Proyecto
Análisis de Datos

Aguayo González Jaime Francisco
Barajas Vega José Trinidad

Diciembre 7, 2021

Sobre el proyecto

En este trabajo analizamos los datos abiertos del programa de renta de bicicletas "Mi Bici" de la zona metropolitana de Guadalajara. En particular nos interesa conocer el impacto que han tenido políticas de ampliación del programa y las políticas de distanciamiento social derivadas de la pandemia. Con este análisis esperamos brindar información que permita evaluar el programa de renta de bicicletas.

Para hacer este análisis, el presente reporte se divide de la siguiente forma:

- En la sección **Análisis Exploratorio de Datos** describimos brevemente la información que incluye los datos y algunos descubrimientos que hicimos explorando y graficando algunas variables
- En la sección **Cambios de Estructura** explicamos cómo es posible obtener puntos de quiebre donde la estructura de la serie de tiempo cambia, además interpretamos los resultados obtenidos al usar dichos métodos
- En la sección **Análisis de las estaciones** tratamos de proporcionar gráficas informativas sobre el uso de las estaciones en las diferentes etapas del programa
- Finalmente, en **Conclusiones y Comentarios Finales** damos algunas conclusiones y trabajo a desarrollar en el futuro

Sin más preámbulo, pasamos a describir la limpieza de los datos y los resultados de su análisis exploratorio.

Análisis Exploratorio de Datos

Los datos proporcionados por Mi Bici están disposibles en agregados mensuales. Cada archivo consiste en una tabla que proporciona para cada viaje:

- Un Identificador único del viaje
- El identificador del usuario que realiza el viaje
- Género del usuario
- Año de nacimiento del usuario
- Fecha y hora del inicio del viaje
- Fecha y hora del fin del viaje
- Identificador de la estación de origen
- Indentificación de la estación destino

Como para responder nuestras preguntas, solo necesitamos el número de viajes por intervalo de tiempo, decidimos contar el número de viajes realizados al día. El resultado fue el siguiente:

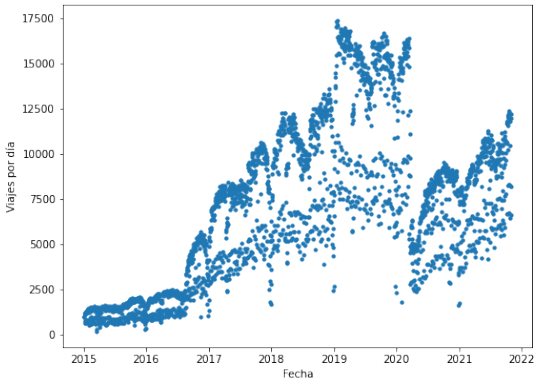


Figure 1: Número de viajes realizados por día.

En esta gráfica se observa el número de viajes por día, desde el enero de 2015 hasta octubre de 2021.

En la gráfica se aprecia como hay una gran variabilidad de un día a otro. De hecho, podemos encontrar al menos dos grupos diferenciados, uno arriba y otro abajo. Para distiguir los días de mejor manera, agrupamos por días de la semana obtieniendo lo siguiente:

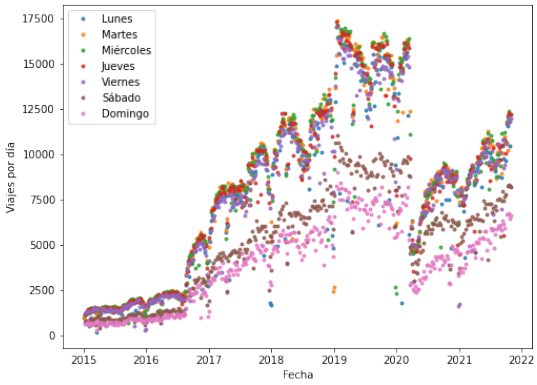


Figure 2: Número de viajes realizados por día, clasificación por día de la semana.

Podemos observar como sábados y domingos tenemos una baja en el número de viajes. Más evidente es en la siguiente gráfica.

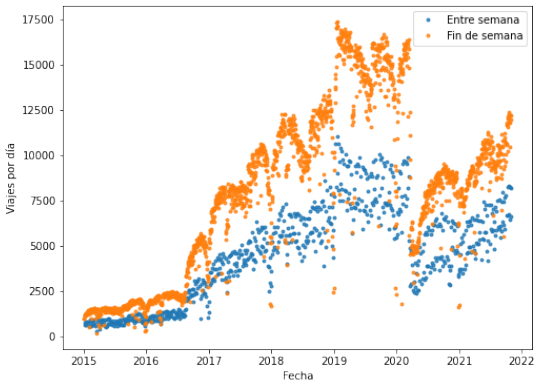


Figure 3: Número de viajes realizados por día, dia laborable vs. fin de semana.

Puesto que no queremos que nuestros métodos para evaluar el cambio de estructura se vean afectados por el cambio en los días de la semana, decidimos agregar el número de viajes realizados cada semana (empezando en lunes). El número de viajes por semana presenta la estructura de la fig. 4.

Se observa una serie de tiempo más uniforme con pequeñas caidas en lo que suponemos son semanas con fines de semana largos o vacaciones. De ya se puede apreciar una caida en el número de viajes a inicio del 2020, seguramente

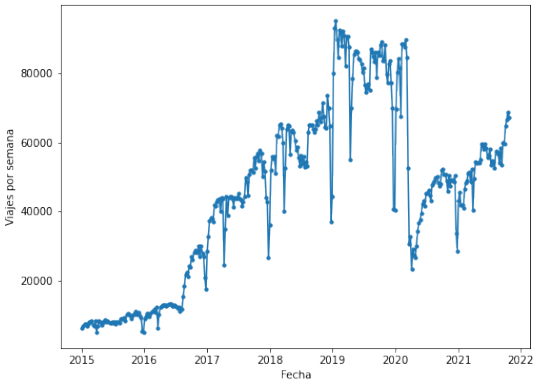
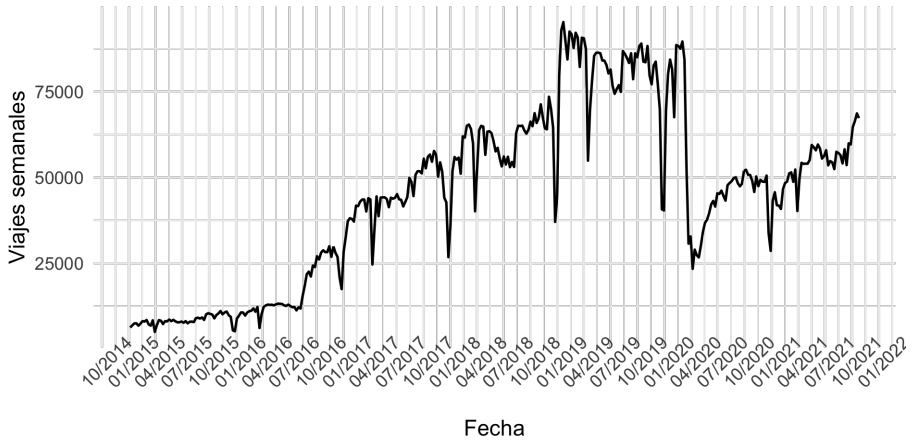


Figure 4: Número de viajes realizados por semana.

provocada por las medidas de confinamiento implementadas en la entidad al inicio de la pandemia.

También nos interesó contabilizar el número de viajes realizados entre pares de estaciones, el prrocesamiento de esos datos se detalla en la penúltima sección.

Cambios de estructura



Dado un conjunto de datos $\{(Y_i, X_i)\}_{i=1}^n$ con $X_i \in \mathbb{R}$, suponemos que existen k intervalos $[a_i, a_{i+1}]$ para $i \in \{1, \ldots, k\}$, de forma tal que es posible dividir las observaciones en estos intervalos, es decir para índices en I_j , $(Y_i, X_i)_{i \in I_j} \in [a_j, a_j + 1]$ de forma tal que hay una relación lineal entre estos datos que puede o no estar relacionada con el modelo lineal de los otros intervalos.

¿Bajo qué condiciones se puede dar este supuesto? Es posible que se tenga un experimento a través del tiempo donde las condiciones del experimento cambian por una cierta temporalidad. En nuestro caso queremos probar la hipótesis de que el comportamiento del número de viajes realizados ha cambiado por la pandemia o no.

De ser cierta la hipótesis, se dice que los datos presentan un **cambio de estructura** y a los puntos $a_2, \ldots, a_{k-1}$ como **punto de quiebre**. El problema cambia dependiendo de la información que poseemos.

Estudiemos primero el caso donde conocemos que existe un solo punto de quiebre y nos interesa evaluar la hipótesis

H_0 : no existe cambio de estructura.

La hipótesis puede ser descrita formalmente de la siguiente forma:

$H_0 : Y_i = \beta X_i + u_i, \quad \forall i \in \{1, \ldots, n\},$

mientras que la hipótesis alternativa sería:

$$H_1 : \exists \quad i_0 \in \{1, \ldots, n\} \text{ tal que } \begin{cases} Y_i = \beta_A X_i + u_i & 1 \leq i \leq i_0 \\ Y_i = \beta_B X_i + u_i & i_0 < i \leq n \end{cases} , \beta_A \neq \beta_B.$$

Chow (1960) propuso un estadísitico que puede ser usado para probar esta hipótesis cuando el valor i_0 es conocido. Para ello propone ajustar dos modelos

de regresión lineal de forma independiente para cada uno de los dos intervalos definidos por i_0, y rechazar la hipótesis alternativa si el cociente:

$$F_{i_0} = \frac{\hat{u}^\top \hat{u} - \hat{e}^\top \hat{e}}{\hat{e}^\top \hat{e}/(n-2)}$$

es mayor que un cierto valor, donde $\hat{e} = (\hat{u}_A, \hat{u}_B)\top$ son los residuales de cada modelo de regresión ajustado de forrma independiente en su intervalo y $\hat{u}$ son los residuales del modelo lineal ajustado a toda la muestra.

Chow demuestra que bajo H_0, el estadísitico se distribuye asintóticamente una χ^2_1 y, si $u_i \sim \mathcal{N}$, el estadísitico F_{i_0} sigue una distribución F con $n-2$ grados de libertad.

Sin embargo, dado que desconocemos el valor de i_0, Andrews (1993) propone calcular el estadístico F para todos los posibles puntos de quiebre y rechazar la hipótesis nula si para alguno se llega a un valor alto. Para deterrminar si alguno de los valores llega a ser alto, se puede usar el estadístico:

$$\sup F = \sup_{1<i<n} F_i.$$

Usando este estadístico de prueba, evaluamos la hipótesis nula para saber si existe evidencia suficiente para suponer que la pandemia ha provocado un cambio en la estructura del número de viajes que se realizan en el sitema Mi Bici de Guadalajara.

Experimentos con un solo punto de quiebre

Usando la librería `strucchange` de `R`, calculamos estadístico F para todo punto en nuestros datos. Esto se puede realizar con la función `Fstats`. Observe que para los modelos de regresión usamos como único predictor el tiempo.

```
res <- Fstats(trips ~ date + 1, data = data) # F statistics

plot(res2, xaxt = "n", xlab = "") # Plot F statistics
axis(1, labels=labels, at = ticks/length(data$X), las=2)
# Print line at i arg sup F
breakpoints(res2)
lines(breakpoints(res2))
```

En los resultados observmos un linea punteada el posible punto de quiebre para el cual se obtiene el valor del estadísitico F más alto.

```
 Optimal 2-segment partition:

Call:
breakpoints.Fstats(obj = res2)
```

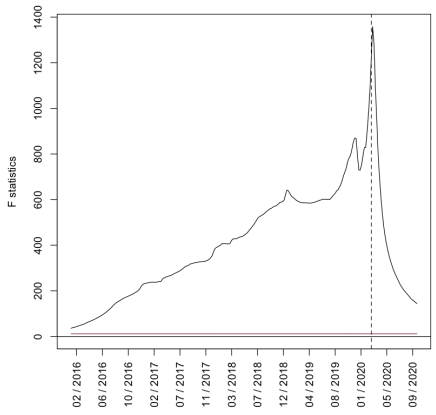


Figure 5: F-estadístico para cada posible punto de quiebre. Observe un máximo a principios de 2020.

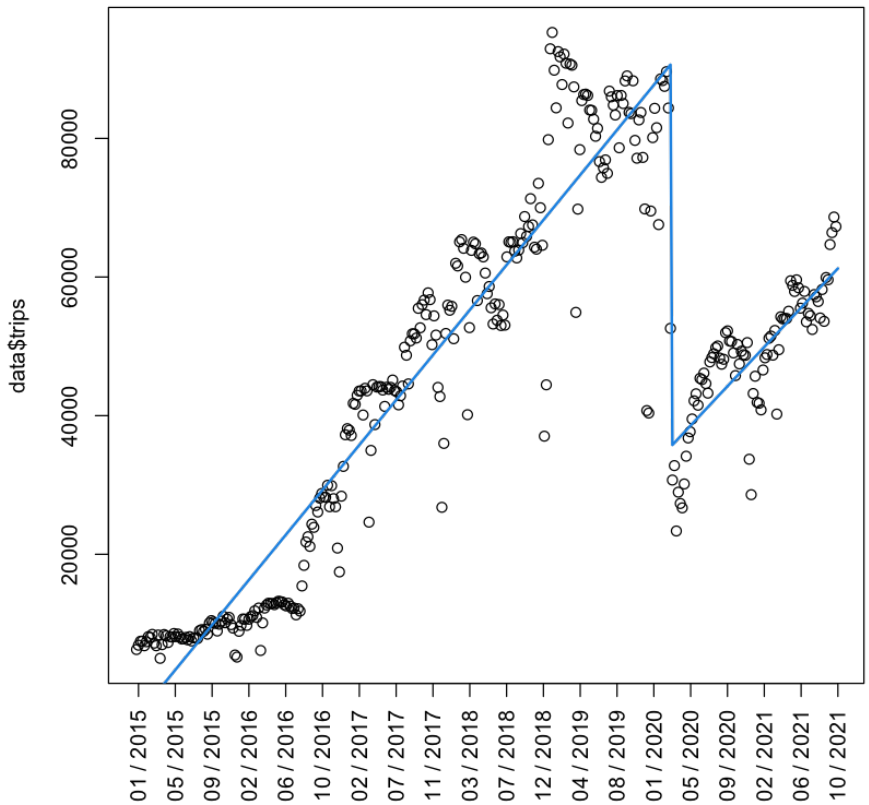
Note que el máximo lo obtenemos en fechas cercanas a febrero de 2020. Además observamos como todos los valores rebasan la linea roja que corresponde al valor máximo de la prueba a nivel $\alpha = 0.05$. Quiere decir que con un nivel del 95 %, rechazamos la hipótesis nula para cualquier punto.

El p-valor sorrespondiente a $\sup F$ es menor a 10^{-15}. Practicamente 0.

```
sctest(res, type="supF")
```

```
supF test

data:  res2
sup.F = 1358.7, p-value < 2.2e-16
```

Podemos graficar los modelos de regresión que resultan para cada intervalo. Estos se ven así:



Se aprecia como ambas rectas aproximan bien los datos. Sin embargo observamos que existe un comportamiento diferente en 2015 y 2016, comparrado con 2017 a 2019. Esto puede ser un indicativo de más cambios de estructura que los que contemplamos originalmente.

Antes de evaluar esta posibilidad, vamos a ubicar con precisión el punto de quiebre obtenido y evaluar la calidad de los modelos lineales ajustados.

```
bp <- breakpoints(res)
```

```
data$date[bp$breakpoint]
```

2020-03-09

Se observa que el punto de quiebre obtenido corresponde a la semana del 9 de marzo de 2020, que corresponde con la semana de suspensión de clases presenciales es el estado de Jalisco.

Por otro lado, los modelos de regresión parecen ajustarse bien a los datos:

```
fm1 <- lm(trips ~ breakfactor(bp)/date - 1, data = data)
```

```
summary(fm1)
```

```
Call:
lm(formula = trips ~ breakfactor(bp)/date - 1, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-46783  -4269    467   5072  25152

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
breakfactor(bp)segment1      -8.191e+05  1.614e+04 -50.743  < 2e-16 ***
breakfactor(bp)segment2      -7.583e+05  9.845e+04  -7.702 1.37e-13 ***
breakfactor(bp)segment1:date  4.963e+01  9.281e-01  53.478  < 2e-16 ***
breakfactor(bp)segment2:date  4.330e+01  5.284e+00   8.195 4.70e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8367 on 352 degrees of freedom
Multiple R-squared:  0.974,Adjusted R-squared:  0.9737
F-statistic:  3297 on 4 and 352 DF,  p-value: < 2.2e-16
```

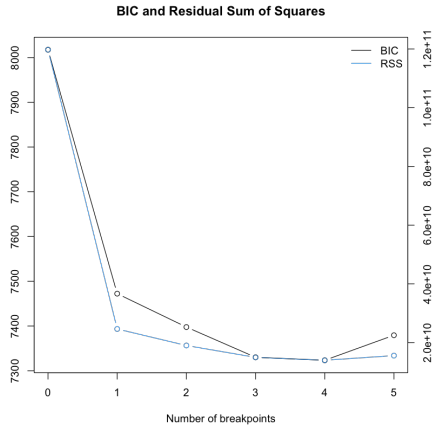
Los p-valores del estadístico T son muy chicos y $R^2 \approx 0.97$, muy cerca de 1.

Experimentos con múltiples puntos de quiebre

Dado que para muchos valores obtenemos un valor alto de F y por la inspección visual de los datos, es razonable pensar en la existencia de más puntos de quiebre en nuestros datos. Para ello podemos usar el mismo razonamiento hecho con anterioridad, pero ahora evaluar el estadísitico para los dos intervalos obtenidos. Esto es, hacer un análisis para el intervalo que va del primero de enero de 2015 al 9 de marzo de 2020, y otro para el intervalo del 9 de marzo de 2020 a la fecha.

Sin embargo este enfoque solo nos va a permitir determinar un punto de quiebre a la vez. Los autores de la librería `strucchange` ponen a nuestra disposición algunos métodos más elaborados para la detección de multiples puntos de quiebre. Dentro de sus propuestas se considera un proceso de fluctuación basado en estimar de forma recursiva modelos de regresión lineal para diferentes ventanas de tiempo. Dado que el estudio de estos procesos de fluctuación están fuera del alcance del curso, nos limitamos a utilizar dicha función y discutir los modelos de regresión que se ajustan con los nuevos puntos de quiebre.

La función `efp(trips ~ date + 1, data = data, type = "ME")` realiza el análisis y regresa los posibles puntos de quiebre. Lo que es de nuestro interés, es analizar el *residual sum of squares* cuando ajustamos un solo modelo de regresión lineal, dos modelos, tres modelos o cuatro modelos dependiendo del número de quiebres que obtenemos. Esto se puede apreciar en la siguiente gráfica:



Podemos destacar que el cambio en el error disminuye considerablemente de 0 a 1 punto de quiebre. Este punto corresponde al ya analizado. Por otro lado, se observa que después del cuarto punto de quiebre, el error aumenta. Después de una inspección visual de los modelos de regresión, decidimos quedarnos con los primeros tres puntos de quiebre, pues se ajustan muy bien a los datos como se muestra en la siguiente imagen:

En rojo se observa el intervalo de confianza del punto de quiebre utilizando el estadístico F. Para evaluar la calidad de nuestro modelos de regresión, observamos

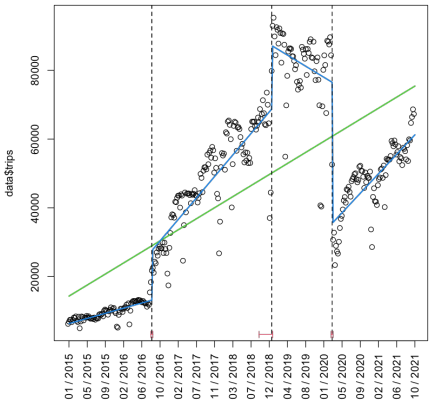


Figure 6:

el resumen que arroja R:

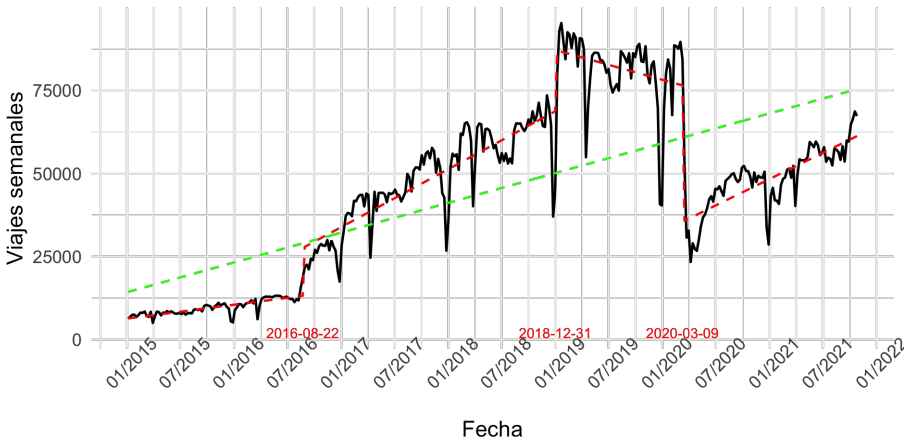
```
summary(fm1)
```

```
Call:
lm(formula = trips ~ breakfactor(bp.bikes2)/date - 1, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-37872  -1903    492   3757  16818

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
breakfactor(bp.bikes2)segment1      -1.816e+05  6.812e+04  -2.666 0.008041 **
breakfactor(bp.bikes2)segment2      -7.867e+05  4.157e+04 -18.925  < 2e-16 ***
breakfactor(bp.bikes2)segment3       5.271e+05  1.205e+05   4.375  1.6e-05 ***
breakfactor(bp.bikes2)segment4      -7.583e+05  7.717e+04  -9.826  < 2e-16 ***
breakfactor(bp.bikes2)segment1:date  1.143e+01  4.070e+00   2.809 0.005249 **
breakfactor(bp.bikes2)segment2:date  4.780e+01  2.379e+00  20.089  < 2e-16 ***
breakfactor(bp.bikes2)segment3:date -2.458e+01  6.649e+00  -3.697 0.000254 ***
breakfactor(bp.bikes2)segment4:date  4.330e+01  4.142e+00  10.455  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6558 on 348 degrees of freedom
Multiple R-squared:  0.9842,Adjusted R-squared:  0.9838
F-statistic:  2711 on 8 and 348 DF,  p-value: < 2.2e-16
```

Vemos que el p-valor del estadístico T para cada coeficiente es muy bajo, menor

a 10^{-3} en la mayoría de los casos. Además vemos que R^2 es muy cercano a 1, lo que habla muy bien de nuestros modelos.

Por último graficamos la serie de tiempo, en rojo los modelos de regresión lineal para cada intervalo, en verde el modelo lineal ajustado a toda la serie y en en rojo las fechas estimadas de los puntos de quiebre.



¿A qué se pueden deber los puntos de quiebre?

Tratando de explicar los cambios de estructura, nos dimos a la tarea de investigar noticias relativas al programa Mi Bici en fechas cercanas a los puntos de quiebre. Los resultados fueron los siguientes:

- El octubre de 2016 se inauguró la segunda etapa del programa, con la ampliación del número de estaciones.
- En noviembre de 2018 se inauguró la tercera etapa del programa.
- El 3 de marzo de 2020 se suspendieron las clases presenciales en el estado de Jalisco.

Salvo la suspensión de clases, los punto de quiebre no corresponden exactamente a las inauguraciones del número de estaciones, pero están fuertemente relacionadas pues, de acuerdo a los datos reportados, las estaciones ya se encontraban en uso semanas antes de la inauguración. Como ejemplo, la semana del 15 de agosoto de 2016 se dio una apliación de aproximadamente 120 a 234 estaciones en uso. Como comparación, en el análsis localizamos un punto de quiebre en la semana del 22 de agosoto del mismo año.

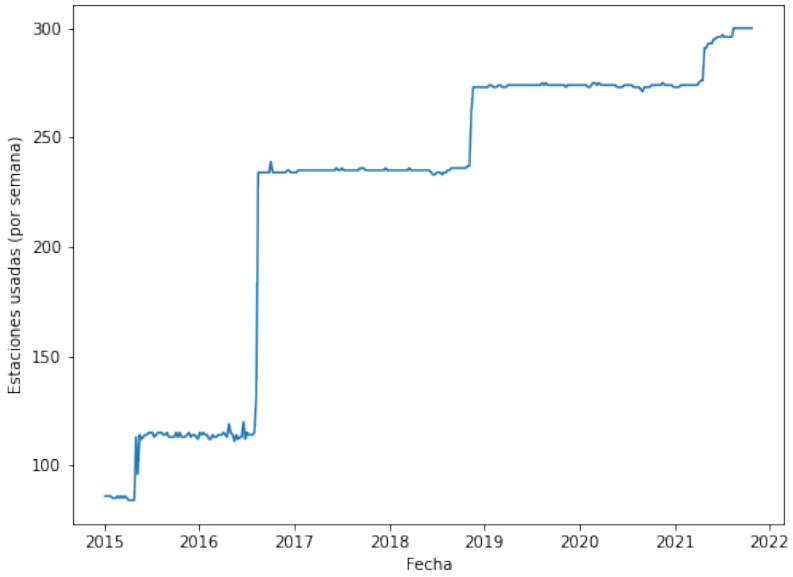
En el siguiente capítulo se abarca un poco más a detalle el impacto en la ampliación del número de estaciones.

Análisis de las estaciones

En esta sección exploramos los viajes entre puntos de recogida de bicicletas, puntos a los que llamaremos estaciones. La idea surge de las observaciones de que hay un comportamiento diferente a partir de la apertura de nuevas estaciones. Para poder hacer observaciones nos apoyamos en herramientas de *R* así como de *Python*.

Se tomó el acumulado de viajes mensuales y con ello se generaron matrices con la siguiente cualidad:
Si había un viaje de la estación *i* hacia la estación *j* entonces a la entrada *(i,j)* de la matriz se le suma 1.

La siguiente gráfica muestra la cantidad de diferentes estaciones que registraron viajes semanalmente.



Observemos que hay 4 aumentos importantes en cuanto a la cantidad de estaciones diferentes, por ello exploramos como se se comporta la interacción en estas etapas.

En las gráficas tipo HeatMap se muestra la interacción de viajes, donde oscuro significa nula o casi nula interacción. Cada una de ellas fue calculada promediando las interacciones mensuales de acuerdo a la duración de la etapa.

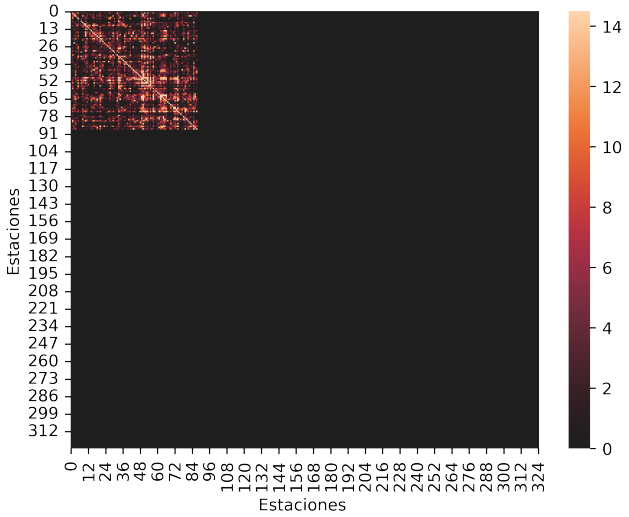


Figure 7: Etapa 1

Etapa Uno tomando los primeros 4 meses del registro, de Enero de 2015 a Abril de 2015.

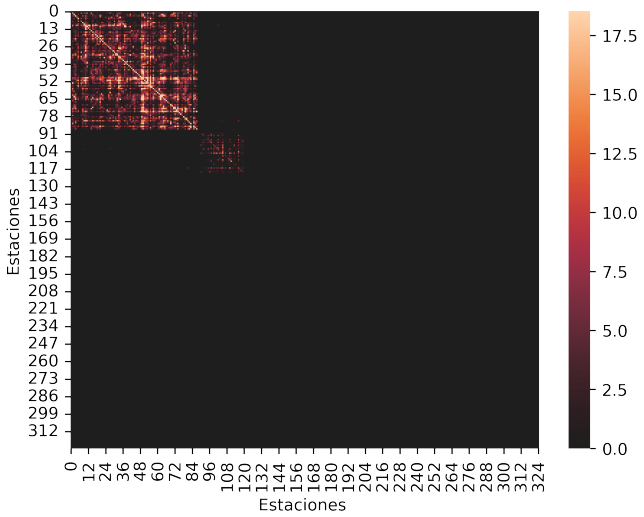


Figure 8: Etapa 2

Etapa Dos tomando de Mayo de 2015 a Julio de 2016.

Podemos observar que la primera ampliación no causa un gran inmpacto y las estaciones agregadas no tienen mucha interacción con las estaciones iniciales, pues al agregarse no se nota un cambio visible en el mapa de calor.

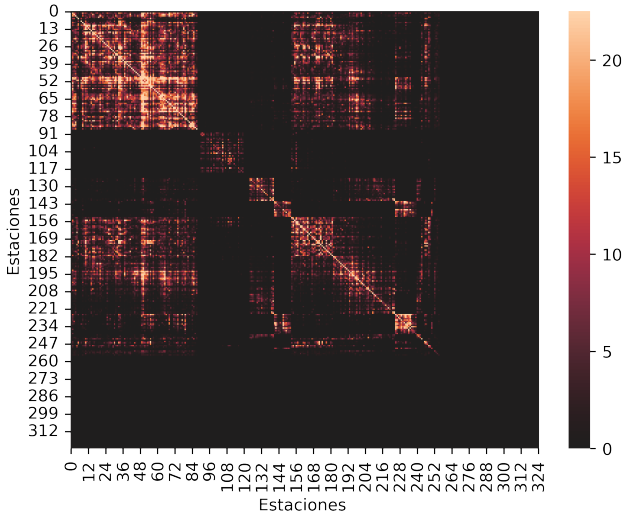


Figure 9: Etapa 3

Etapa Tres tomando de Agosto de 2016 a Octubre de 2018.

Después de la segunda ampliación se observa un claro incremento en la actividad, no solo entre las nuevas estaciones, sino también con las estaciones de la primera estapa pues se nota una iluminación en los puntos correspondientes a esta comunicación. También hay un incremento en las cantidas de viajes entre estaciones, lo que sugiere que incrementó la actividad en general, podemos asumir que hay evidencia visual de que en esta ampliación hubo un cambio significativo en las cantidades de viajes en general.

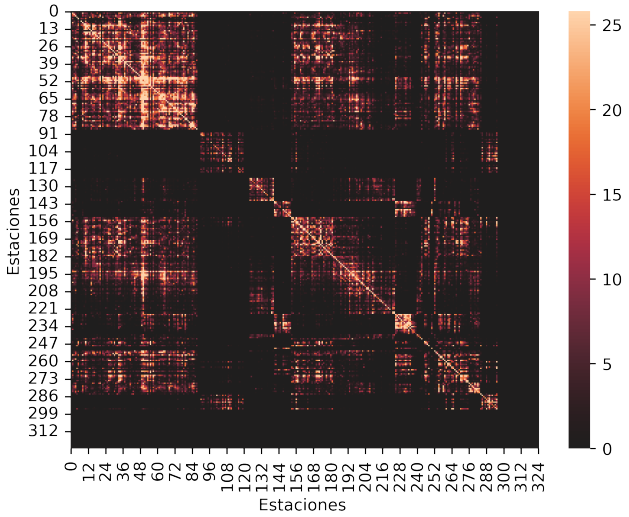


Figure 10: Etapa 4

Etapa Cuatro tomando de Noviembre de 2018 a Febrero de 2020.

Decidimos hacer un corte aquí por el evento del confinamiento dictado a inicios de Marzo de 2020.

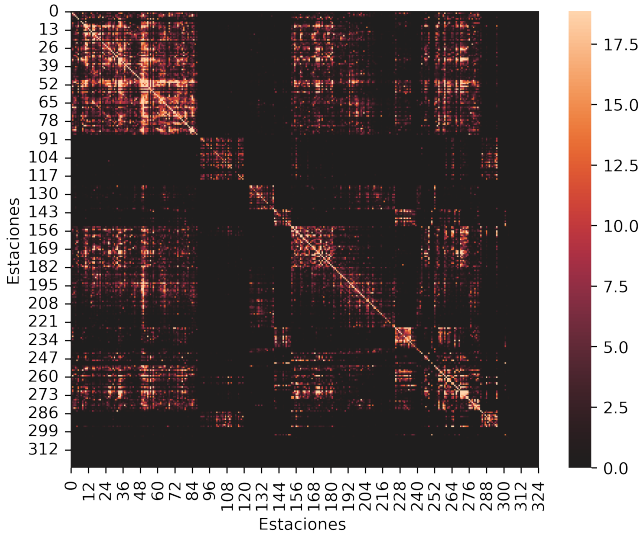


Figure 11: Etapa 5

Etapa Cinco tomando de Marzo de 2020 a Abril de 2021.

Después de la tercera ampliación, podemos ver que no hay un gran cambio entre el comportamiento de las estaciones anteriores, sin embargo si hay un nuevo aporte pues las estaciones agregadas en esta etapa tienen intereacción con las estaciones antiguas, pero casi nula con las estaciones de la segunda etapa.

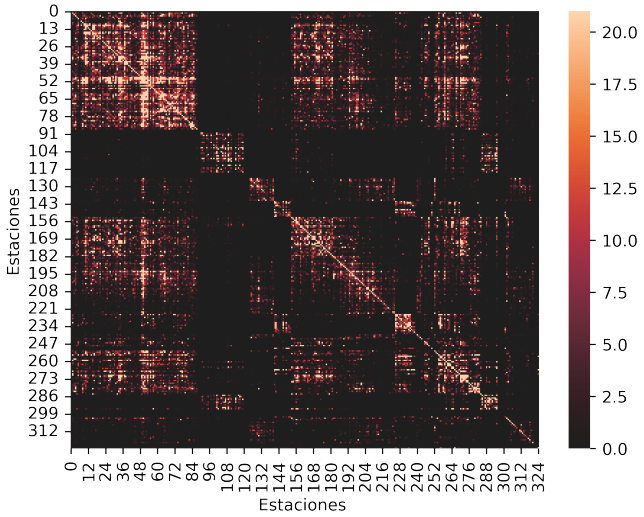


Figure 12: Etapa 6

Tomado de Mayo 2021 en adelante.

Finalente al realizarse la cuarta expansión de estaciones podemos ver que no hay un gran cambio entre el comportamiento de las estaciones anteriores, así como tampoco representa una gran interacción con las estaciones establecidas anteriormente.

Hay evidencia visual para justificar los cambios en el comportamiento de los viajes de acuerdo a los aumentos en la cantidad de estaciones, salvo por la primera ampliación, esto puede deberse tal vez a su ubicación geográfica pues el mapa de calor muestra que esas estaciones solo mantienen interacción entre ellas prácticamente.

Para la realización de estos gráficos usamos una librería de *python* llamada **seaborn**, que asigna el color de acuerdo a una escala lineal. Para evitar que los gráficos fueran muy opacos decidimos reasigar las observaciones de acuerdo al intervalo al que pertenecían tomando los cortes a partir de los quantiles observados e ignorando los ceros. De esta manera, los puntos mas iluminados serán aquellos que se encuentren por encima del 90% de las observaciones, siguiendo en orden descendente los que estén por encima de 80% y así descendentemente. Se decidió hacer de esta manera para poner más atención a la distribución y no tanto a los máximos y mínimos.

Analizando el evento Pandemia

A continuación el HeatMap del promedio de viajes promedio entre estaciones antes del anuncio del paro a las actividades **Marzo 2020**

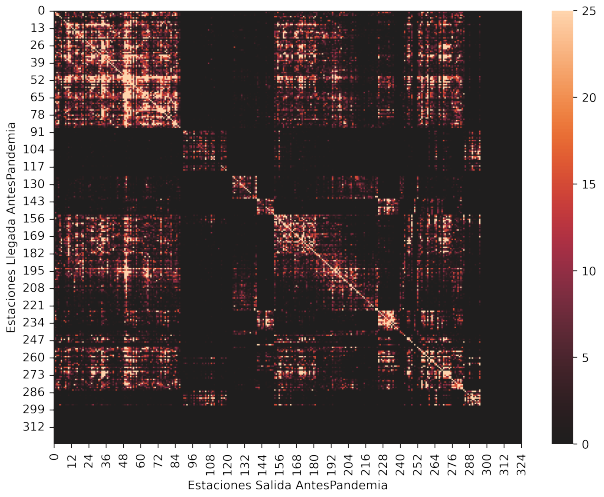


Figure 13: Antes del Marzo 2020

Y ahora despues del llamado a quedarse en casa.

Fueron calculados en la misma escala de calor para hacer ver la diferencia, se vuelve muy evidente la gran disminución en cuanto a viajes.

A continuación una visualización de como se fue dando el crecimiento de la intereacción de las estaciones.

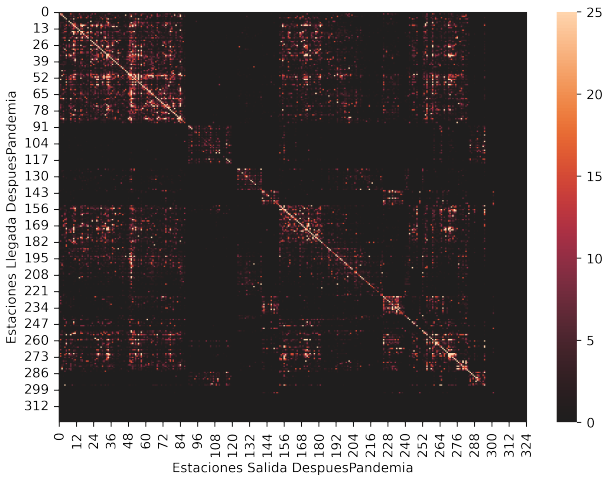


Figure 14: Mes siguiente después de Marzo 2020

Conclusiones y Comentarios Finales

De acuerdo al análisis elaborado, encontramos evidencia significativa para afirmar que el aumentos del número de estaciones de las etapas 3 y 4 tuvieron un impacto *positivo* en el programa. Se destaca en particular la etapa 3, donde observamos un aumentos en el número de viajes correspondiente a un gran aumento del número de estaciones. Los eventos de aumento de estaciones coinciden con los puntos de quiebre obtenidos en el análisis de estructura.

Tambien se observa que el impacto del aumento en el número de estaciones fue inmediato, pues en cada modelo de regresión ajustado se observa un intercepto más álto, lo que representa aumento abruto en los viajes. Esto pudo deberse a razones varias, tales como divulgación de las nuevas rutas o incorporación de algún programa que ya estaba en arranque y no necesariamente al simple hecho de abrir nuevas estaciones; no obstante el análisis de este fenómeno es algo que queda fuera de nuestro alcance pues se requiero un mayor conocimiento de las políticas aplicadas al inicio de cada etapa.

Tambien nos gustaría recalcar que aumentar el número de estaciones no garantiza un mayor uso, el mejor ejemplo de esto es que ocurrió una ampliación en el número de estaciones en 2015 pero no significó una diferencia, y esto lo sustentan las observaciones visuales así como su no detección en los tests para puntos de quiebre.

En los heatmaps se pueden observar regiones fuertemente conexas, en el sentido de que la mayoría de viajes ocurren entre estaciones de estos conjuntos denominados regiones.

El impacto de las medidas de distanciamiento social es notorio en el uso de las bicicletas del programa, puesto que hay una disminución tanto en los viajes realizados así como en la comunicación de viajes entre estaciones. Por otro lado de acuerdo al modelo de regresión lineal, se observa un aumento gradual de viajes en los últimos meses, esto es observable en la progresión [1] de los heatmap así como en otras gráficas.

[1] Se adjunta un giff animado con los heatmaps