INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO
ITAM

Caracterización de las cicloestaciones del programa Ecobici
Una implementación del Algoritmo EM

Estadística Computacional
Profesor Mauricio García Tec

Denisse Martínez 159780
Ariana López 160281
Stephane Keil 160559

18 de diciembre de 2015

CONTENIDO

I. ANTECEDENTES

ECOBICI es un sistema de bicicletas públicas de cuarta generación, que implementó el Gobierno del Distrito Federal como parte de la Estrategia de

Movilidad en Bicicleta. Desde la puesta en marcha el 16 de febrero del 2010, es gestionado por la Secretaría del Medio Ambiente del Distrito Federal.

Inició operaciones a través de 85 cicloestaciones (1,114 bicicletas) separadas por una distancia de 300 metros entre una y otra en la cual daba servicio a las colonias Cuauhtémoc, Juárez, Roma Norte, Hipódromo Condesa y Condesa. En octubre 2011 se ampliaron operaciones al Centro Histórico de la Ciudad de México con 5 cicloestaciones y 12 reubicaciones de las existentes, más Polanco y zonas aledañas. En 2015 cuenta con 444 cicloestaciones (6,000 bicicletas) con un área de cobertura de 32 km2 en 42 colonias de las Delegaciones Benito Juárez, Cuauhtémoc y Miguel Hidalgo.

ECOBICI, una alternativa de movilidad, que funciona como eficaz complemento a los sistemas de transporte y ayuda a resolver problemas de movilidad en el Distrito Federal que es una de las ciudades más grandes del mundo.

El sistema es operado por Clear Channel Outdoors a través de su división SmartBike, quien alrededor del mundo ha implementado sistemas similares en España, Francia, Noruega, Suecia e Italia.

A. PALABRAS CLAVE: Cicloestación, usuario ecobici, caracterización, EM

B. PROBLEMA A RESOLVER

Actualmente existe poca información detallada sobre el programa Ecobici que permita conocer las características de las cicloestaciones y de los usuarios que la frecuentan, por este motivo se plantea el siguiente análisis de reconocimiento con el fin de obtener mayor conocimiento acerca de estos temas.

C. OBJETIVO DEL PROYECTO Y PREGUNTAS CLAVE

Utilizando un algoritmo EM determinar los tipos de cicloestaciones existentes en el Distrito Federal durante el 2012 que puedan ser clasificadas por sus características y las de los usuarios ecobici que las frecuentan dependiendo del horario del día en que se efectúen los traslados.

¿Cómo se observa el flujo de retiros y arribos de unidades en las cicloestaciones de la ciudad?
¿Se puede clasificar las cicloestaciones con base al flujo de usuarios que presentan?
¿Existen patrones en los tiempos y distancias de traslado entre cicloestaciones?
¿Cuál es el perfil de los usuarios ecobici según las cicloestaciones que frecuentan?

II. ANÁLISIS EXPLORATORIO

El análisis fue realizado utilizando los conjuntos de datos proporcionados por el Gobierno del Distrito Federal a través del portal Ecobici (Sistema de Transporte Individual). Se obtuvieron 3 conjuntos en formato de texto completos para el año 2012 correspondientes a las llegadas-salidas por cicloestación, coordenadas geográficas y distancia entre cicloestaciones con alrededor de 3 millones de registros.

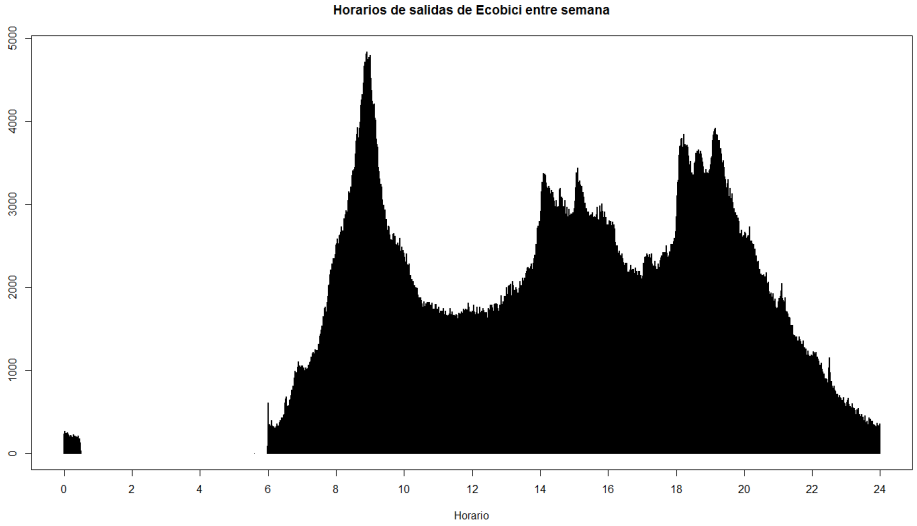
A partir del conjunto de datos original generamos 13 variables como parte del análisis exploratorio:

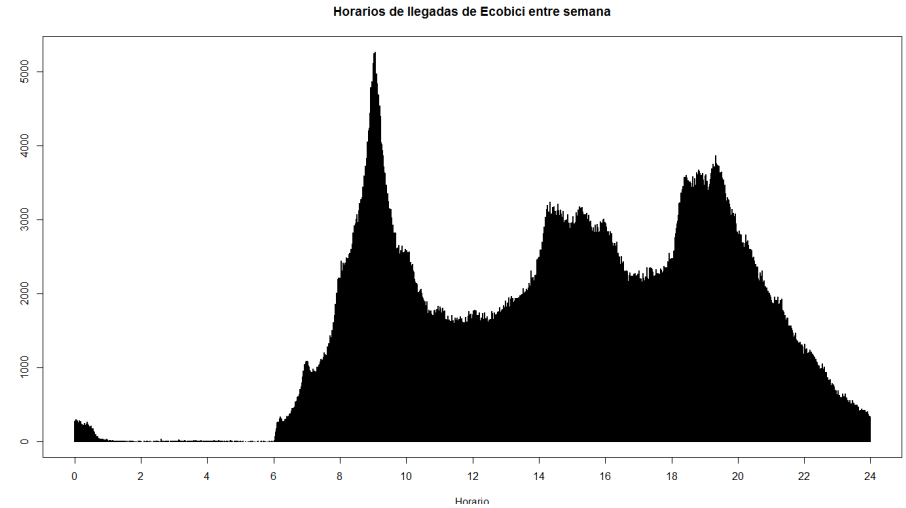
1. estación Número de estación (262 estaciones a diciembre del 2012)
2. sal_mor Frecuencia de salidas en el horario de 6:01 am a 12:00
3. lleg_mor Frecuencia de llegadas en el horario de 6:01 am a 12:00
4. sal_mid Frecuencia de salidas en el horario de 12:01 a 17:00
5. lleg_mid Frecuencia de llegadas en el horario de 12:01 a 17:00
6. sal_tar Frecuencia de salidas en el horario de 17:01 pm a 21:00 pm
7. lleg_tar Frecuencia de llegadas en el horario de 17:01 pm a 21:00 pm
8. sal_noc Frecuencia de salidas en el horario de 21:01 pm a 6:00 am
9. lleg_noc Frecuencia de llegadas en el horario de 21:01 pm a 6:00 am
10. distancia Distancia promedio entre cicloestaciones
11. dur_via Duración del viaje promedio
12. edades Edad promedio del usuario Ecobici
13. prop Proporción promedio de género de los usuarios ecobici

```
    estacion         sal_mor         lleg_mor         sal_mid         lleg_mid
 Min.   :  1.00   Min.   :   26   Min.   :   23.0   Min.   :   32   Min.   :   33.0
 1st Qu.: 65.75   1st Qu.:  337   1st Qu.:  230.8   1st Qu.:  339   1st Qu.:  327.5
 Median :132.50   Median :  683   Median :  842.0   Median :  815   Median :  745.0
 Mean   :131.69   Mean   : 2821   Mean   : 2820.9   Mean   : 2816   Mean   : 2815.7
 3rd Qu.:197.25   3rd Qu.: 5031   3rd Qu.: 4678.8   3rd Qu.: 5107   3rd Qu.: 5155.5
 Max.   :262.00   Max.   :23021   Max.   :21780.0   Max.   :17184   Max.   :18964.0
    sal_tar          lleg_tar         sal_noc          lleg_noc
 Min.   :   35.0   Min.   :   24.0   Min.   :   6.00   Min.   :   5.00
 1st Qu.:  280.5   1st Qu.:  356.8   1st Qu.:  68.75   1st Qu.:  81.75
 Median :  817.0   Median :  722.5   Median : 159.00   Median : 170.50
 Mean   : 2610.6   Mean   : 2610.6   Mean   : 673.40   Mean   : 673.40
 3rd Qu.: 4375.5   3rd Qu.: 4393.8   3rd Qu.:1174.75   3rd Qu.:1057.50
 Max.   :16778.0   Max.   :19139.0   Max.   :5843.00   Max.   :5123.00
   distancia         dur_via          edades           prop
 Min.   :0.9638   Min.   : 8.172   Min.   :31.46   Min.   :0.1397
 1st Qu.:1.6818   1st Qu.:10.142   1st Qu.:34.20   1st Qu.:0.2463
 Median :2.2551   Median :11.408   Median :34.92   Median :0.2730
 Mean   :2.3198   Mean   :12.075   Mean   :34.94   Mean   :0.2737
 3rd Qu.:2.8084   3rd Qu.:13.208   3rd Qu.:35.65   3rd Qu.:0.3032
 Max.   :5.1591   Max.   :21.675   Max.   :39.48   Max.   :0.3764
```

Como primer paso para el análisis graficamos la frecuencia de llegadas y salidas a cada cicloestacion por hora del día, separando días entre semana y fines de semana. Esto con el fin de dividir nuestro estudio en períodos que mostraran un comportamiento semejante.

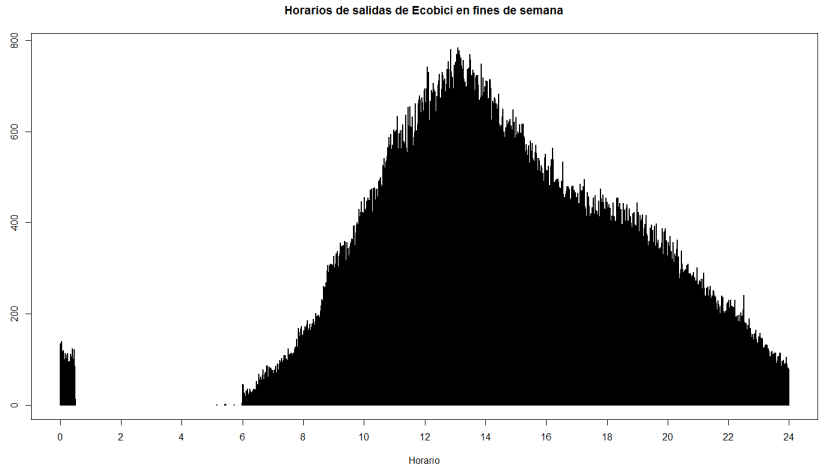
Observamos en los días entre semana picos de frecuencias en los horarios de las 9:00, 15:00, 19:00 y prácticamente nula frecuencia a partir de medianoche, por lo que agrupamos nuestros horarios de análisis en 6:01 a 12:00, 12:01 a 17:00, 17:01 a 21:00 pm y 21:01 a 6:00. En este punto no encontramos diferencias relevantes en los datos observados para las salidas o llegadas a las estaciones, se comportan de manera similar.

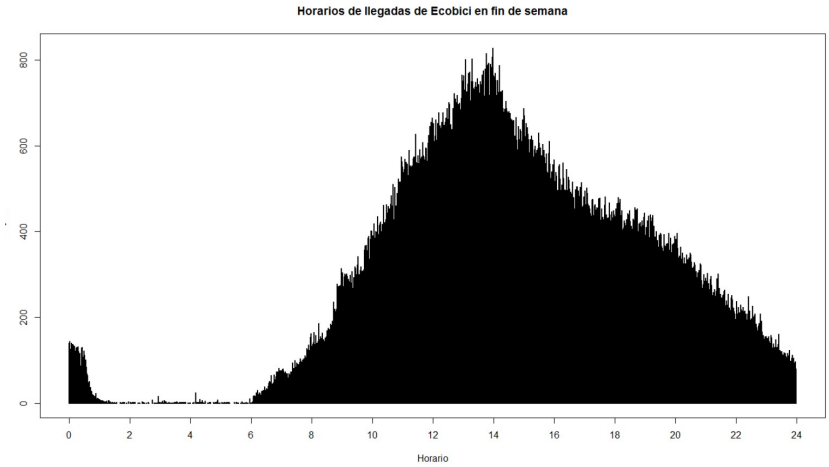




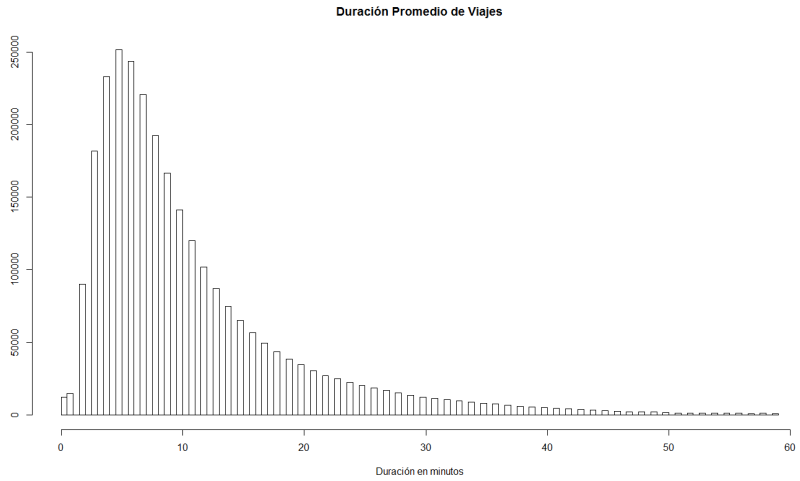
Las frecuencias observadas en fines de semana difirieron considerablemente con respecto a los días entre semana. Los datos se agrupan similares a una normal, concentrando la mayoría de los viajes en el horario entre 12:00 a 16:00 pm.

Igualmente, no encontramos diferencias significativas en los datos observados para las salidas o llegadas a las estaciones, se comportan de manera similar.

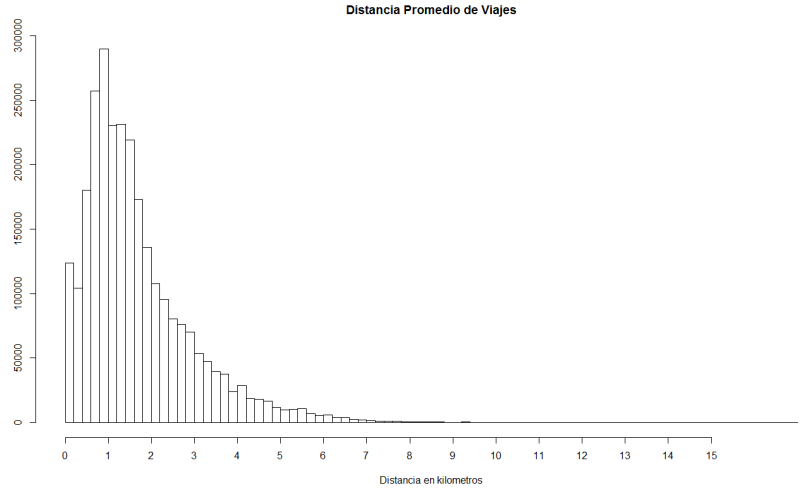




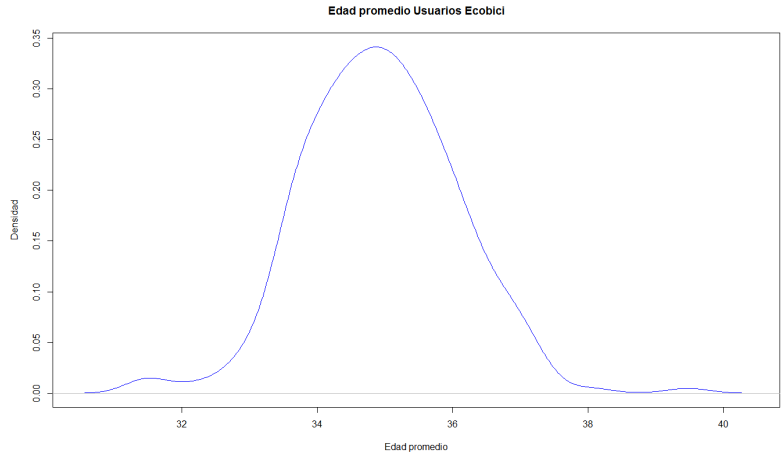
La duración máxima de los viajes entre estaciones es de 60 minutos, sin embargo, en promedio no dura más de 12 minutos.



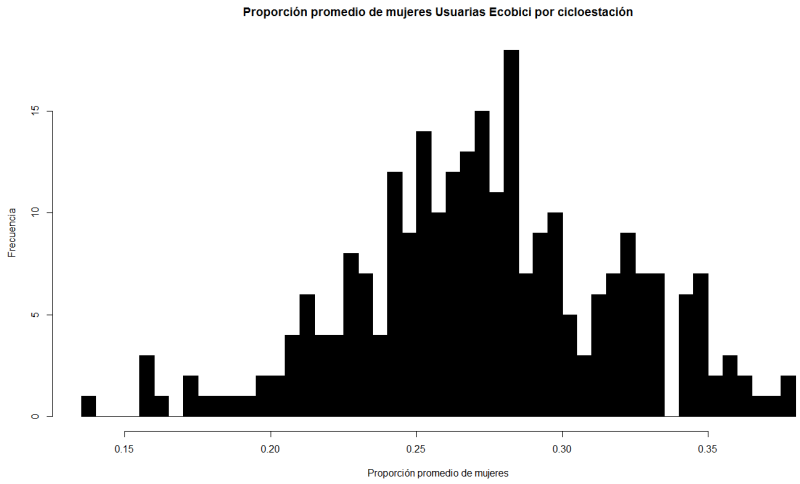
La distancia máxima de los viajes entre estaciones es de 8 kilómetros, sin embargo, la distancia promedio recorrida es de 1.7 kilómetros.



Respecto a las características de los usuarios ecobici encontramos que aproximadamente 90% de los usuarios ecobici se encuentran en un rango de edad de entre 33 y 37 años. El 33% de los usuarios tiene 35 años.

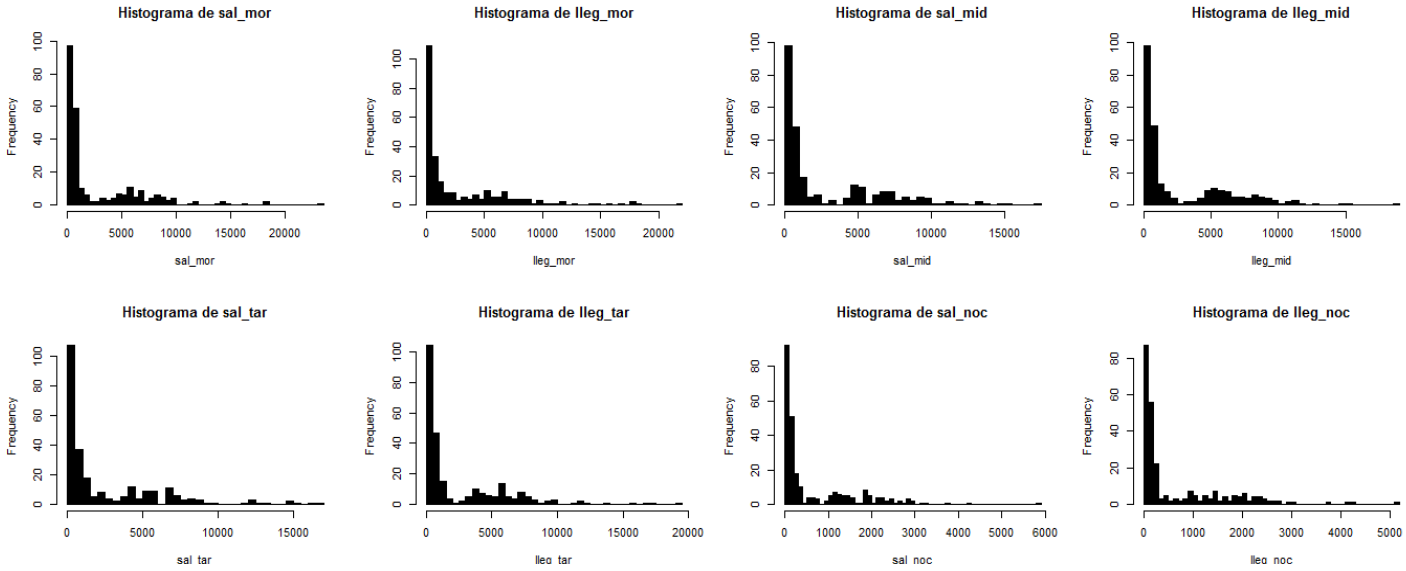


Los usuarios ecobici en su mayoría son hombres, sólo entre un 25% y 30% de los usuarios totales son mujeres.



En base al análisis exploratorio de los datos decidimos construir variables a nivel cicloestación que nos permitieran realizar una clasificación en función del uso típico que se le da a cada ubicación. Partimos de la hipótesis que ciertas estaciones dan origen a los desplazamiento de las personas hacia sus centros de trabajo (salidas en la mañana y llegadas por la tarde) mientras que otras estaciones son el destino de las personas para llegar a su trabajo (llegadas en la mañana y salidas por la tarde). Es probable también que aquellas personas que viven y se desplazan de manera local en las áreas cubiertas por el sistema de ecobici, utilizan este sistema como su principal método de transporte. Para poder analizar estos comportamientos es necesario obtener los valores promedio de viajes de salida y de viajes de llegada a cada una de las cicloestaciones durante la semana laboral. Se transformó la base de datos de viajes para generar valores agregados a nivel cicloestacion para realizar la caracterización de éstas en diferentes grupos significativos.

A continuación se presentan los histogramas del volumen de viajes de salida y de llegada promedio de cada una de las cicloestaciones en los cuatro horarios definidos (Mañana, Mediodía, Tarde y Noche). Podemos observar que existe un importante sesgo a la derecha en la distribución de frecuencia de viajes para todos los horarios, es decir que un grupo pequeño de cicloestaciones concentra la gran mayoría de viajes realizados mientras que la gran mayoría de las estaciones presenta una pequeña proporción de los viajes totales.



III. METODOLOGÍA

- Análisis exploratorio de datos disponibles y alcance del proyecto

Se exploraron las bases de datos disponibles para identificar variables relevantes que contribuyen al objetivo del proyecto. Además, se definió el período de análisis y se delimitó el alcance de éste.

- Limpieza, creación y selección de atributos

Una vez identificadas las variables se preprocesan los datos, limpiando e integrando la información. Posteriormente se calculan atributos que pueden ser útiles como la distancia entre una estación y otra, duración del trayecto, proporción de mujeres usuarias de ecobici, entre otras.

- Análisis de atributos

Con el fin de entender la dinámica de nuestras variables y reducir la dimensionalidad, se efectúa un Análisis de Componentes Principales.

- Definición del algoritmo de agrupamiento

A través del “criterio del codo” de Bezdek se elige el número de clases óptimas que representan de la mejor forma nuestros datos. El método elegido para clasificar las diferentes clases de estación de ecobici fue un Algoritmo EM.

- Caracterización de grupos

Una vez identificados los parámetros (centroides) de las distribuciones resultantes de la estimación del Algoritmo EM, se etiqueta cada observación de acuerdo la distancia euclidiana entre los centroides y las observaciones.

- Visualización e interpretación de resultados

Para facilitar la interpretación de las clases de estaciones se imprime en un mapa su ubicación coloreando según su clase. Se analizan las métricas de cada grupo y se busca entender la dinámica que presenta cada uno.

IV. ANÁLISIS DE COMPONENTES PRINCIPALES

Se realizó un Análisis de Componentes Principales (PCA por sus siglas en inglés) para entender cómo es que diferentes variables explican la dinámica de la información. Además de reducir la dimensionalidad y la redundancia en los datos.

Supuestos:

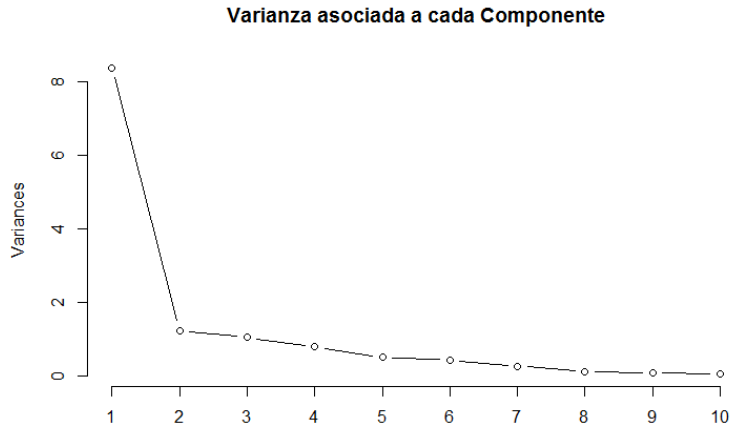
- Sólo encuentra relaciones lineales.

- Asume que la covarianza es un buen indicador de las relaciones interesantes entre atributos. Por lo cual si los datos no pueden ser adecuadamente descritos a través de su media y su varianza el método es inadecuado (si las observaciones no se distribuyen de forma gaussiana o exponencial).

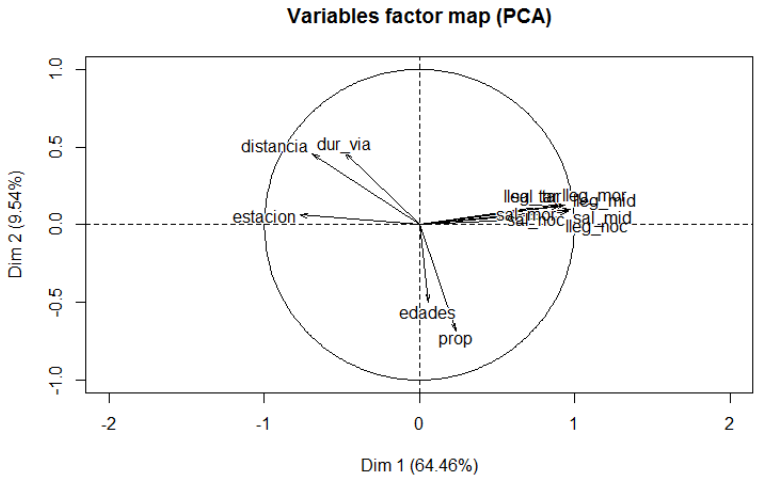
- Componentes ortogonales.

En general, el método PCA consiste en encontrar una matriz P de tal forma que la matriz de covarianza PX es diagonal y las entradas de esa diagonal tengan orden descendente. PX representará las variables transformadas que son combinaciones lineales de los atributos originales y estarán conformadas por los pesos calculados en P. Dado que la matriz de covarianza es diagonal, la covarianza de cualquier par de variables distintas será cero, por lo que las varianzas de las nuevas variables estarán sobre la diagonal.

Los resultados que arrojó el PCA para los 13 atributos, indican que el 82.1% de la varianza lo explican los primeros tres componentes.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Desviación Estándar	2.895	1.113	1.027	0.890	0.712	0.660	0.520	0.343	0.294	0.247	0.182	0.134	0.065
Proporción de la varianza	0.645	0.095	0.081	0.061	0.039	0.034	0.021	0.009	0.007	0.005	0.003	0.001	0.000
Proporción Acumulada	0.645	0.740	0.821	0.882	0.921	0.955	0.975	0.984	0.991	0.996	0.998	1.000	1.000



En el gráfico anterior se muestran las dos dimensiones (PC1 y PC2) que acumulan el 74% de la varianza y cómo es que se relacionan las variables. Para el análisis correspondiente se consideró que la representatividad del 82% de la varianza en 3 componentes es suficiente para explicar el modelo.

Tabla de Correlación y Composición de PC1, PC2 y PC3

	PC1		PC2		PC3	
	Correlación	Composición	Correlación	Composición	Correlación	Composición
estacion	0.770904062	0.26629734	0.062562528	0.05619089	0.310552974	0.302471562
sal_mor	-0.902406617	-0.31172294	0.120652886	0.10836508	-0.033805195	-0.032925494
lleg_mor	-0.90191118	-0.3115518	0.133872849	0.12023866	-0.015000862	-0.0146105
sal_mid	-0.97214606	-0.33581339	0.099368607	0.08924848	-0.03523409	-0.034317206
lleg_mid	-0.972655412	-0.33598934	0.100109342	0.08991377	-0.048937301	-0.047663822
sal_tar	-0.938239177	-0.32410076	0.124264841	0.11160917	-0.002207274	-0.002149835
lleg_tar	-0.921402571	-0.31828481	0.12909324	0.11594583	-0.048878768	-0.047606813
sal_noc	-0.958889306	-0.33123404	0.082948562	0.07450072	0.043621968	0.042486809
lleg_noc	-0.923601857	-0.31904452	0.041109128	0.0369224	0.0250002	0.024349628
distancia	0.689854758	0.23830006	0.458525208	0.41182703	-0.165816376	-0.161501394
dur_via	0.480644364	0.16603144	0.461355622	0.41436918	-0.523980964	-0.510345589
edades	-0.058696813	-0.02027594	-0.498677987	-0.44789047	-0.794737105	-0.774055937
prop	-0.234721922	-0.0810812	-0.686223615	-0.61633565	0.118837491	0.115745024

La tabla anterior muestra la correlación de los PC’s y las variables originales y cómo está compuesto cada componente. Entre las relaciones más destacadas se encuentran:

- PC1 tiene una correlación alta negativa con las variables que representan el nivel de tráfico de viajes promedio.

- PC2 tiene una correlación media positiva con la distancia y la duración de los viajes promedios y una relación negativa con la edad promedio del usuario y la proporción de mujeres entre usuarios totales.

- PC3 tiene una correlación alta negativa con la duración de los viajes y la edad promedio de los usuarios.

V. ALGORITMO DE AGRUPAMIENTO

La tarea de agrupamiento busca encontrar qué elementos del objeto de estudio se parecen entre sí y al mismo tiempo asegurar que los grupos encontrados sean diferentes entre sí. Existen diferentes estrategias para encontrar estos grupos, agrupamiento por particiones, por densidades o de manera jerárquica. Dentro de los algoritmos de agrupamiento por partición existen aquellos que asignan de manera única a cada objeto de estudio a un grupo y aquellos que se basan en lógica difusa dónde cada objeto de estudio tiene una función de membresía a cada uno de los grupos del sistema.

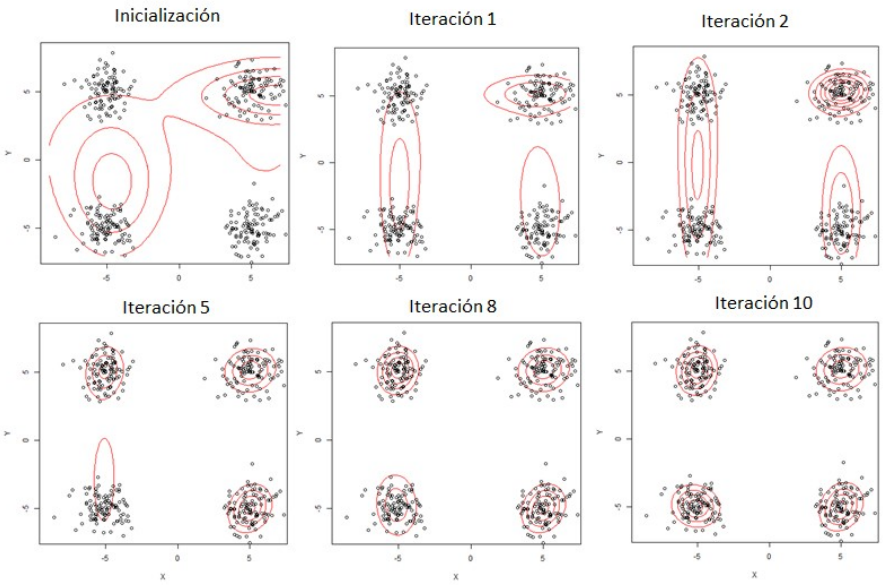
Decidimos realizar una implementación del algoritmo EM para el agrupamiento difuso de las cicloestaciones del sistema ECOBICI para ejemplificar el uso de esta técnica estadística para encontrar variables latentes no observadas, en este caso los parámetros de una mezcla de distribuciones gaussianas de probabilidad que ayuden a determinar los diferentes grupos presentes en la muestra. Asumimos que los datos de las cicloestaciones pueden caracterizarse por una mezcla de distribuciones normales multivariadas y que es necesario encontrar las proporciones de cada una de las distribuciones usadas en la mezcla así como los parámetros que describen a cada una de ellas.

El Algoritmo de Expectation-Maximization (EM) es un método para encontrar estimadores de máxima verosimilitud para parámetros de distribuciones de probabilidad que dependen de variables no observables. Este algoritmo resulta interesante cuando se desconoce la distribución de la cual proviene cada observación y se desea hacer una clasificación de estas. Este método consta de los pasos: Expectation y Maximization.

El proceso consiste en definir una esperanza o expectativa en particular y posteriormente maximizarla. El procedimiento inicializa con parámetros aleatorios de las distribuciones los cuales son utilizados para calcular las probabilidades de que cada observación pertenezca a un grupo. Estas probabilidades o valores esperados se actualizan en cada iteración maximizando la expectativa en esa iteración particular. La maximización repetida converge al máximo de la propia función de verosimilitud. El algoritmo en su fase de Expectation evalúa la probabilidad de pertenencia de cada observación de la muestra a cada una de las distribuciones normales multivariadas de la mezcla de distribuciones. Una vez que se determinan las distribuciones condicionales de cada observación dados los parámetros encontrados en la iteración anterior, se procede a l paso de Maximization, dónde se obtienen los estimadores de máxima verosimilitud de las variables latentes, en este caso los parámetros y proporciones de cada una de las distribuciones que componen nuestra mezcla.

Realizamos la implementación del algoritmo EM en el software R en su versión 3.2.2 utilizando el paquete base y el paquete mvtnorm, éste último para calcular las distribuciones de probabilidad normales multivariadas. Uno de los principales problemas del algoritmo EM es la necesidad de generar parámetros de inicialización que afectan el resultado del algoritmo.

miento del algoritmo implementado para una muestra de datos generada sintéticamente dónde se tienen claramente cuatro agrupaciones principales. Se presentan las gráficas de contorno de las distribuciones normales multivariadas que componen la mezcla en este ejemplo de dos dimensiones. Nótese que el algoritmo converge rápidamente en tan solo 10 iteraciones.

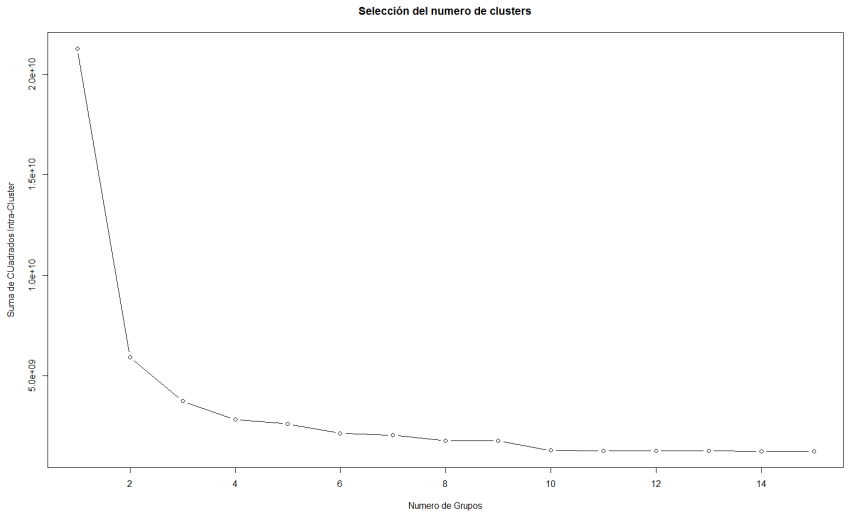


Uno de los problemas principales de este método es la decisión de los parámetros de inicialización para los parámetros iniciales de las distribuciones que componen la mezcla. Decidimos realizar una inicialización aleatoria de las distribuciones de la mezcla y las proporciones de mezcla fueron dejadas equiprobables entre todas las distribuciones. El vector de medias de cada distribución se elige de manera aleatoria y uniforme en el rango minimo y maximo de cada uno de los atributos o dimensiones del análisis, para la matriz de varianza y covarianza se inicializan todas las covarianzas en el valor de 0 y el valor de la varianza se elige de manera aleatoria entre 0 y la varianza de la muestra para cada atributo. A pesar de todos los esfuerzos existe siempre la probabilidad de una mala elección de alguna de las distribuciones que causa que ninguna observación tenga probabilidad alguna de pertenecer por lo que en el paso de maximización en ocasiones el cálculo de la nueva matriz de covarianzas falla debido a que todos los valores para una distribución valen 0. Este problema ocurre con mayor frecuencia cuando el número de dimensiones se incrementa, cayendo en la maldición de la dimensionalidad. Logramos reducir este problema utilizando una técnica de reducción de la dimensionalidad y de selección de atributos que describimos anteriormente en la sección de Componentes Principales.

Otro elemento de importancia es la decisión de cuentas distribuciones deben componer la mezcla, en el caso del uso del algoritmo EM para agrupamiento cada distribución adicional que se introduce para componer la mezcla de distribuciones corresponde a un grupo adicional en el que se particionan las observaciones de la muestra. No existe un criterio determinístico para saber cual es el número óptimo de agrupamientos que se requieren para caracterizar un conjunto de datos. Sin embargo existen ciertos criterios heurísticos que permiten tener una idea aproximada, además del conocimiento experto que siempre se requiere en un análisis no supervisado. Uno de estos criterios tiene que ver con minimizar la varianza entre las observaciones de cada agrupación y otro en

maximizar las diferencias entre los diferentes grupos, es decir maximizar la medida de distancia utilizada entre los grupos. Aunque existen varios criterios para medir la calidad de un agrupamiento, decidimos utilizar el criterio del codo, donde se analiza la varianza intra-cluster en función del número de grupos que se buscan.

A continuación se presentan los resultados de esta selección para el problema de agrupamiento de las cicloestaciones, es notable como existen varios puntos donde la reducción marginal de la varianza se vuelve pequeña, en este problema en particular notamos estos puntos de inflexión en 4 y 10 grupos. Decidimos utilizar 4 grupos basados en nuestras hipótesis y conocimiento previo del tema, aunque 10 grupos probablemente darían mayor precisión en cuanto a las características de las cicloestaciones del sistema ECOBICI.



Para efectos de nuestro análisis se corrió el algoritmo que desarrollamos con cuatro distribuciones normales multivariadas con valores iniciales aleatorios con las características que mencionamos anteriormente. Es importante notar que no se utilizaron los datos en su formato original sino que se aplicó el Análisis de Componente Principales y se incluyeron aquellas que explican el 82% de la varianza en los datos, únicamente se incluyeron las tres componentes principales, lo cual permitió reducir la dimensionalidad del problema de 13 dimensiones de análisis a tan solo 3 dimensiones de análisis. A pesar de estos esfuerzos el algoritmo EM tuvo dificultades en converger adecuadamente debido a malos parámetros de inicialización, estas fallas del algoritmo se incrementan con el número de grupos que se buscan. Por ejemplo, para optimizar la función de verosimilitud completa con cuatro distribuciones únicamente 1 de cada dos intentos fue exitoso.

VI. RESULTADOS

En base a nuestro análisis las cicloestaciones pueden ser clasificadas en cuatro grupos principales de acuerdo a la información disponible. Una vez clasificadas las cicloestaciones de acuerdo a los resultados del algoritmo de agrupación, nos dimos a la tarea de caracterizar cada uno de los cuatro grupos. A continuación se presentan las características principales:

			Clusters			
Atributo	**Estadistico**	**Total**	**1**	**2**	**3**	**4**
Numero de Estación	Media	132	178	49	56	32
	Desviación	76	49	24	31	23
Salidas Mañana	Media	2,821	523	6,600	7,050	8,251
	Desviación	3,902	454	1,753	6,919	4,085
Llegadas Mañana	Media	2,821	554	6,278	5,379	10,125
	Desviación	3,939	589	2,159	5,343	4,746
Salidas Mediodia	Media	2,816	562	6,615	6,222	8,534
	Desviación	3,599	438	1,892	4,886	3,199
Llegadas Mediodia	Media	2,816	553	6,749	6,967	7,682
	Desviación	3,549	429	1,814	4,927	2,833
Salidas Tarde	Media	2,611	556	5,896	5,202	8,659
	Desviación	3,470	526	1,702	4,851	4,003
Llegadas Tarde	Media	2,611	520	6,059	7,397	6,758
	Desviación	3,481	431	1,607	5,702	3,712
Salidas Noche	Media	673	116	1,680	1,300	2,111
	Desviación	934	92	588	1,136	1,168
Llegadas Noche	Media	673	119	1,803	1,422	1,700
	Desviación	922	89	570	1,360	1,000
Distancia viaje promedio	Media	2.3	2.7	1.3	2.5	1.7
	Desviación	0.8	0.7	0.2	0.5	0.2
Dureción viaje promedio	Media	12.1	12.6	9.6	15.9	10.3
	Desviación	2.7	2.3	0.9	3.4	1.1
Edad promedio de los usuarios	Media	34.9	34.8	35.2	35.6	34.7
	Desviación	1.1	1.2	0.9	1.0	0.8
Proporción de Mujeres usuarias	Media	27.4%	26.9%	30.6%	24.9%	25.7%
	Desviación	4.5%	4.6%	3.6%	3.0%	3.1%
Observaciones		**260**	**169**	**51**	**18**	**22**

*

El primer grupo se caracteriza principalmente por un bajo número de viajes de salida y de llegada en todos los horarios estudiados, estas cicloestaciones tienden a ser las de última instalación y muestran un volumen de viajes que alcanza apenas el 10% de viajes de cualquiera de los otros tres grupos, este grupo tiene 169 estaciones, es decir más de dos tercios del total de estaciones analizadas. Este grupo es el que muestra la mayor distancia promedio en kilómetros de sus viajes, esto nos indica que son las estaciones más alejadas de los centros de gravedad principales del sistema de cicloestaciones.

El segundo grupo muestra un volumen de viajes por mucho superior en todos sus horarios, es notable sin embargo que el volumen de viajes de salida y de llegada es balanceado en los cuatro rangos horarios analizados y una varianza baja en términos de los viajes entre las cicloestaciones que lo componen, esta cicloestaciones se caracterizan por una alta proporción de viajes realizados por

usuarios del sexo femenino comparado con los demás grupos y tienen los valores más bajos en términos de las distancias y duraciones de los viajes generados.

El tercer y cuarto grupo muestran valores superiores a los del segundo grupo en términos de viajes totales generados comparados con el segundo grupo. Sin embargo ambos grupos tienen comportamientos diferentes en términos de las proporciones entre viajes de salida y viajes de llegada, lo cual nos indica un comportamiento de conmutación de los habitantes de la ciudad desde sus viviendas y hacia sus lugares de trabajo por las mañanas y al contrario por las tardes. El tercer grupo de cicloestaciones tiene un mayor volumen de viajes de salida por la mañana respecto a los viajes de llegada y este comportamiento se invierte por las tardes donde los viajes de llegada superan a los viajes de salida, esto nos indica que estas cicloestaciones son utilizadas por los usuarios del sistema para desplazarse hacia sus lugares de empleo posiblemente conmutando desde otros medios de transporte público del sistema de nodos de transporte de la ciudad. Estas estaciones tienen un promedio de distancia y de duración de los viajes superior a la media y en general a cualquiera de los otros tres grupos además de mostrar la menor proporción de viajes realizados por usuarios del sexo femenino y corresponden a viajes realizados por personas de mayor edad.

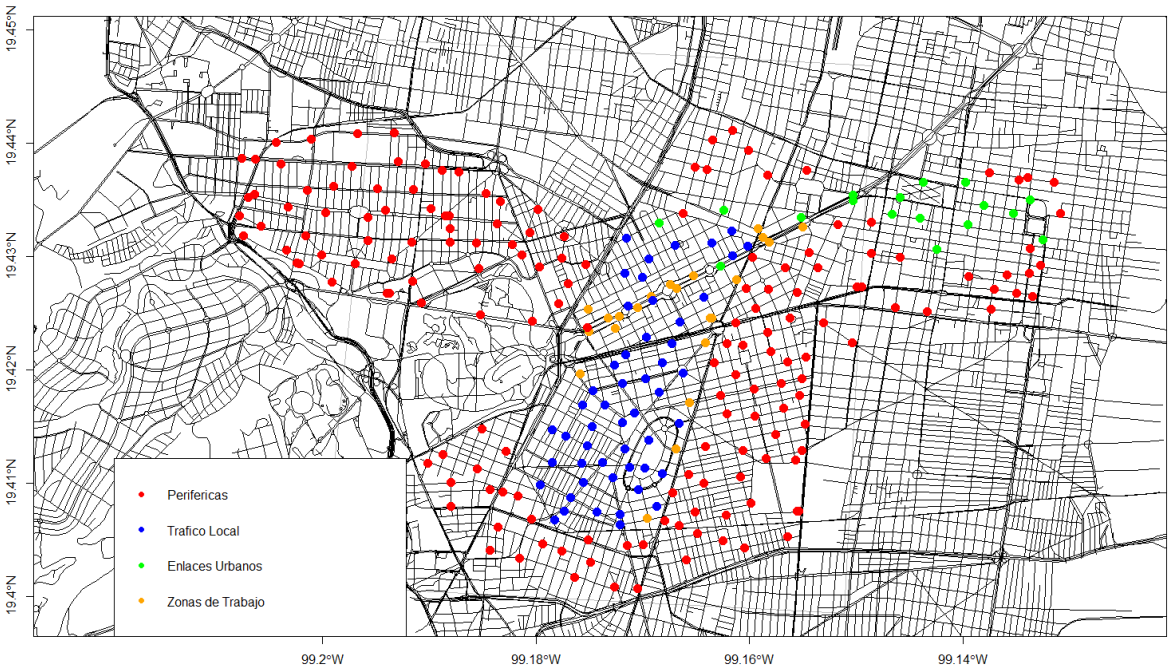
El cuarto grupo es el que muestra el mayor número de viajes totales de todos los grupos y tiene la característica principal de mostrar un mayor número de viajes de llegada por la mañana que de salida y inversamente un mayor número de salidas por la tarde que de llegadas. Esto nos indica que estas cicloestaciones se encuentran cerca de los lugares de trabajo de los usuarios del sistema y son las cicloestaciones que se comportan como los mayores centros de gravedad en términos de los viajes. El volumen de viajes por la noche en estas estaciones es superior al de cualquiera de los otros grupos lo cual nos permite suponer que se encuentra en centros nodales donde existe vida nocturna.

Para poder revisar de manera visual estas características, decidimos codificar cada uno de estos grupos por color de la siguiente manera.

Grupo	Color
1 - Pocos viajes, distancias y duraciones grandes. Estaciones más recientes.	Rojo
2- Comportamiento balanceado de salidas y llegadas, alta proporción de viajes de mujeres.	Azul
3- Flujo hacia fuentes de empleo, distancias largas, baja proporción de	Verde

viajes realizados por mujeres.	
4- Flujo destino de viajes por la mañana, centros de gravedad y alto trafico a todas horas	Anaranjado

Utilizamos las ubicaciones geográficas de las cicloestaciones y el grupo al que pertenecen de acuerdo a nuestro análisis para incorporar el contexto geográfico a nuestra clasificación. A continuación se presenta un mapa con las cicloestaciones coloreadas por el grupo al cual pertenecen.



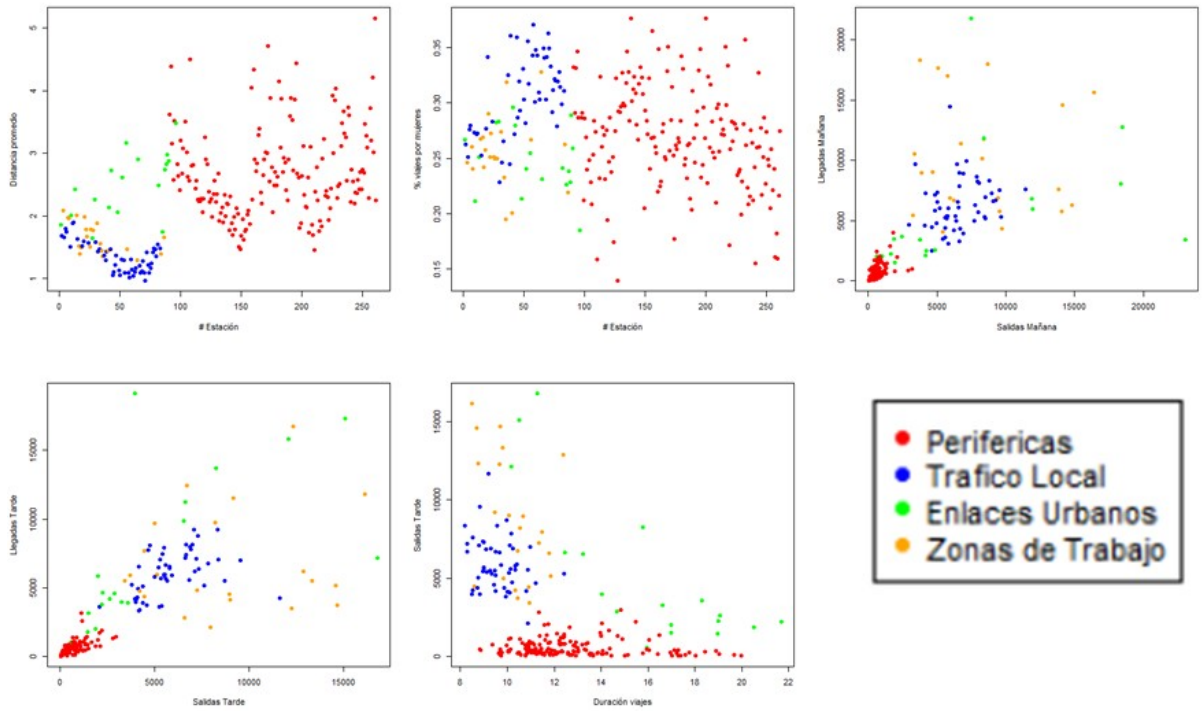
Es notable como las estaciones anaranjadas se encuentran casi todas cerca de la avenida Reforma y varios puntos de la Avenida de los Insurgentes, esta cicloestaciones reciben un mayor número de llegadas por la mañana y un mayor número de salidas por la tarde, esto nos confirma que son estaciones principales cercanas a centros importantes de empleo de la población de la ciudad. Decidimos llamarlas estaciones en zonas de trabajo.

Las estaciones del grupo 2, que muestran viajes de salida y llegada balanceados a los largo del día se concentran en las colonias condesa, cuauhtémoc y juárez, zonas de vivienda donde se concentran jóvenes de nivel socioeconómico medio alto y un gran número de comercios, restaurantes y bares. Asimismo esta cicloestaciones son las que tienen los viajes más cortos y la mayor proporción de viajes realizados por mujeres. De acuerdo a la ubicación y comportamiento de estas estaciones decidimos llamarlas estaciones de tráfico local dado que los usuarios de estas estaciones las utilizan como su medio de transporte principal.

Las estaciones del grupo 3 se concentran en el primer cuadrante de la ciudad, y en puntos nodales del sistema colectivo de transporte como el metro o el metrobus, estas estaciones muestran un comportamiento de conmutación con altas salidas por la mañana y altas llegadas por la tarde asi como la distancia y tiempos promedio más grandes de los grupos. Estas estaciones parecen ser principalmente utilizadas por personas que se desplazan desde otras zonas de la ciudad y que incorporan el sistema ECOBICI como parte de su trayecto diario para llegar al trabajo. Las llamamos estaciones de enlace urbano ya que son aquellas que permiten que los usuarios transbordan a otros elementos del sistema de transporte colectivo de la ciudad.

FInalmente las estaciones rojas, se encuentran casi en su totalidad en la periferia del sistema y muestran un uso muy limitado en términos de viajes, los viajes tienen una distancia promedio grande por lo que estas estaciones parecen pertenecer a la periferia del sistema. Sin embargo también son las estaciones más nuevas por lo que dado que nuestro análisis se restringe al año 2012 es posible que aún no fueran incorporadas por los habitantes de estas zonas como medio de transporte regular. Este grupo concentra más de ⅔ de las estaciones por lo que existe una gran diversidad en sus ubicaciones y uso por lo que simplemente las llamamos periféricas a falta de mayor información al respecto.

A continuación se presentan gráficas de dispersión para algunas de las variables analizadas.



Se presenta a continuación un esquema que resume las características de cada uno de los grupo**s analizados.**

VII. REFERENCIAS

Kuri-Morales, A. 2014. “Data Base Analysis using a Compact Data Set”. International Congress on Big Data. México, D.F.

Kovács-Ferenc & Iváncsy-Renáta. 2006. “A Novel Cluster Validity Index: Variance of the Nearest Neighbor Distance.” *Wseas Transactions on Computer*: 477-483. Vol. 5 March 2006, ISSN: 1109: 2750.

“Algoritmos EM”. Econometric Laboratoty. University of Berkeley <<http://eml.berkeley.edu/books/choice2nd/C14.pdf>>