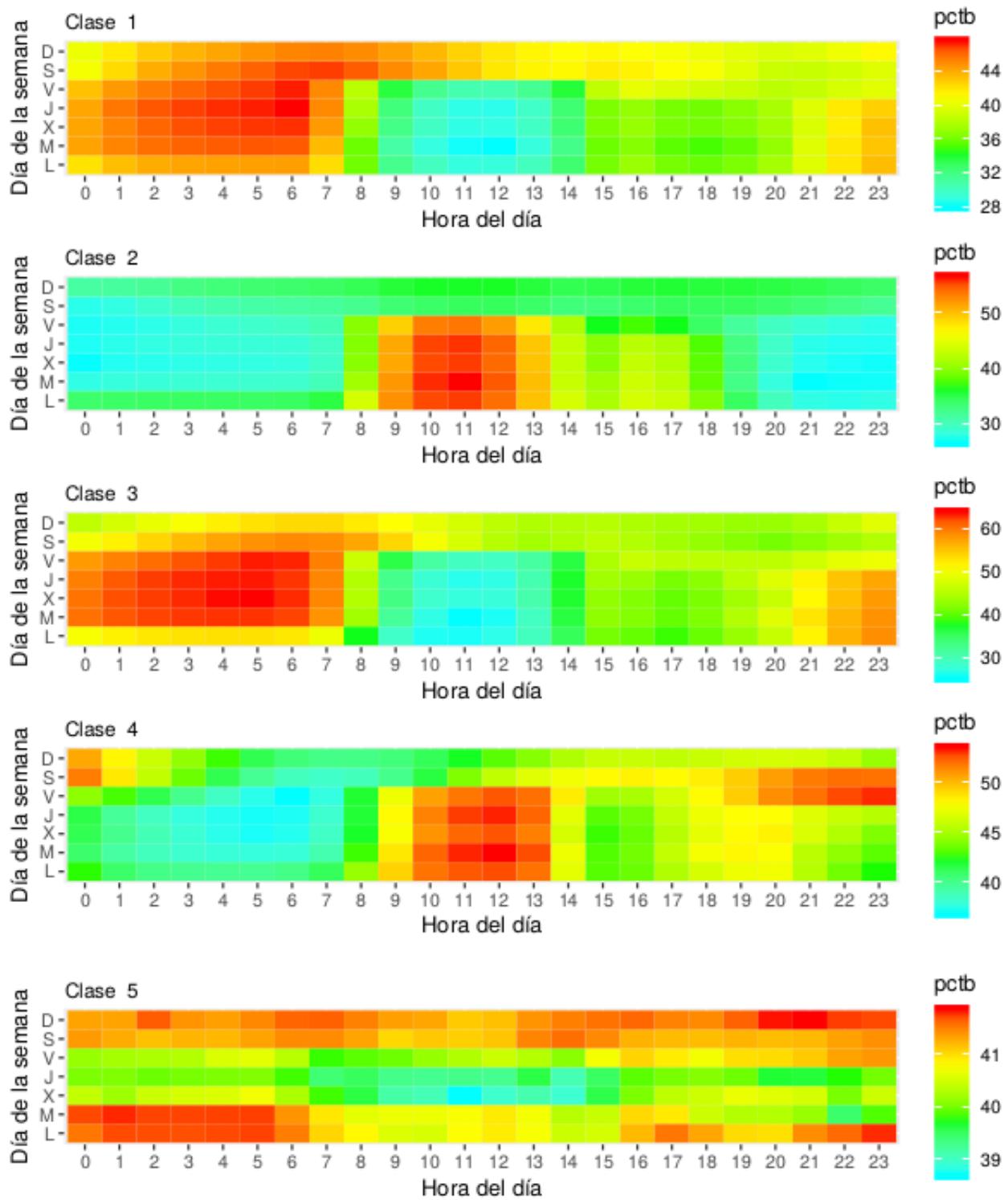


Figura 27: % de Bicis disponibles por hora del día y día de la semana. Patrones por clase de estación.

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.

### 4.3. Modelo predictivo

#### 4.3.1. Persistencia de modelos

Las tablas *modelos*, *betas* y *residuos* recogen la información derivada del ajuste, entrenamiento y testeo de los distintos modelos. Las cabeceras de dichas tablas dan cuenta del modo en que se ha organizado.

```
> 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
```

Cuadro 12: 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

#### 4.3.2. Bondad de ajuste de los modelos

La tabla *modelos* incorpora tres indicadores de la bondad de ajuste, uno referido a la etapa de entrenamiento ( $r^2$ ) y otros dos (RMSE y R2test) calculados en base al contraste de las predicciones con los valores reales en el conjunto test.

En la tablas siguientes se muestran los indicadores señalados tanto para el conjunto de todos los modelos como por tipo de modelo.

Cuadro 13: 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

Cuadro 14: 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

Los resultados muestran un buen ajuste para el conjunto de los modelos, con un R2test que para más de la mitad de ellos superan el 0.67. Pero más importante es que RMSE, raíz del error medio cuadrático, tiene un valor muy bajo, 17.59 en media y mediana de 15.51. Hay que tener en cuenta que RMSE se mide en las mismas unidades que la variable objetivo de predicción, en nuestro caso porcentaje de bicicletas disponibles.

La bondad de los modelos por tipo es también bastante buena, si bien, como era esperable con un notable incremento del error a medida que se dispone de menor información para la predicción. En los modelos con disponibilidad de información muy reciente (15min), la bondad del ajuste, se dispara con R2 muy próximo a

1 y RMSE entorno a 5. Para un horizonte de predicción de 4h, RMSE se sitúa entorno a 20, con 24h en 26 y para horizontes más lejanos, el RMSE está entorno a 32.

La figura siguiente muestra el progresivo incremento de RMSE a medida que se alarga el horizonte de predicción y han de utilizarse modelos con menor información reciente.

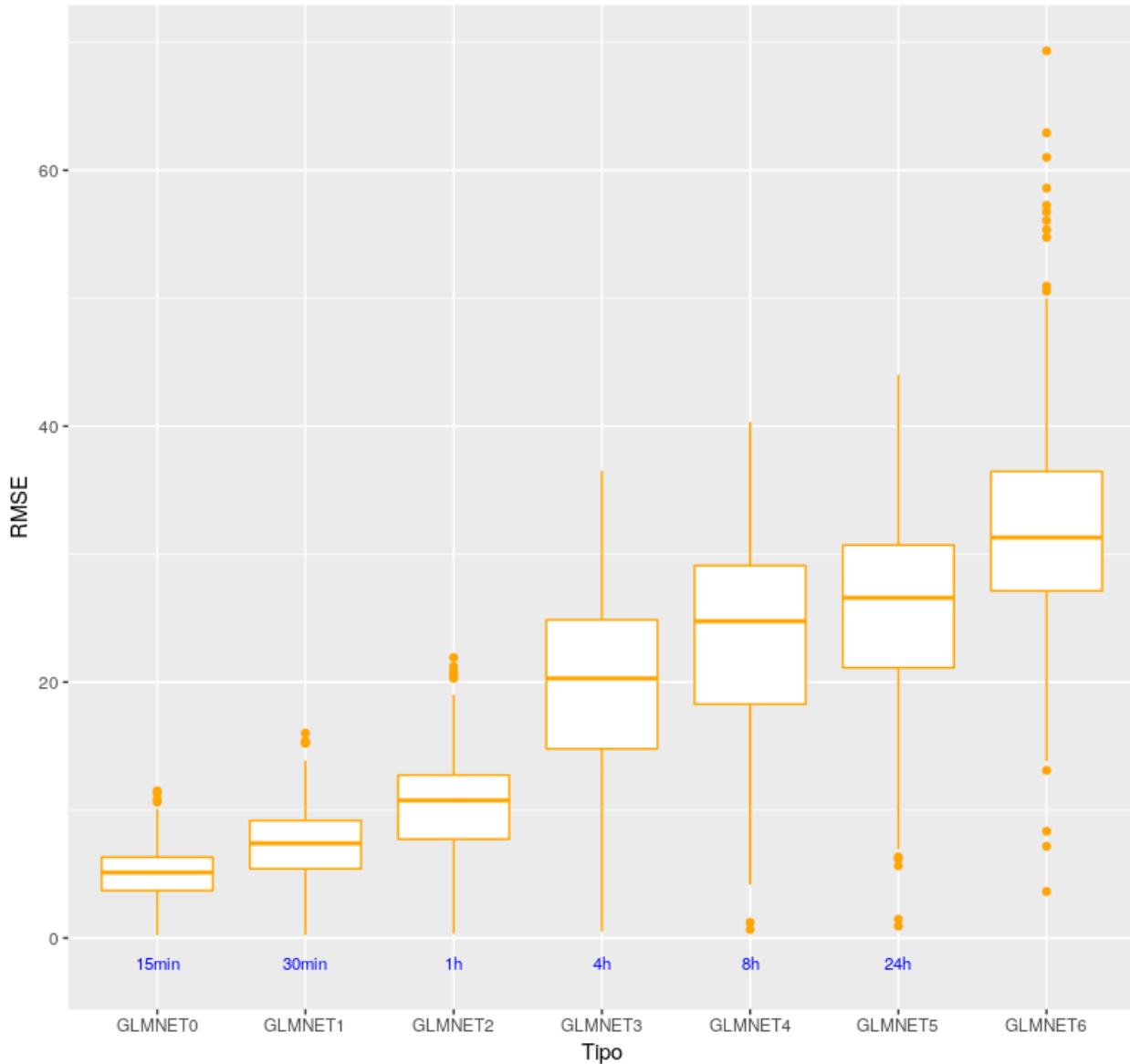


Figura 28: Bondad de ajuste. Raíz del error cuadrático medio (RMSE) por tipo de modelo.

Para cada estación se han estimado siete modelos, la figura siguiente muestra para cada uno de ellos su error (RMSE).

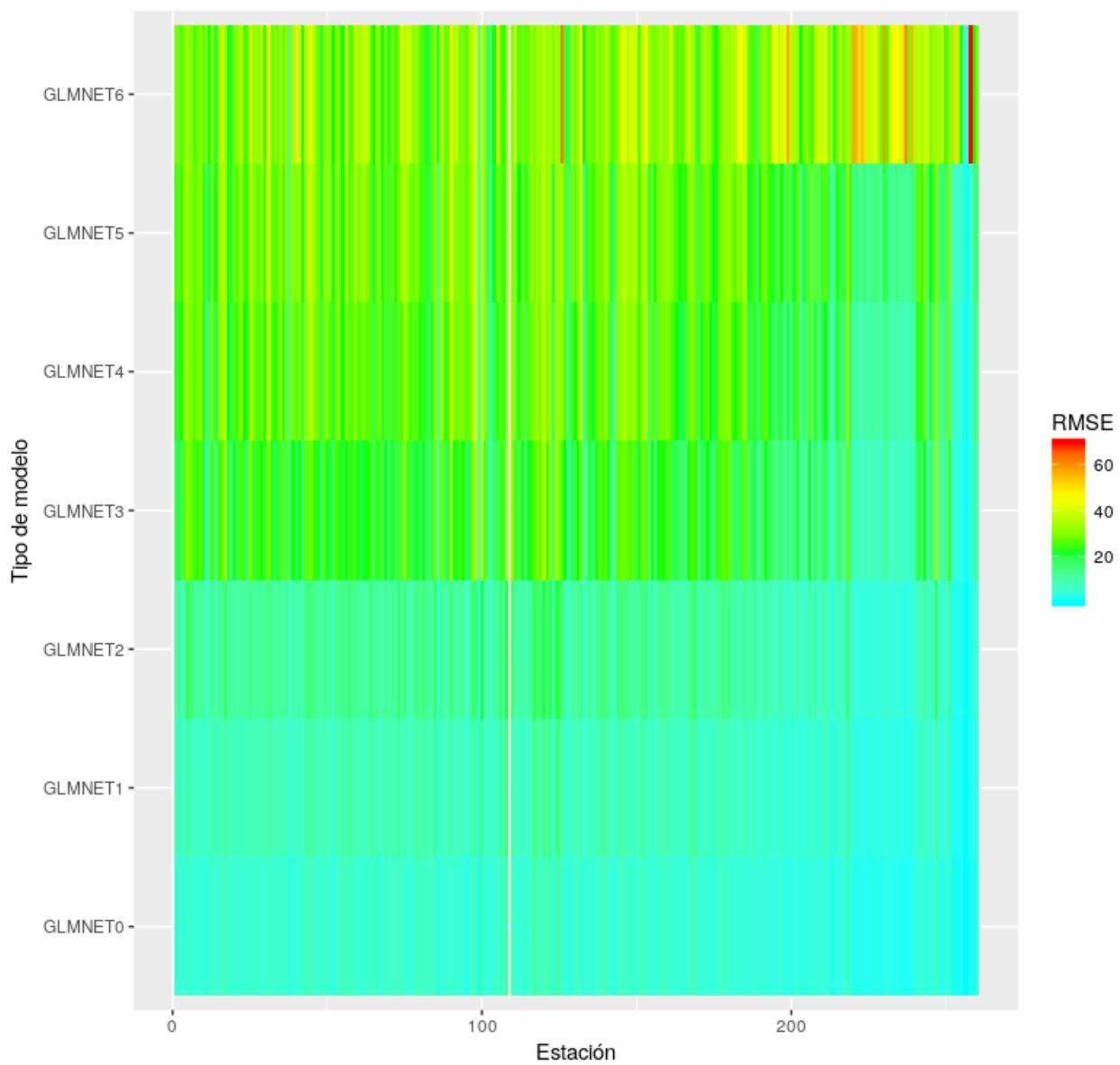


Figura 29: Errores (RMSE) por estación y tipo de modelo

#### 4.3.3. Residuos

La figura siguiente muestra la distribución de residuos por tipo de modelo.

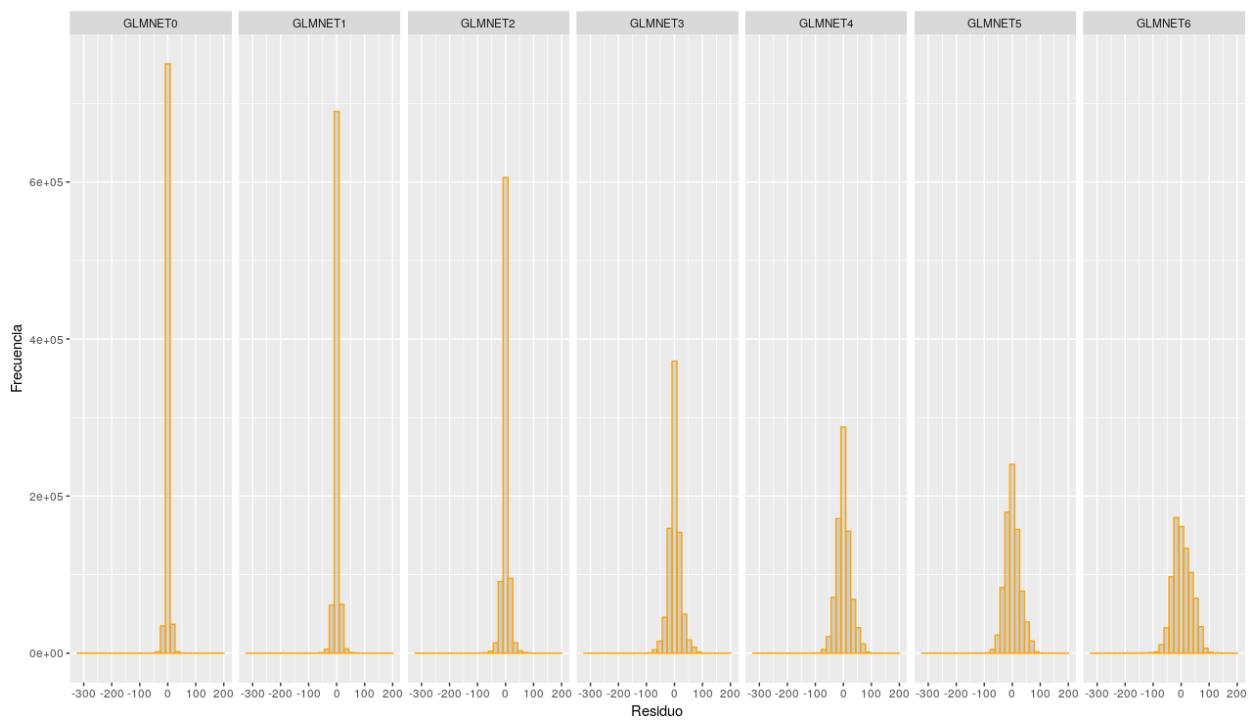


Figura 30: Distribución de residuos por tipo de modelo.

#### 4.3.4. Regresores significativos

La importancia relativa de cada uno de los regresores en el conjunto de modelos estimados se muestra en la figura siguiente, en la que se representa el número de modelos en los que dicho regresor aparece como significativo, según tipo de modelo.

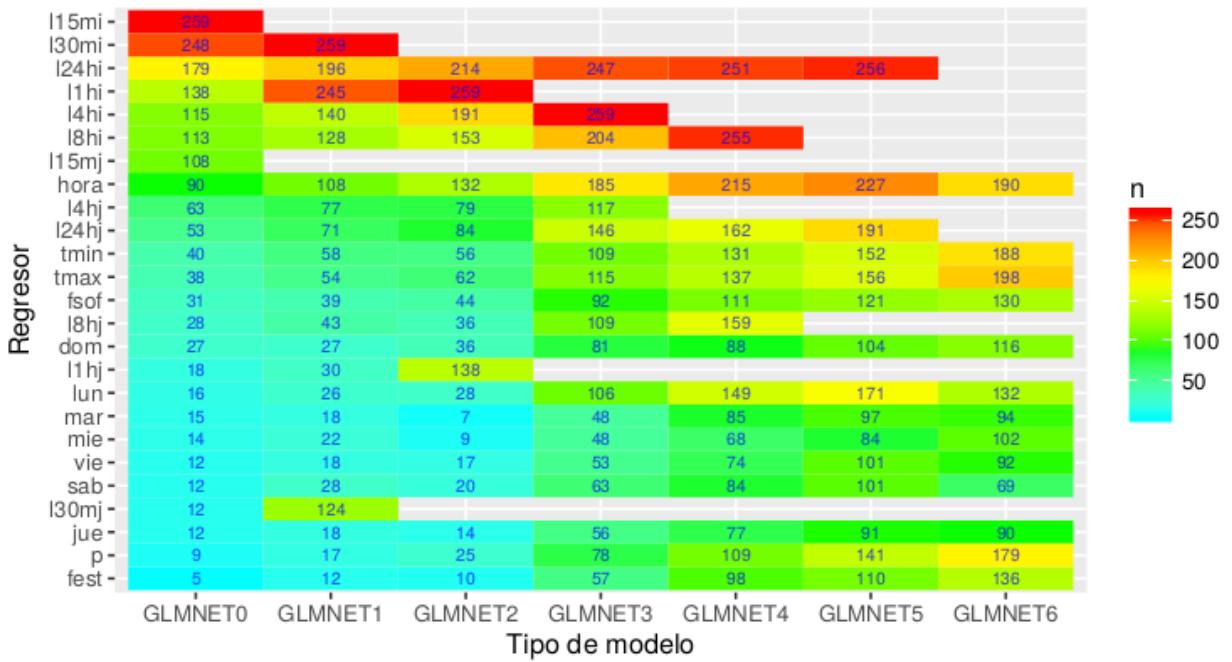


Figura 31: Frequencia de modelos con regresor significativo por tipo de modelo.

Vemos que para cada uno de los tipos de modelo, los regresores de retardo propio (de la misma estación) más recientes disponibles son los que más importancia tienen, apareciendo en la totalidad de estaciones. Esto es cierto para horizontes de predicción de hasta 4h (GLMNET3). En el caso de un horizonte de 8h (GLMNET4), tiene más importancia el retardo propio de 24h que el mismo de 8h.

El retardo propio de 24h tiene una gran importancia apareciendo en tercer lugar con 179 estaciones para modelos completos (GLMNET0).

Los regresores de retardo de vecino más próximo (j) tienen una importancia más limitada, el orden de importancia de los mismos es: 15min,4h,24h,8h y 30min.

La variable no retardada que aparece como más importante en conjunto es *hora*, alcanzando su máxima representación en los modelos con horizonte de 24h (GLMNET5), le siguen temperaturas; *tmin*, *tmax*.

En los modelos sin variables retardadas (GLMNET6) el orden de importancia de las variables, en los términos aquí considerados, es el siguiente: *tmax* > *hora* > *tmin* > *p* > *fest* > *lun* > *fsof* > ...

Tras todo lo expuesto se puede concluir que el modelo predictivo desarrollado responde de forma satisfactoria a los objetivos planteados inicialmente y es una buena base para un desarrollo y despliegue más ambicioso que puede ser contrastado con usuarios potenciales y gestores.