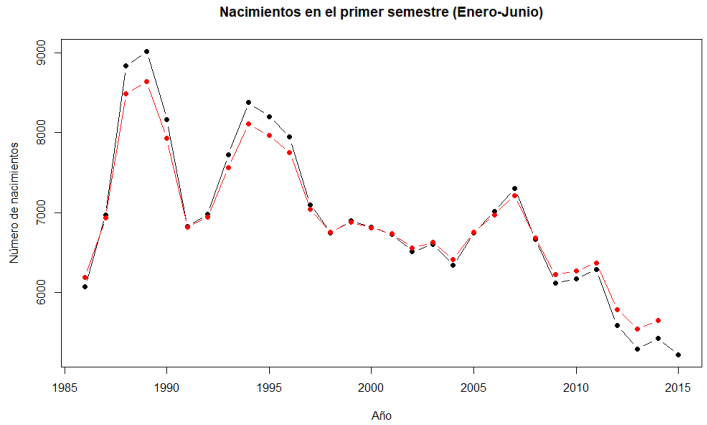
Examen OPI – Data Science
Soluciones por:
Vidal Alí González Cucurachi

Sección 1: Bebés

Para cada AGEB de la delegación Álvaro Obregón estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. Enlista tus fuentes y presenta los resultados.

Sea X el número de bebés de 0 a 6 meses en total en la delegación Álvaro Obregón, y sea p_j la probabilidad de que un bebé sea del AGEB j. Así, bajo estos supuestos, en total hay $p_j X$ bebés en el AGEB j. Para obtener los valores $p_j X$ para cada uno de los AGEBS, debemos estimar X y los p_j.

Primero, para estimar X vamos a considerar el número de nacimientos en la delegación, desde 1986 hasta 2015. Estos datos están disponibles en la página del INEGI en la sección de Natalidad y Fecundidad. Notemos que un bebé tiene entre 0 y 6 meses si y sólo si nació entre Enero y Junio de 2017. Así, lo que queremos estimar, X, es el número de bebés que nacieron en el primer semestre del año. Vamos a considerar la serie de tiempo $\{X_n\}_{n=1986}^{2015}$, donde X_n es el número de nacimientos en el primer semestre del año n. Ajustando un modelo ARMA(1,0) (bajo el criterio de Akaike es el que mejor se ajustó), obtenemos una predicción para los años 2016 y 2017 de 5467 y 5670 nacimientos en el primer semestre respectivamente. En la siguiente figura se muestra la serie de tiempo $\{X_n\}_{n=1986}^{2015}$ en negro, y el ajuste del modelo ARMA(1,0) en rojo.



Ahora bien, también debemos considerar la tasa de mortalidad infantil, la cual también se estimó para el 2017. En la página de la Comisión Nacional de Población se encontró para la delegación Álvaro Obregón una tasa del 12.66 muertes por cada 1000 nacimientos en el año 2005. Consultando información del INEGI, se tiene que dicha tasa va en disminución año con año, haciendo una extrapolación se obtuvo una tasa para el 2017 del 10.56 muertes por cada 1000 nacimientos en el año 2017. Con esto, obtenemos una estimación final de $\hat{X} = \mathbf{5640}$ nacimientos en el primer semestre del 2017.

Para estimar las probabilidades p_j para cada uno de los 199 AGEBS en Álvaro Obregón, se utilizaron los datos del CPV 2010. Se extrajeron el número de bebés de 0 a 2 años n_j para cada AGEB j. Así, una estimación para p_j es

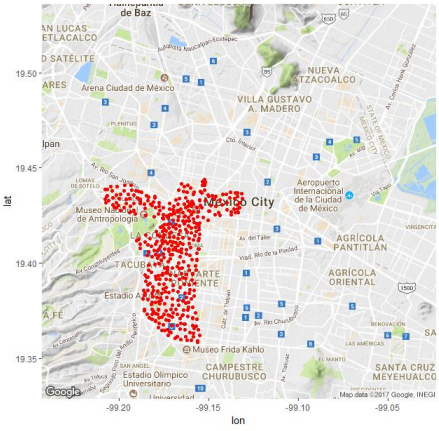
$$\hat{p}_j = \frac{n_j}{\sum_{i=1}^{199} n_i}.$$

Como es de esperarse, los valores $\hat{p}_j \hat{X}$ no son enteros. Para dar una estimación final para el número de bebés por AGEB, se redondearon estos números.

Datos de nacimientos
http://www.inegi.org.mx/sistemas/olap/Proyectos/bd/continuas/natalidad/nacimientos.asp
Tasa de mortalidad infantil
http://www.conapo.gob.mx/es/CONAPO/Base_de_datos
http://www.inegi.org.mx/saladeprensa/aproposito/2016/ni%C3%B1o2016_0.pdf
Datos poblacionales
http://www.beta.inegi.org.mx/proyectos/ccpv/2010/

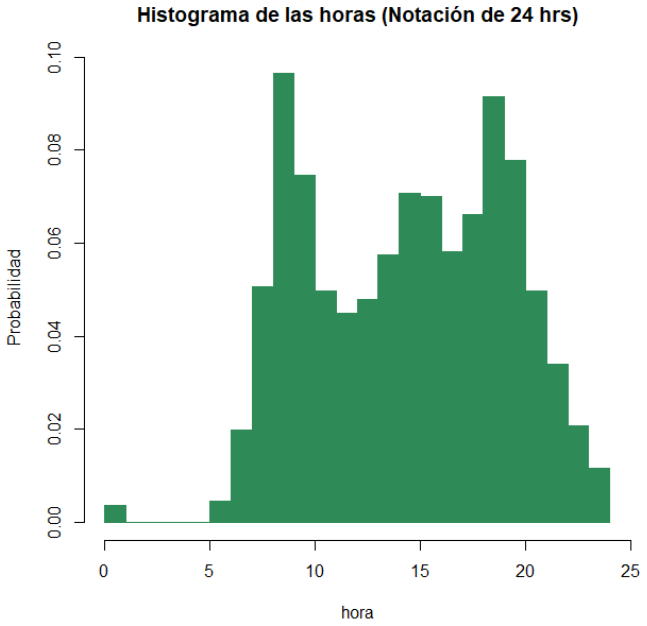
Sección 2.a: Ecobici

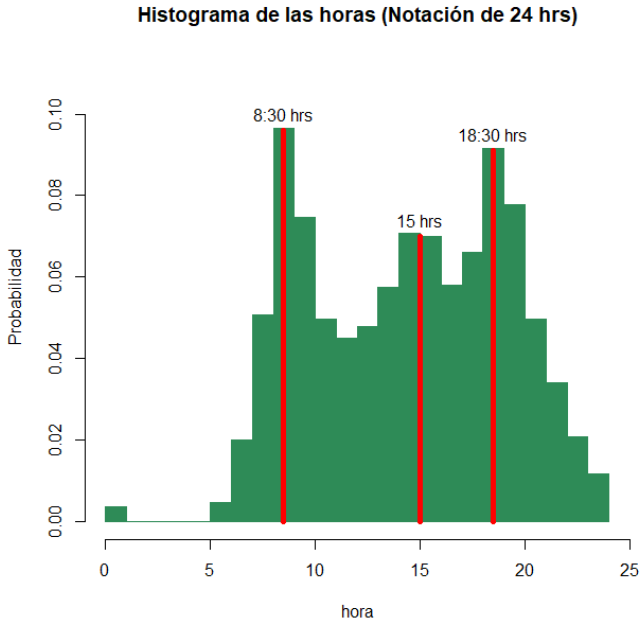
Se obtuvieron los datos correspondientes a Abril, Mayo y Junio con un total de 2,431,474 registros. Se tiene que hay 452 estaciones de ecobici en la Ciudad de México, distribuidas de la siguiente manera



1. ¿En qué horarios hay mayor afluencia y en qué estaciones? Da una breve descripción de por qué crees que es así

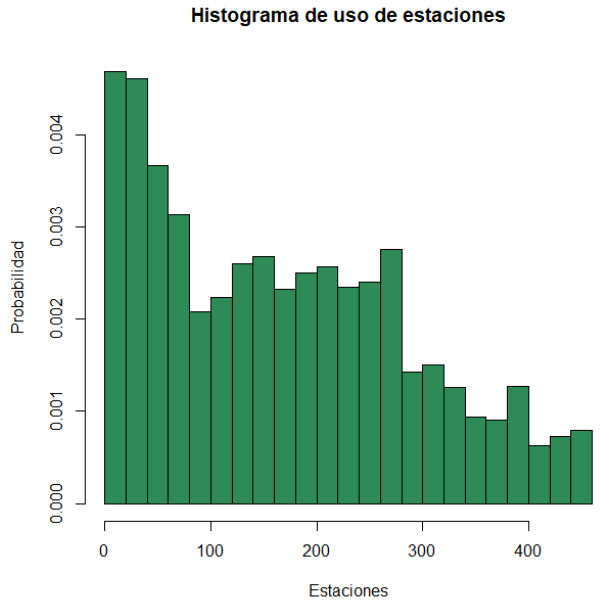
A continuación se muestran los histogramas de las horas en las que se retiró o se arribó una bicicleta a alguna estación.

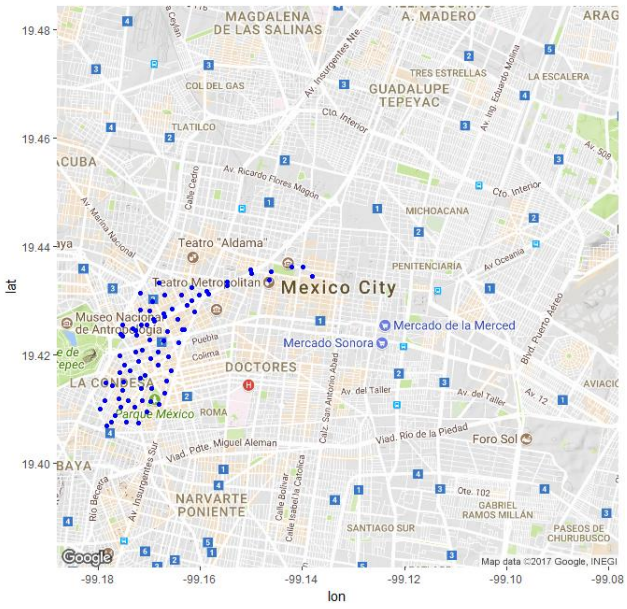




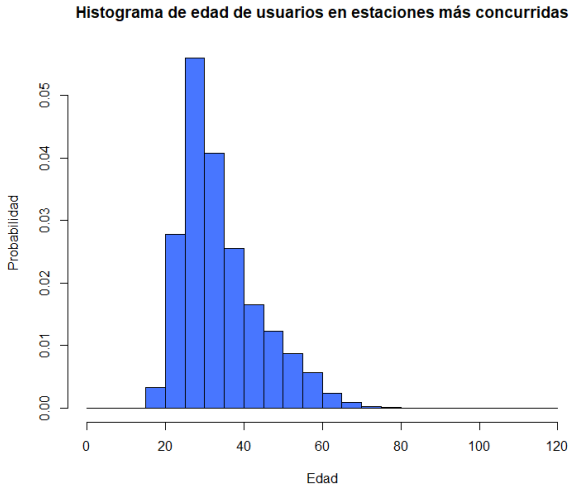
Se puede observar que en las horas en las que hay mayor actividad son: Entre 8 y 9 hrs, entre 14 y 16 hrs y por último entre 18 y 19 hrs. El motivo por lo que se concentra mayor actividad en estos intervalos es porque entre estos se encuentran los horarios (más comunes) de entrada, comida y salida en el trabajo, respectivamente. Nótese que entre 14 y 16 hrs hay un poco menos de actividad que en los otros dos intervalos, ya que muchas personas tienden a preferir comer dentro o cerca de su lugar de trabajo.

Ahora bien, a continuación se muestra el histograma del uso de cada estación. La mayor actividad se presenta en las estaciones que van de la 1 a la 80, las cuales corresponden a las estaciones marcadas con azul en el mapa.





A continuación se muestra el histograma de las edades de los usuarios que ocupan las estaciones en los horarios más concurridos. Observamos que las personas de entre 25 y 35 años son las que más registraron visitas a las estaciones. Notemos que la zona en la que se concentran las estaciones con mayor ínidice de uso, es una zona en la que viven y/o trabajan un gran número de personas precisamente con el rango de edad recién mencionado. Es debido a esto que sean las estaciones marcadas con azul en el mapa de arriba, las que presenten mayor afluencia de ususarios en las horas de mayor registro de actividad.



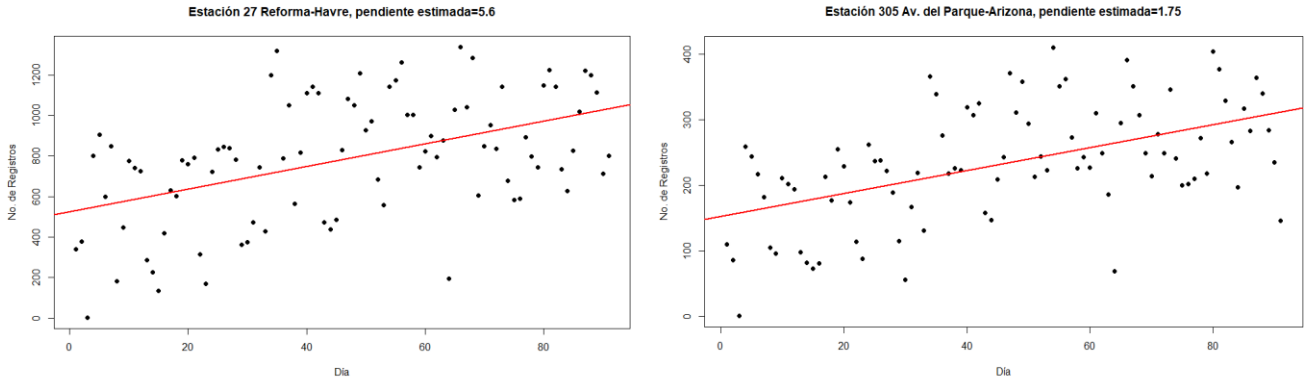
2. A partir de un análisis temporal:
a. ¿En qué estaciones puedes observar una tendencia de uso a la alta?
b. ¿Puedes categorizar las estaciones con base en su tendencia de uso?
c. Demuestra tus conclusiones gráficamente

Para responder a esta pregunta se realizó lo siguiente lo siguiente. Para cada estación se extrajo una serie de tiempo $\{X_n^j\}_{n=1}^{91}$, en donde $j = 1,2,\dots,452$ representa a cada estación y n a cada uno de los 91 días que comprenden entre Abril y Junio. Así, X_n^j representa el número de registros de actividad, tanto de retiro como de arribo, en la estación j el día n.
Para analizar la tendencia, se ajustó un modelo de regresión lineal simple tiempo vs. Registro de actividad para que con base en el valor de la estimación de la pendiente, decidir si se tiene una tendencia a la alta significativa.

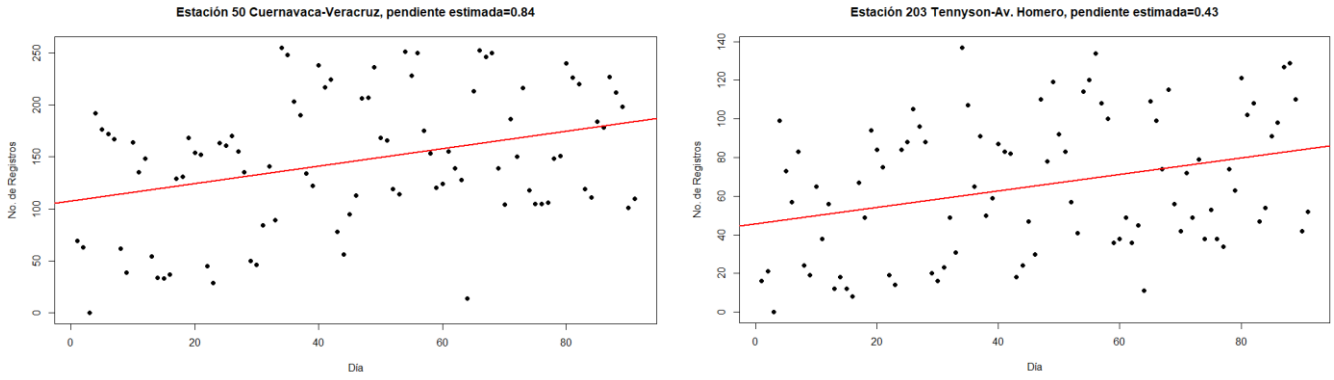
Cabe mencionar que hubo 3 estaciones: 31 (Hamburgo-Insurgentes), 213 (Solón-Cicerón) y 365 (Holbein-Av. Revolución), que no tuvieron registros en los tres meses que se analizaron. Consultando en el mapa de disponibilidad, se observó que actualmente la estación 31 y 365 se encuentran en estatus "No Operativa", de lo que se deduce que posiblemente esta sea la causa de la falta de registros en dichas estaciones. Además, hubo 3 estaciones: 40 (Oaxaca-Puebla), 315 (Alabama-Montana) y 405 (División del Norte-Municipio Libre) que obtuvieron una estimación de la pendiente negativa. Observando el comportamiento de las series de tiempo, en los tres casos se presentaron intervalos de varios días sin registros, lo que hace pensar que estuvieron fuera de servicio en esos periodos, posiblemente debido a mantenimiento.

Poniendo de lado las estaciones mencionadas, todas las restantes obtuvieron un valor estimado positivo de la pendiente. Una observación importante es la siguiente. Bajo el modelo de regresión lineal simple, una pendiente mayor a 1 representa un incremento por día de al menos un registro. Por esta razón, es razonable considerar al valor 1 como referencia para decidir si es significativa una tendencia a la alza en cada serie de tiempo. Así, si para una estación se obtuvo una estimación de la pendiente menor a 1, se considerará como una tendencia poco significativa o moderada, mientras que una estimación mayor a 1 será sustento para concluir una tendencia creciente significativa. Con el criterio que se acaba de describir es posible categorizar a cada una de las estaciones. A continuación se muestran dos ejemplos de tendencia a la alta significativa y dos de tendencia a la alta moderada (note la diferencia en el rango del eje Y para cada gráfica).

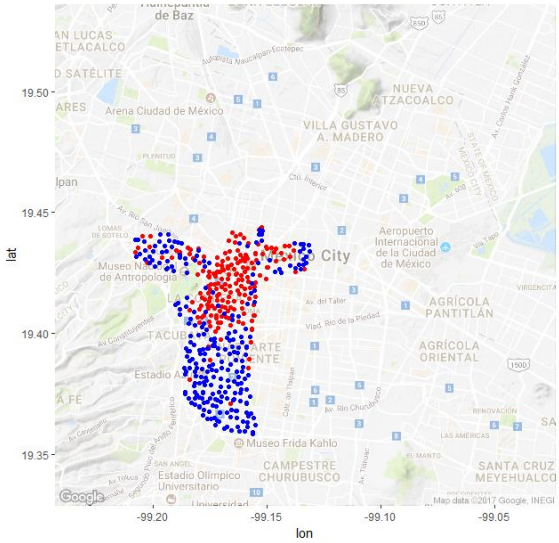
Tendencia a la alta significativa



Tendencia a la alta moderada

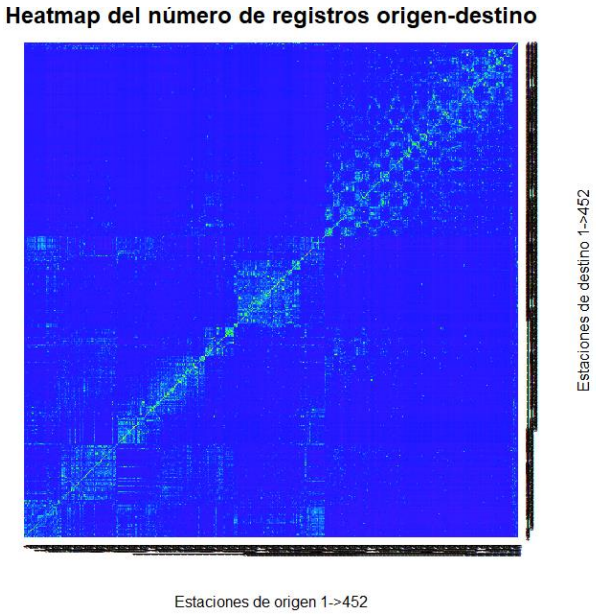


En el siguiente mapa se muestran en rojo las estaciones con tendencia a la alta significativa (en total 192), y con azul las estaciones con tendencia a la alta moderada (en total 254).



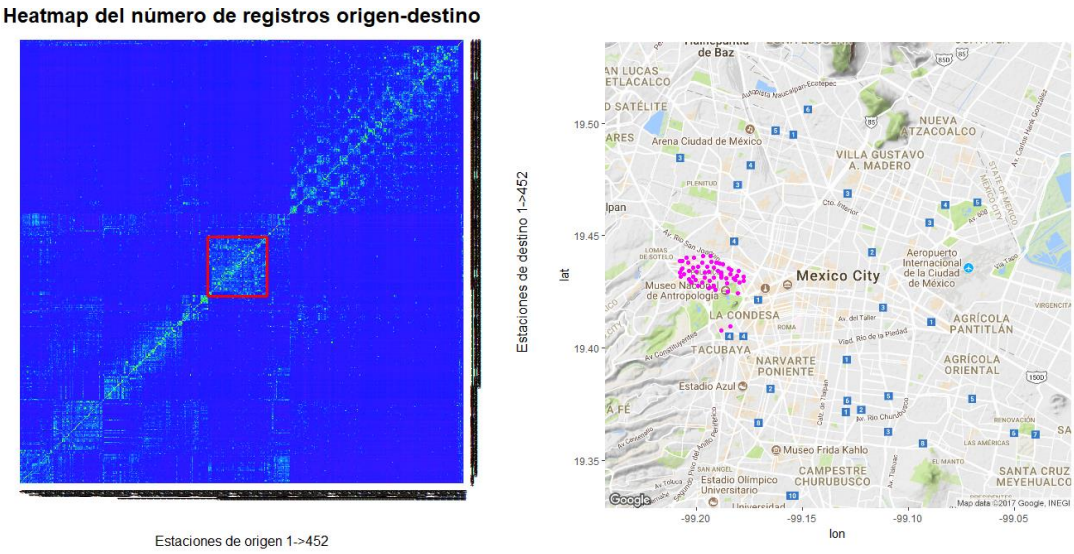
3. Por cada estación de Ecobici, identifica cómo están correlacionadas las entradas-salidas entre las otras estaciones (Hint: Puedes usar un heatmap para mostrar la correlación o matrices de origen-destino).

A continuación se muestra un heatmap del número de registros (estandarizados) con origen-destino en cada una de las estaciones. Se muestran los orígenes de izquierda a derecha, y los destinos de abajo hacia arriba. Tanto los orígenes como los destinos son el número de estaciones del 1 al 452. Se excluyeron las estaciones 31, 213 y 365, que son las que no tenían registros.



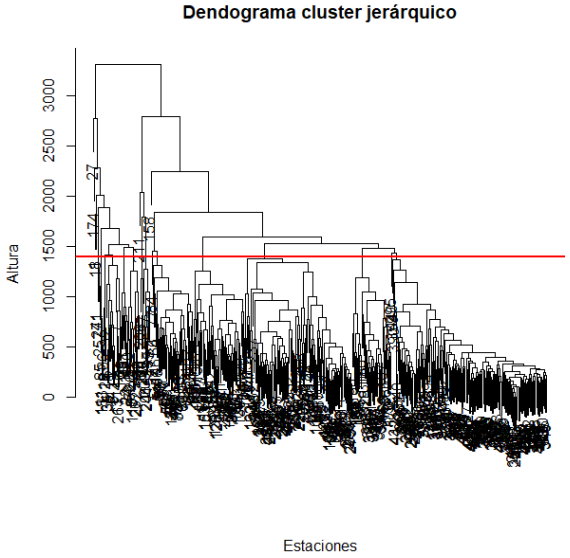
Como se puede observar alrededor de la diagonal del gráfico, existen mayor interacción entre estaciones con números cercanos. Esto corresponde a estaciones que por lo general se encuentran cercanas entre ellas. Como también podemos observar en el heatmap, hay algunas interacciones significativas alejadas de la diagonal, pero como se analizará en el siguiente inciso, son estaciones que se encuentran próximas geográficamente.

(BONUS) A continuación se muestra un mapa con las estaciones que comprenden los números del 190 al 250 (en rosa), correspondientes a las interacciones encerradas en el recuadro rojo. Como se puede analizar, dichas estaciones se encuentran geográficamente muy cercanas, que intuitivamente es la razón por la cual tienen un gran número de interacciones entre ellas, ya que por lo general una persona no recorre distancias demasiado largas en bicicleta. Veremos en la solución al siguiente inciso que, efectivamente, tomando en cuenta el comportamiento de entradas y salidas de cada estación, los grupos de estaciones que se forman se pueden explicar con su ubicación geográfica.



4. Usa un método de aprendizaje no supervisado para encontrar “perfiles de uso” de las estaciones. Lo que debes de hacer es categorizar a las estaciones en diferentes grupos a partir de su comportamiento de entradas y salidas. Explica qué método usaste y por qué. De los grupos que encontraste describe las características que puedes inferir de estos a partir de lo descubierto en el inciso anterior.

Vamos a considerar que cada columna y cada renglón de la matriz origen-destino representa el comportamiento de entradas y salidas de cada estación, respectivamente. Utilizaremos métodos de clustering para encontrar perfiles de uso de las estaciones. Primero, para explorar el número de grupos (clusters) que se forman, haremos uso del método de clústering jerárquico. A continuación se muestra el dendograma resultado del algoritmo. Haciendo un corte a la altura de 1400, obtenemos 8 clusters.



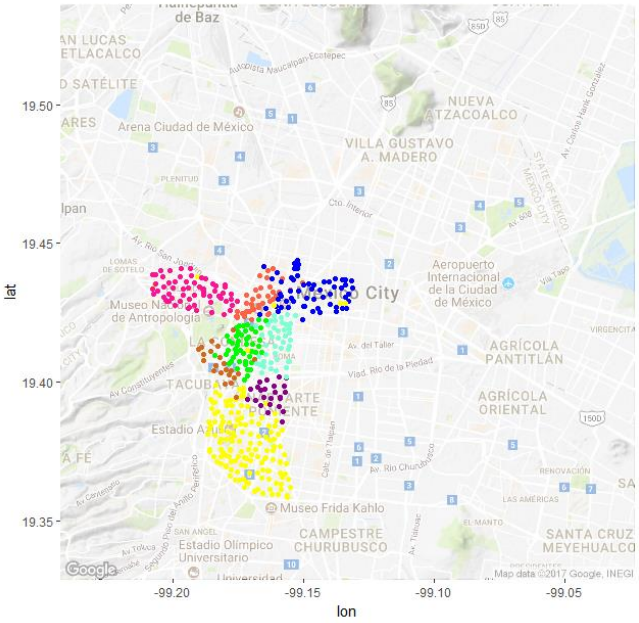
Ahora bien, identificando 8 potenciales agrupamientos, haremos uso del método de clustering espectral para encontrar los grupos en los que se dividen las estaciones. Se prefirió el método de clustering espectral, porque se ha demostrado que incluso cuando los datos presentan agrupamientos con geometría más compleja que simples elipsoides, se obtienen buenos resultados, en contraste con el método de k medias y jerárquico. (En Luxburg, U Von (2007) *A Tutorial on Spectral Clustering*, arXiv:0711.0189 se detallan algunas ventajas y expone algunas otras referencias sobre la comparación con otros métodos y sobre algunos resultados de consistencia). Debido a que nuestros datos tienen dimensión 452, no es posible visualizar qué tipo de geometría presenta.

Se analizaron los datos de comportamiento de salidas (los renglones de la matriz de origen-destino) y los de comportamiento de llegadas (las columnas de la matriz de origen-destino). A su vez, se juntó toda la información para obtener una matriz de *interacción*. Es decir, en dicha matriz de interacción, $(i,j) = (j,i)$ representa el número de

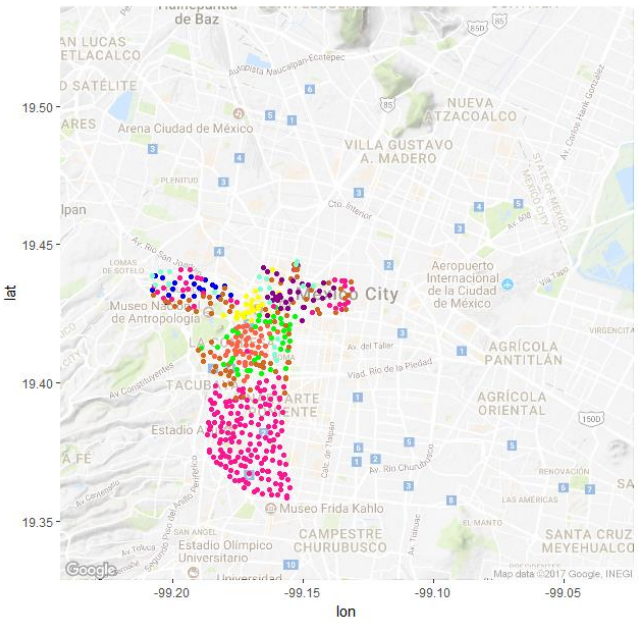
interacciones entre la estación i y la estación j, la suma de las llegadas y las salidas. A continuación se muestran, divididos por colores, los grupos que resultaron del análisis. Como se puede esperar, dichos grupos corresponden a regiones geográficas. Es consistente pensar que los destinos de las personas que toman una ecobici de cierta estación sean similares, y éstos se encuentren dentro de cierto radio geográfico. Por otro lado, haciendo uso del comportamiento de entradas, no se obtienen grupos tan bien identificados geográficamente para algunos clusters, debido a que a pesar de estar cerca dos estaciones, los orígenes de los cuales provienen los usuarios pueden ser muy diferentes.

Por último, agrupando haciendo uso de las *interacciones* (combinando entradas y salidas), geográficamente están bien identificados los clusters. Como se ha venido mencionando a lo largo del análisis, los viajes en bicicleta por lo general no son demasiado largos, por lo que una estación tiene mayor número de interacciones con estaciones cercanas a ella.

Clusters obtenidos con base en el comportamiento de salidas



Clusters obtenidos con base en el comportamiento de entradas



Clusters obtenidos con base en los comportamientos de entradas y salidas combinados