VNIVERSIDAD NACIONAL
AVTNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

TRANSPORTE PÚBLICO: ¿PREDECIR O ADAPTAR?

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

PRESENTA:
JAIR CASTRUITA GASTÉLUM

DIRECTOR DE TESIS:
Dr. CARLOS GERSHENSON GARCÍA, GUILLERMO SANTAMARÍA BONFIL
Facultad de ciencias, UNAM

Ciudad Universitaria, CD. MX. Enero, 2017

A mi familia, por todo el esfuerzo, apoyo y amor incondicional.
A mis amigos, por enseñarme algo nuevo cada día.
A mis tutores, por alimentar mi deseo de aprender.
A la Universidad, por la formación que me han dado.
Es gracias a ustedes que es posible el presente trabajo.
En verdad, gracias.
Yo.

Reconocimientos

También quisiera reconocer a CONACYT, PAPIIT / etc. Por el apoyo y confianza que ha depositado en mí y en el de muchos compañeros para culminar nuestros estudios. Porque cada día hay más preguntas y se necesitan más personas para contestarlas.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Jair Castruita Gastélum. Ciudad de México, 2017

Resumen

Este trabajo realiza un estudio sobre el sistema de bicicletas compartidas de Ecobici en la CDMX, para la obtención de conocimiento y capturar las tendencias que se han manifestado a partir de las impresiones digitales considerando los primeros tres periodos en el funcionamiento del sistema. Dado que las estaciones de Ecobici se encuentran en constante interacción, éstas pueden manifestar comportamientos diferentes con respecto a su ubicación y periodo analizado. Convirtiendo el desarrollo de herramientas para su gestión, pronósticos en la oferta y demanda difícilmente posible. Más aún el comportamiento de algunas estaciones dentro del sistema puede ser pronosticado con un menor grado de error por el rol que juega en la ciudad. Se propone el uso de dos técnicas: la agrupación de estaciones basados en los conteos de los viajes utilizando una técnica de clustering conocida como mixture models y una herramienta para segmentar regiones de predictibilidad con el uso de diversidad de rango, que es una técnica para la medición de variabilidad de los rangos dependiente del tiempo utilizada en sistemas complejos. Por último se realiza una categorización de estaciones con las regiones obtenidas por la diversidad de rango para medir y comparar los pronósticos realizados.

Índice general

1. Introducción **1**
- 1.1. Planteamiento del problema . . . 1
- 1.2. Objetivo general . . . 3
  - 1.2.1. Objetivos específicos . . . 3
- 1.3. Motivación . . . 3
- 1.4. Contribuciones . . . 4
- 1.5. Estructura de la tesis . . . 5

2. Estado del arte **7**

3. Datos **11**
- 3.1. Bike sharing systems . . . 11
- 3.2. Ecobici . . . 12
- 3.3. Fuente de datos . . . 12
- 3.4. Limpieza, pre-procesamiento y análisis de los datos . . . 13
- 3.5. Remoción de datos atípicos . . . 13
- 3.6. Separación de periodos: fase 1, 2 y 3 . . . 17
- 3.7. Obtención de perfiles para las estaciones de Ecobici utilizando clustering 19
  - 3.7.1. Comportamiento promedio para cada estación . . . 20
  - 3.7.2. Agrupación y medidas de evaluación – fase 1 . . . 21
  - 3.7.3. Agrupación y medidas de evaluación – fase 2 . . . 23
  - 3.7.4. Agrupación y medidas de evaluación – fase 3 . . . 26
  - 3.7.5. Reducción de dimensionalidad aplicando Análisis en Componentes Principales (PCA) . . . 27
  - 3.7.6. Agrupación y medidas de evaluación para la fase 1 – PCA . . . 30
  - 3.7.7. Agrupación y medidas de evaluación para la fase 2 – PCA . . . 32
  - 3.7.8. Agrupación y medidas de evaluación para la fase 3 – PCA . . . 33
- 3.8. Resumen . . . 34

4. Diversidad de rango (rank diversity) **37**
- 4.1. Construcción de diversidad de rango para Ecobici . . . 38
- 4.2. Frecencia de rango . . . 39
- 4.3. Resumen . . . 42

5. Diseño experimental **43**
5.1. Gradient Boosted Regression Trees . . . 43
5.2. Conjunto de prueba y entrenamiento . . . 43
5.3. Definición de las regiones de cabeza, cuerpo y cola . . . 44
5.4. Ajuste y predicción para GBRT . . . 45
5.4.1. Evaluación y prueba de hipótesis con RMSE . . . 47
5.4.2. Evaluación y prueba de hipótesis con NDCG . . . 54
5.5. Resumen . . . 58

6. Conclusiones y discusión **61**
6.1. Resumen . . . 61
6.2. Trabajo futuro . . . 62

A. Código/Manuales/Publicaciones **65**
A.1. Algoritmos complementarios . . . 65
A.1.1. Bayesian Information Criteria . . . 65
A.1.2. Reducción de dimensionalidad aplicando PCA . . . 67
A.1.2.1. Estandarización . . . 67
A.1.2.2. Ordenamiento y selección de los eigenpares . . . 67
A.1.2.3. Varianza explicada . . . 68
A.1.2.4. Matriz de proyección . . . 68
A.1.2.5. Proyección en el nuevo espacio de características . . . 68
A.2. Técnicas en complejidad . . . 69
A.2.1. Definición de las regiones de la cabeza, cuerpo y cola utilizando diversidad de rango . . . 69
A.3. Medidas de errores para evaluación . . . 70
A.3.1. Normalized Discounted Cumulative Gain (NDCG) . . . 70
A.4. Pruebas estadísticas . . . 70
A.4.1. ANalysis Of VAriance (ANOVA) . . . 70
A.4.2. Tukey's test . . . 71
A.5. Técnicas de machine learing . . . 73
A.5.1. Gradient Boosted Regression Trees (GBRT) . . . 73

Bibliografía **75**

Capítulo 1

Introducción

1.1. Planteamiento del problema

Las ciudades son fuentes ricas e interesantes para la investigación y descubrimiento de patrones. El crecimiento en las ciudades viene acompañado de un incremento de la población, y por consecuencia, de la demanda de transporte para agilizar el proceso de traslado de sus habitantes. Esto último también provoca el incremento de factores indeseables como lo son las emisiones de carbono (contaminación) o embotellamientos de tráfico. Este tipo de problemas lleva a las ciudades a considerar formas alternativas de movilidad en áreas urbanas como una medida para enfrentar los problemas de demanda en movilidad, como caminar o el ciclismo. Los sistemas de bike sharing o bicicletas compartidas (BSS por sus siglas en inglés) son servicios implementados en distintas metrópolis del mundo que proveen a los suscriptores acceso a bicicletas para viajar distancias cortas a través de su red de estaciones. Estas bicicletas están disponibles en quioscos automatizados siendo un servicio público/concesionado. Para su uso, membresias de diferentes tipos pueden ser adquiridas pagando la cuota correspondiente. Esto permite mantener un registro del historial de viajes el cual utiliza sistemas identificadores de radio frecuencia (RFID por sus siglas en inglés) integradas en las estaciones; Debido a la flexibilidad en las dinámicas de los BSS, los viajes en bicicleta realizados por los suscriptores pueden conducir a dos tipos de problemas:

- Los problemas de largo plazo como la planeación urbana donde un análisis y predicción en el actual BSS debe ser realizado para realizar mejoras en el servicio. Esto involucra la agregación y/o eliminación de estaciones en diferentes puntos de la ciudad, así como ajustar los tamaños de las estaciones, e.g. ampliar los estacionamientos en estaciones que se quedan rápido sin espacios.
- Los problemas de corto plazo se refieren a las interacciones entre las estaciones. Entre estos problemas se considera que no todas las estaciones son usadas de una manera uniforme; existen estaciones con un tráfico pesado de bicicletas a ciertas horas del día, dejando estaciones llenas donde sus usuarios no pueden

realizar devoluciones. Por otro lado existen algunas estaciones con un tráfico de salidas pesado dejando estaciones vacías donde los usuarios no pueden tomar bicicletas. La tarea aquí consiste en balancear la oferta y la demanda; tanto de bicicletas como de espacios disponibles generada a partir de estas interacciones (bike balancing problem) [Chemla et al.].

Una importante necesidad para los BSS ha sido el encontrar una manera de predecir cómo se comportará una estación, cuáles factores están presentes en una estación con alta demanda y también qué tipo de comportamiento tendrá. Este conocimiento es crucial para hacer una mejor y más eficiente planeación en el servicio ofrecido por la ciudad [Vogel et al.], puede ayudar a planeadores urbanos a lidiar o incluso prevenir los problemas antes expuestos, por ejemplo, en la creación de nuevas políticas para incentivar a los suscriptores a colaborar en la distribución de bicicletas. Las estaciones tienen un comportamiento distinto de acuerdo a su ubicación donde muchas de estas pueden ser utilizadas para transporte de trabajo u ocio. Este tipo de demanda afecta en gran parte el sistema, haciendo el servicio muy ineficiente y poco placentero. Con tal de obtener una mejor experiencia de parte del servicio es muy importante la habilidad para predecir una situación antes de que ésta ocurra, conduciendo a mejores respuestas frente a la demanda en movilidad como el transporte de bicicletas en exceso de una estacion llena a una vacía o, si es posible, la adición de estaciones de tal manera para evitar que esta situación se manifieste en primer lugar. Para lograr ésto es importante el entender la manera en cómo este servicio reacciona a la dinámica demanda dentro de la ciudad.

Con respecto a lo anterior se puede decir que algunas soluciones ya han sido aplicadas para el manejo de este problema, como lo son: el hacer uso de camiones para transportar bicicletas de estaciones llenas a estaciones vacías para asegurar la disponibilidad de bicicletas, el uso de aplicaciones móviles que muestren el estado actual de las estaciones y, recientemente, flexibilidad de dar tiempo extra a los usuarios para alcanzar otra estación de bicicletas, si la estación en la que se intenta realizar la devolución se encuentre llena. La adopción de los BSS va en aumento y con ello también el conocimiento generado de las huellas digitales que está a disposición en las ciudades que ya cuentan con ellas. Ésto puede contribuir con una estimación factible para obtener un mejor entendimiento de la movilidad urbana que puede ser utilizada para hacer futuras estimaciones sobre sus características como: costos de operación, redistribución de bicicletas sobre las estaciones, una mejor estimación sobre los viajes a estaciones claves y la determinación para su capacidad. Pero incluso con la disponibilidad de grandes cantidades generadas por los historiales de viajes es difícil obtener conocimiento de éstos sin el uso de algoritmos automatizados que revelen patrones urbanos ocultos y fenómenos generados por la dinámica urbana.

1.2. Objetivo general

Utilizar tanto técnicas nuevas como tradicionales en la creación de perfiles para las estaciones del BSS de Ecobici y aprovechar esta nueva información en mejorar la calidad en los pronósticos realizados.

1.2.1. Objetivos específicos

- Modelar el uso de las estaciones de servicio directamente.
- Caracterizar el comportamiento de las estaciones utilizando técnicas de agrupamiento.
- Aplicar diversidad de rango sobre el sistema para conocer y corroborar su comportamiento.
- Detectar y establecer zonas de predictibilidad para los rangos de las estaciones en función de la diversidad de rango.
- Demostrar que las regiones establecidas favorecen en la predicción de los rangos esperados en el sistema.
- Sustentar que el uso de diversidad de rango en conjunto con técnicas de predicción ofrecen importantes características que ayudan para el pronóstico de un BSS, haciendo uso de múltiples enfoques para la observación en propiedades de un sistema dinámico que ayudan en la adaptación periódica de regiones de predictibilidad.

1.3. Motivación

Los problemas viales que surgen en las ciudades producto de la interacción entre sus habitantes generan una serie de redes interactivas y dinámicas que cambian constantemente su esquema con el tiempo. las herramientas de predicción convencionales tienen la limitación de ajustar un modelo a un fenómeno estático, reduciendo la aplicabilidad para generar soluciones a problemas de gran envergadura al estar sometidos a constantes cambios a través del tiempo.

No obstante, dichas herramientas pueden complementarse de otras técnicas existentes para que en conjunto se obtengan soluciones adaptativas más robustas para problemas donde el constante cambio y la incertidumbre es inevitable.

Este trabajo surge como un proyecto de investigación cuya intención es utilizar y comparar en conjunto dos herramientas que provienen de diferentes áreas. Como objeto de estudio se seleccionó el BSS de la ciudad de México con el objetivo de observar si dicho sistema exhibe comportamientos regulares que ayuden en el entendimiento y mejor planificación de dicho sistema. Se propone encontrar regularidades en las estaciones

para tener un mayor entendimiento sobre los aspectos a considerar en la mejora de la infraestructura.

La medida de diversidad de rango [Cocho et al.], de la que se hace uso en este trabajo ha mostrado que sistemas que cuentan con una alta interacción entre sus componentes suelen desarrollar una estructura la cual podemos obtener provecho al conocerla. Para este caso una estructura jerárquica se construye haciendo uso de dos factores: una métrica obtenida con base en el número de bicicletas que interactúan con las estaciones la cual pueda medirse la magnitud de uso en las estaciones y un ordenamiento creado a partir de dicha métrica tomando en cuenta un periodo comprendido. Finalmente, tras conocer si Ecobici manifiesta una estructura descrita por la diversidad de rango lo que se propone es establecer que existe una relación entre el lugar que las estaciones ocupan en la diversidad de rango y el nivel de predictibilidad del rango de las estaciones durante un periodo comprendido se realiza el perfilaje sobre las estaciones descrita en los capítulos 3 y 4 para comprobar si los elementos con el mismo perfil comparten características similares haciendo uso de técnicas de aprendizaje automatizado.

1.4. Contribuciones

La hipótesis en realizar este pronóstico es el de sugerir que existen diferentes grados en la predictibilidad del sistema distintos de acuerdo a los rangos formados por la actividad generada de sus estaciones. Las contribuciones que este trabajo aporta son las siguientes:

- El uso de las regiones en diversidad de rango (rank diversity en inglés) conocidas como cabeza, cuerpo y cola [Apéndice A – Obtener regiones de cabeza, cuerpo y cola] para la creación de perfiles en las estaciones de Ecobici.
- Un nuevo tipo de visualización a partir de diversidad de rango el cual no se había realizado anteriormente nombrado frecuencia de rango que ayuda a observar a nivel individual la presencia que tiene una estación sobre los rangos.
- Se realiza un proceso de agrupamiento utilizando los conteos producto de las interacciones entres las estaciones y los viajes realizados en durante los horarios operativos de Ecobici para cada uno de sus diferentes periodos en sus fases de expansión, cosa que anteriormente no se había llevado a cabo en ningún BSS.
- Se ofrece un análisis cualitativo de los resultados obtenidos de estas dos diferentes técnicas, apuntando las similitudes que resaltan entre ellas.
- Se hace uso de un modelo de predicción propios del área de aprendizaje automatizado para realizar el pronóstico del ordenamiento de uso (ranking) de las estaciones que forman parte de una misma categoría y obtener comparativas entre la calidad de las mediciones del error en el pronóstico realizado sobre estos diferentes grupos.

1.5. Estructura de la tesis

Este trabajo está dividido en seis capítulos donde se explica de manera detallada los procedimientos realizados para el tratamiento y limpieza de los datos, el análisis exploratorio de datos, la diversidad de rango que presentan las estaciones de BSS y la selección y la predicción por uso y rango de la actividad de las estaciones.

En el capítulo dos se habla del estado de arte vinculado con BSS en otros paises. Sobre cómo la información obtenida en estos trabajos ayuda a construir un mejor entendimiento sobre los BSSs para realizar una mejor gestión en el diseño o para implementación en una ciudad que no cuenta con uno o como medidas para gestionar uno ya existente.

El capítulo tres describe la procedencia de los datos, el tipo de información y la organización con los que éste cuenta, el periodo de tiempo que comprenden los registros, los diferentes eventos espacio-temporales que son capturados a lo largo de este periodo, también se detalla realizando un análisis de datos exploratorio las diferentes tendencias culturales y sociales que suelen tener los subscriptores en el sistema cuando realizan sus recorridos mediante Ecobici. La transformación a la que estos datos fueron sometidos en orden de ofrecer una estructura en una forma que resultara más sencilla de analizar y experimentar. También se habla sobre el fenómeno que representa para Ecobici la expansión del sistema hacia otros sectores de la ciudad y su debida separación para sus diferentes fases tomando en cuenta factores importantes presentes en una ventana de tiempo correspondiente. Por último se obtienen perfiles sobre la actividad tomando los conteos generados por el flujo de entradas y salidas a lo largo del día en las estaciones de Ecobici con el uso de una herramienta de agrupamiento conocida como clustering. Adicionalmente el análisis sobre la correlación entre los factores que se agrupan es realizado para mejorar el desempeño en la agrupación y explicar las similitudes y diferencias que comparten las estaciones dentro del sistema.

El capítulo cuatro detalla el concepto de diversidad de rango, sobre la idea principal que esta técnica establece y se da una definición sobre los diferentes elementos que éste maneja. Este capítulo también relata cómo se ha utilizado la diversidad de rango para establecer jerarquías sobre los sistemas que son objetos de estudio. Se realiza la construcción de diferentes diversidades de rango en periodos correspondientes a la apertura de diferentes fases en Ecobici.

El capítulo cinco detalla el modelo empleado en este trabajo para el ajuste de los datos históricos obtenidos para realizar pronósticos en el ordenamiento del sistema realizados para un conjunto de días no observados. Se detalla cómo el conjunto de datos del que se dispone fue separado en dos partes necesarias para comprobar el poder predictivo del modelo: el conjunto de entrenamiento y de prueba. Haciendo uso de las estaciones etiquetadas por su pertenencia según su distribución la diversidad de rango, el conjunto de entrenamiento-prueba y el modelo para realizar los pronósticos sobre el rango de las estaciones se realiza una comparativa tomando en cuenta los errores que fueron obtenidos haciendo uso de dos métricas: La raíz del error medio (RMSE) y la ganancia descontada acumulativa normalizada (NDCG) aplicando técnicas estadísticas

para comprobar la hipótesis aquí planteada.

El capítulo seis expone las conclusiones a las que se llegaron dados los resultados, las posibilidades que se podrían explotar utilizando el conocimiento generado por los perfiles obtenidos con la técnica de clustering, las características exhibidas descritas por la diversidad de rango y el trabajo futuro que puede desarrollarse dados los resultados.

Capítulo 2

Estado del arte

Los BSS son sistemas relativamente nuevos en el mundo. Los problemas que se han visto emerger de estos sistemas son ofrecer una mejor gestión y el aseguramiento de la calidad del servicio tanto en disponibilidad de bicicletas. Investigaciones anteriores han sido realizadas para establecer un contexto de entendimiento entre el sistema y los múltiples factores que afectan a éste como:

- Caracterizar los viajes realizados por los diferentes tipos de usuarios que hacen uso del sistema.
- El perfilaje de las estaciones para establecer relaciones de comportamiento inherentes en diferentes secciones de la ciudad.
- El ajuste de modelos predictores a diferentes escalas para entender los flujos en el sistema a lo largo del día, así como los factores que están involucrados con las magnitudes y flujos.

El tipo de datos que pueden llegar a manipularse a partir de estos BSS pueden variar de acuerdo a cual sea el objeto de estudio. Algunos de los estudios que se han realizado con anterioridad se mencionan los siguientes:

El estudio realizado por [Froehlich et al.] en el que realiza un análisis espacio-temporal de la demanda en los BSS en la ciudad de Barcelona, España conocido como Bicing, haciendo uso de datos correspondientes a 13 semanas del estado de disponibilidad de las estaciones, creando una medida normalizada obtenida de la división de las disponibilidad entre la capacidad de bicicletas con la que cuenta. En esta investigación la métrica utilizada para medir la distancia entre la demanda de las estaciones en diferentes horas de tiempo fue Dynamic Time Warping debido al interés que se tiene en comparar patrones temporales a lo largo de una ventana de tiempo. También se hace uso de técnicas de modelos de regresión para predecir la actividad de demanda en las estaciones a cierta hora del día el cual corresponde al flujo de tráfico en la ciudad. Este estudio además extiende una caracterización de patrones temporales entre las estaciones para investigar y relacionar los comportamientos de uso que existen y cómo

están distribuidos sobre la ciudad utilizando una técnica de agrupamiento conocida como agrupamiento jerárquico, aplicando esta técnica para dos segmentos de la semana distintos: días entre semana y días en fin de semana. Al final de este estudio se realiza una predicción en el uso de las estaciones haciendo uso de modelos predictivos simples donde se mostró que con capaces de clasificar el estado de una estación (estación llena o estación vacía) con un 80 % de eficiencia en un pronóstico de 2 horas en el futuro.

En el trabajo realizado por [Borgnat et al.] analizan un conjunto de datos correspondiente a un periodo de actividad de 2 años del BSS de Lyon, velo'v, Francia, compuesto por 13 millones de viajes registrados. Éstos fueron analizados con técnicas para procesamiento de señales para el pronóstico cuantitativo de los viajes realizados por los usuarios en el sistema. Se construyen dos modelos de regresión diferentes: un modelo de regresión que se ajustara a la actividad generada por el sistema de los conteos por hora. Por una parte, un modelo autoregresivo de orden 1 tomando en cuenta entradas exógenas fue utilizado y otro modelo conteos por día donde se utilizan para construir el vector de características así como la temperatura, volumen de lluvia, días festivos y días de huelga. El segundo enfoque es acerca de la distribución espacial del sistema. Dado que la demanda en entradas como salidas de las estaciones no es uniforme, el objetivo aquí fue utilizar modelos para aprender de manera automática las dinámicas de movimientos generados en la ciudad a varias horas del día. Interpretando al BSS como una red donde las estaciones son nodos y los viajes realizados como aristas entre las estaciones A y B, se transformó la dinámica del sistema a un problema de grafos dirigidos con diferentes escalas de tiempo. Posteriormente se implementa un algoritmo de agrupamiento jerárquico se agrupan aquellas estaciones cuya interacción entre ellas sea lo suficientemente estrecha para detección de comunidades. Finalmente, este estudio resalta el uso de suelo en diferentes localidades de la ciudad utilizando el flujo de tráfico característico que tienen las estaciones característicos con respecto a sus horarios.

En [Etienne and Latifa.] con los registros generados del BSS de Paris, Vélib que constan de un periodo de 30 días, con alrededor de 2,500,000 viajes totales registrados. hace uso de un modelo generativo de mezclas (Mixture models en inglés) de variable latente utilizando una distribución de Poisson para realizar una partición entre las estaciones en términos de su dinámica temporal sobre el día respecto al número de bicicletas rentadas y entregadas para describir perfiles de comportamiento basados en clustering de conteos en series de tiempo. Los viajes son transformados para crear un tensor con dimensiones N x D x T donde: N representa el número de estaciones, D el conjunto disponible de tiempo y T la longitud del vector de descriptores construidos, 48 para este caso. Debido a que la actividad es diferente para los días entre semana y los fines de semana se realiza tal separación y su análisis por separado. Los resultados ofreces información que puede ayudar a develar el valor de suelo en diferentes regiones situadas en la ciudad así como una explicación sobre los factores que rodean a dichas estaciones.

En el estudio de [Rixey] se realiza un análisis sobre variables demográficas e infraestructura seguido de la construcción de un modelo de regresión lineal multivariado que pudiera explicar la actividad mensual sobre los BSS ubicados en tres ciudades distintas

de Estados Unidos: Capital Bikeshare en Washington, D.C; Nice Ride MN en Minneapolis, Minnesota y Denver B-cycle en Denver, Colorado. Utilizando datos obtenidos de distintas fuentes: conteos mensuales de los diferentes BSSs, variables demográficas obtenidas del censo 2010 en conjunto con otros órganos gubernamentales realizados en cada ciudad: densidad de población, densidad de áreas de trabajo, media en nivel socioeconómico, educación, presencia de ciclopistas entre otras, polígonos georeferenciados de rutas, carreteras y manzanas, en conjunto de los valores demográficos propios de cada zona provenientes de información geográficas (GIS en inglés).

Los hallazgos en dichos estudios han estado presentes en muchos otros BSS que existen en otras ciudades repartidas en el mundo, como que las estaciones que se encuentran ubicadas más alejadas de el centro de la ciudad suelen contener una tasa de actividad más baja. También, al tratarse de una técnica de regresión lineal, cuenta con aquellas ventajas presentes que ofrece la herramienta como la interpretabilidad dada por el peso de los coeficientes ajustados para cada descriptor, como por ejemplo, la influencia en eficiencida en predicción con la modificación en el radio que se tiene para tomar en cuenta a las estaciones vecinas del sistema, evidenciando que la proximidad que tiene una estación con respecto a sus vecinas influye en el poder de predicción, apuntando la alta influencia que tiene ser un sistema interconectado.

La idea de anticipar las futuras demandas de las estaciones es una necesidad de gran interés para el balanceo en las estaciones, es por eso que [Dias et al.] realiza una predicción para el estado de demanda que las estaciones tendrán hasta con 72 horas de anticipación, haciendo uso de una colección de metadatos obtenidos por la API del sistema de bicicletas compartidas en Barcelona, Bicing (con alrededor de 10 gigabytes de información) y utilizando un método de ensambles conocido como random forest para pronosticar de entre cinco etiquetas que se asignan al grado de demanda que cada estación tendrá. Este modelo de ensambles es comparado con otra técnica conocida como SARIMA (Seasonal Auto Regressive Integrated Moving Average). Con la finalidad de medir el desempeño de ambas técnicas se realizan predicciones en el estado de las estaciones, midiendo la sensibilidad (la relación entre el total de verdaderos positivo y el de positivos) y la especificidad para cada método donde un positivo es considerado si el modelo predice que la estación estará vacía o llena. Los resultados obtenidos favorecen a la técnica de random forest, pudiendo generar una frontera de decisión mucho más flexible prediciendo cerca del 50 % de los estados, además de que el modelo puede tomar en cuenta factores extra además del estado de disponibilidad de las estaciones en el pasado, como los factores climáticos. Dicho estudio también apunta a que existe un grado de predictibilidad distinto en diferentes estaciones, indicando que cada estación puede tener un 'perfil' donde algunos factores influencían más el uso de unas estaciones que a otras.

Capítulo 3

Datos

En este capítulo, se presenta el origen y descripción de los datos utilizados, se decriben algunas fórmulas que fueron utilizadas para transformar los registros generados por las estaciones en contéos. También se hacen uso de herramientas de visualización para tener una mejor idea de las dinámicas comprendidas dentro del sistema.

3.1. Bike sharing systems

Bike sharing system o sistemas de bicicletas compartidas es un servicio donde bicicletas se ponen a la disponibilidad de individuos bajo uno términos base muy accesibles. Los esquemas de bicicletas compartidas permiten a las personas tomar una bicicleta del punto A y hacer entrega de ella en el punto B. Muchos sistemas de bicicletas compartidas ofrecen suscripciones para viajes donde los primeros 30 a 45 minutos (dependiendo de que BSS se esté hablando) son de uso gratuito o de muy bajo costo, alentando su uso como medio de transporte. Esto permite a cada bicicleta servir a varios usuarios por día. En la mayoría de las ciudades donde se tiene implementado un BSS han implementado el mapeo vía Smartphone mostrando estaciones cercanas con bicicletas y estacionamientos disponibles.

Los BSS pueden ser divididos en dos categorías generales: “Programas de bicicletas comunitarias” organizadas en gran parte por comunidades u organizaciones sin fines de lucro; y “programas de bicicletas inteligentes” implementadas por agencias de gobierno y algunas veces en asociación pública-privada. El concepto central de estos sistemas es el de proveer acceso gratuito o accesible de bicicletas para viajes de cortas distancias en un área urbana como alternativa al transporte motorizado público o privado, así reduciendo los factores contaminantes como congestión de tráfico, ruido y emisión de gases. Los BSS han sido citados como una forma para resolver el problema de la última milla [for local goverment] y conecta a sus usuarios con redes de tránsito público. Las razones por la cuales las personas utilizan BSSs varían de manera considerable. Algunos utilizarían bicicletas propias, pero les preocupan los temas de robo y vandalismo, estacionamiento y requisitos de mantenimiento. Con límites para el número de lugares donde las bicicletas

pueden ser rentadas o regresadas, el servicio se parece a tránsito público y ha sido criticado como menos conveniente que el uso de una bicicleta propia. Los gobiernos que cuentan con programas BSS también han mostrado ser un servicio costoso a no ser que sea subsidiado por intereses comerciales, típicamente en mostrada en forma de publicidad en las estaciones o en las bicicletas mismas. En años recientes, en un esfuerzo por reducir pérdidas por robo y vandalismo, muchos esquemas de bicicletas compartidas ahora requieren que el usuario provea un depósito monetario u otra garantía, como convertirse en un suscriptor mediante la realización de pagos.

3.2. Ecobici

Ecobici es el primer BSS implementado en la ciudad de México. Comenzando en febrero del 2010 con 90 estaciones y 2000 bicicletas, desde entonces el programa ha sido expandido en 4 fases adicionales, incrementando la cobertura en diferentes delegaciones y el número de estaciones y bicicletas a 444 y 6000 respectivamente(17). Los miembros pueden pagar por diferentes tipos de membresía con diferentes costos de acuerdo a su duración, que puede ser desde anual o temporal (por 1, 3 o 7 días). Los viajes son libres de cargos por viajes con duración menor a los 45 minutos, agregando una cuota de penalización por exceder esta duración. La ruta más común corre desde Reforma hasta el Zócalo, en el centro de la ciudad.

3.3. Fuente de datos

Lab CDMX es un programa dedicado a motivar a participantes a involucrarse en los problemas de la ciudad de México con el objetivo de promover la colaboración con cualquier persona interesada en generar nuevas propuestas e ideas para mejorar las políticas del sistema, dinámica urbana y diseño de la ciudad. En el año 2014 lanzó un concurso haciendo públicos los conjuntos de datos de los registros de viajes de Ecobici(3), después liberados por el gobierno y ahora disponibles desde la página oficial de Ecobici. La base de datos de registros temporales consiste en aproximadamente 4 años de actividad (desde el 2 de febrero del 2010 hasta el 31 de diciembre del 2013). La tabla 3.1 muestra los campos del historial de registros y la tabla 3.2 contiene la descripción sobre las estaciones. También existe una tabla que contiene distancias entre todas las estaciones de Ecobici. La medida es una distancia simétrica Euclidiana, tomando una forma de matriz triangular superior.

Sin importar el hecho de que el conjunto también contiene datos con la descripción acerca de los suscriptores y las series de conteos relacionadas con una ruta específica sobre la calle Reforma, estos conjuntos de datos no fueron utilizados debido a que el principal objetivo en este trabajo es la obtención de perfiles de las estaciones tomando en cuenta su posición en la distribución de rangos utilizando diversidad de rango, los

cust_id	bike	date_removed	station_removed	date_arrived	station_arrived	action
6	89	02/02/2010,01:57:00	56	02/02/2010,01:58:00	56	C

Tabla 3.1: Registros temporales de Ecobici

id	principal	secundario	referencia	colonia	delegacion	longitud	latitud	nombre
1	rio balsas	rio sena		Cuauhtemoc	Cuauhtemoc	-99.16848	19.43293	rio balsas-rio

Tabla 3.2: Estaciones de Ecobici.

únicos archivos de interés son aquellos que contengan los registros de viajes realizados y la ubicación de las estaciones.

3.4. Limpieza, pre-procesamiento y análisis de los datos

3.5. Remoción de datos atípicos

La información obtenida por Lab CDMX contenía algunos registros que tuvieron que ser removidos primero en orden de comenzar un análisis con ellos. Los viajes cancelados tuvieron que ser removidos para eliminar todos aquellos registros cuyo origen y destino fueran la misma estación y cuya duración de viaje haya sido menor a 1 minuto. Éstos posiblemente expliquen devoluciones realizadas debido al malfuncionamiento. Por lo tanto, del total de registros fueron removidos aproximadamente el 5 % (724 000 registros).

Los registros fueron transformados primero en orden de realizar un mejor análisis en la actividad de las estaciones en el sistema. Esta transformación consiste en convertir los conteos de entradas y salidas por estación sobre un periodo de tiempo Δ Ec. (3.1), donde d representa múltiplos del intervalo de periodo Δ, s es la ID de la estación y x representa el total de viajes de entrada/salida de una estación s en d, dado por la Ec. (3.2).

$$X_{\Delta} = \begin{bmatrix} x_{11} & \cdots & x_{1s} \\ \vdots & \ddots & \vdots \\ x_{d1} & \cdots & x_{ds} \end{bmatrix} \tag{3.1}$$

$$x_{ds} = \sum_{d \in \Delta}^{d} INCOME_{ds} + \sum_{d \in \Delta}^{d} OUTCOME_{ds} \tag{3.2}$$

Δ = 24 horas, entonces el total de entradas/salidas de las estaciones por día es obtenido. Desde la apertura del sistema hasta el final del periodo de registros de viajes comprendido en este estudio, solo existían 275 estaciones cuyas IDs están enumeradas del 1 al 275. Los conteos de actividad para las estaciones con IDs superiores a los 275 no fueron contemplados para el posterior análisis, removiendo 7 estaciones con IDs no válidos.

En la fig.(3.1)-A se muestra la tendencia global del sistema con un intervalo de confianza del 95 % usando la desviación estándar móvil con 20 observaciones pasadas y considerando solo el conjunto de días entre semana. Puede apreciarse también que algunos picos de baja actividad sobrepasan el intervalo de confianza en conteos del sistema. Tomando las fechas correspondientes a esos días de actividad inusuales se observa que algunas de esas fechas corresponden a días festivos nacionales o internacionales, por ejemplo, el 16 de septiembre del 2010 que corresponde al día de independencia de México o el 24 de diciembre del 2012 que corresponde a navidad. Estos días también fueron excluidos del análisis para suavizar la señal en las estaciones, reduciendo datos atípicos tal como se realiza en [Froehlich et al.]. La fig.(3.1)-B muestra el mismo periodo de actividad sin estos días.

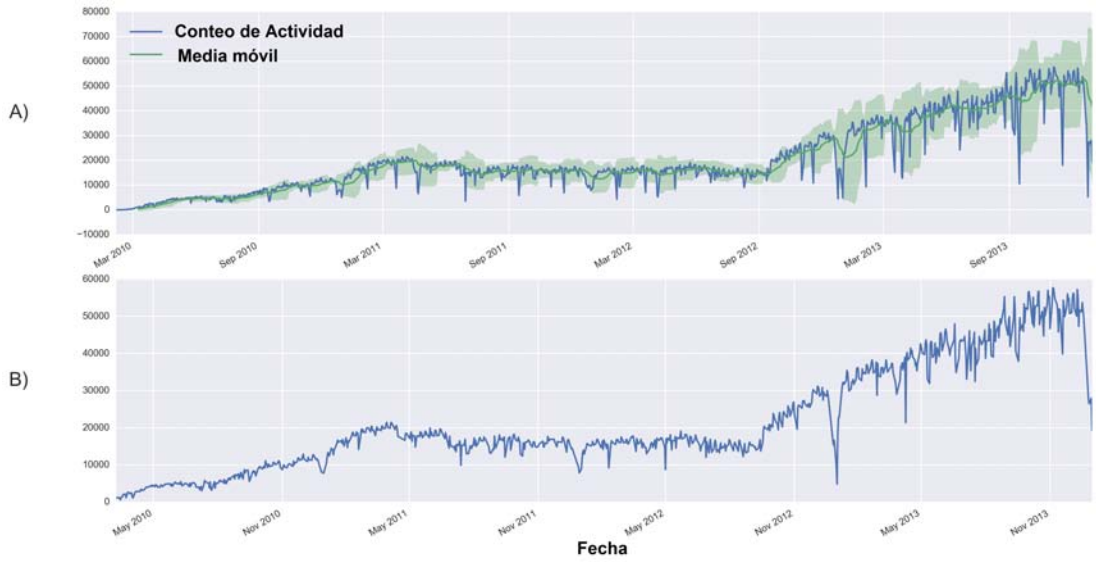

Figura 3.1: Tendencia de conteos en días entre semana. - A) La actividad global de Ecobici junto con la media móvil y desviación estándar móvil de 20 observaciones pasadas. B) La actividad global de Ecobici removiendo aquellos días cuyos conteos estén fuera del intervalo de confianza.

El número de bicicletas y miembros en el sistema ha ido en aumento desde el que se inició este servicio. La fig.(3.2) muestra el número de suscriptores y bicicletas en función del tiempo, claramente la adición de nuevas fases también ha incrementado la

pendiente de tanto suscriptores como bicicletas. Este gráfico no puede distinguir entre suscriptores de largo plazo de los de corto plazo, sin embargo, una clara tendencia puede ser observada.

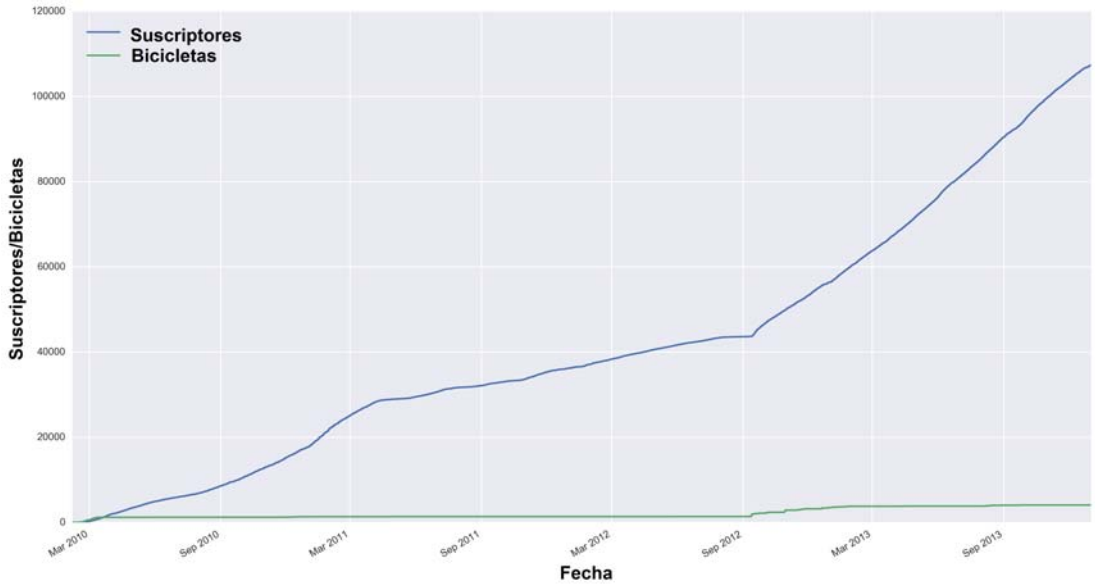

Figura 3.2: Suscriptores/Bicicletas en función del tiempo. - El número único de suscriptores y bicicletas en función del tiempo.

También es interesante notar la media de la distribución sobre la cual se centra la duración de los viajes realizado en Ecobici. En la fig.(3.3) un histograma de la duración de viajes de Ecobici tomado de todo el periodo de tiempo es presentado. La tendencia está situada en alrededor de 8 minutos y la línea roja punteada vertical marca la política de tiempo límite de Ecobici. Esto indica que la mayoría de los viajes realizados son para recorrer distancias relativamente cortas.

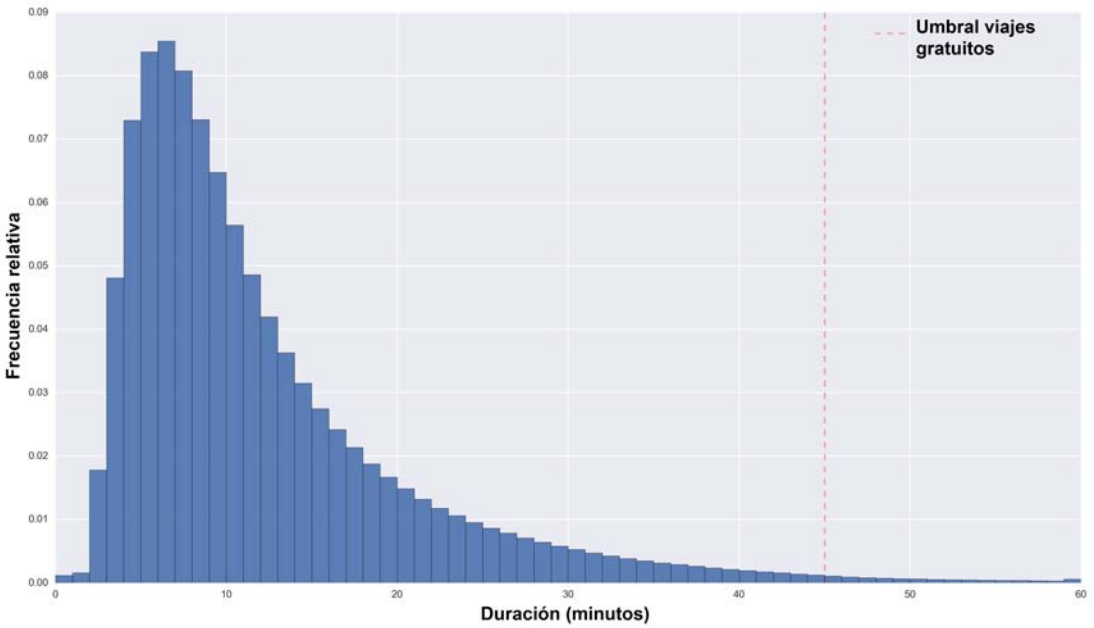

Figura 3.3: Histograma de la duración de viajes. - El histograma muestra la distribución de los viajes realizados por Ecobici. La línea roja punteada marca el límite de tiempo para un viaje antes de incurrir en una cuota de penalización.

También, la elevación en los viajes puede ser un factor importante en la planeación de sistemas de bicicletas compartidas como lo es Ecobici. Los viajes colina abajo suelen ser más populares que los viajes colina arriba. Para obtener la diferencia en la elevación para este conjunto de datos se hizo uso de una herramienta para obtener la elevación en cada estación(12) después se realizó una resta sustrayendo el valor de la elevación en la estación de origen del valor de la estación destino para cada viaje realizado. En la fig. (3.4) se muestra un histograma de la ganancia de elevación para los viajes realizados, una ganancia negativa corresponde a viajes realizados cuesta abajo mientras que una ganancia positiva son viajes realizados cuesta arriba. Los valores de cero o muy cercanos representan viajes sin ninguna elevación importante. Una posible explicación para los resultados obtenidos puede ser debido a que México se encuentra en un valle, y debido a eso en un área plana.

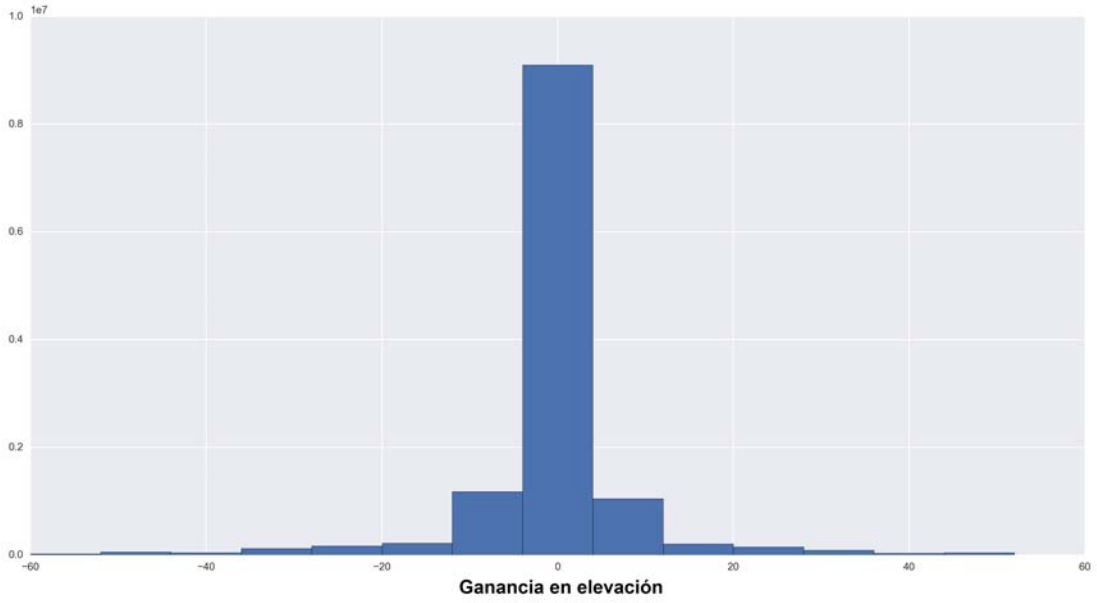

Figura 3.4: Histograma de elevaciones por viaje. - Ganancias negativas significan viajes cuesta abajo. Ganancias positivas significan viajes cuesta arriba.

3.6. Separación de periodos: fase 1, 2 y 3

Existen algunas discrepancias sobre la fecha exacta en la que las nuevas fases fueron incorporadas y no existe un registro oficial que indique la adición de las mismas, así que se decidió extraerlas de manera arbitraria. Para ello se obtuvieron las fechas iniciales, las estaciones fueron agrupadas en conjuntos de acuerdo a la pertenencia de su fase, obtenida desde el sitio oficial, Haciendo uso de X con $\Delta = 24$ horas en (Ec. (3.3)). De esta manera se construye una razón entre las estaciones usadas sobre el total de estaciones que pertenecientes del conjunto correspondiente de su fase, una estación usada puede entenderse como $x_{ds} > 1$ (Ec. (3.4)) donde x corresponde al conteo de bicicletas que la estación s tuvo en el día d. Fue decidido de manera arbitraria que una fase se consideraba activa si ésta contenía el 60 % de estaciones usadas de su fase correspondiente P durante el día d (Ec. (3.5)) donde S_p son todas las estaciones con una pertenencia a la fase p, obteniendo la tabla 3.3. La fig. (3.5) representa la tasa de actividad para las tres fases desde su inauguración hasta el último día que comprenden los registros. La fase 1 (Rojo) tiene una muy alta tasa de actividad desde el comienzo, como es de esperarse. Sin embargo, puede apreciarse que no hay una separación clara entre la fase 2 (verde) y la fase 3 (azul) cuyo umbral de activación no está lejos la una de la otra, haciendo difícil el realizar una separación clara de las fechas de comienzo entre estas dos fases.

$$Dates = f(X_{p\Delta=24})', \text{ para cada p} \in \text{P} \tag{3.3}$$

$$f(x) = \begin{cases} 1, & \text{si } x_{ds} > 0 \\ 0, & \text{de lo contrario.} \end{cases} \tag{3.4}$$

$$f(X)' = \begin{cases} 1, & \text{si } \frac{1}{S_p} \sum_{i:S_i=P}^{S_p} f\left(X_{ds}\right) > 0.6 \\ 0, & \text{de lo contrario.} \end{cases}, \forall d \in D \tag{3.5}$$

Tabla 3.3: Fechas de comienzo de acuerdo al criterio.

Fase	Fecha de comienzo
1	16/02/2010
2	29/10/2012
3	29/11/2012

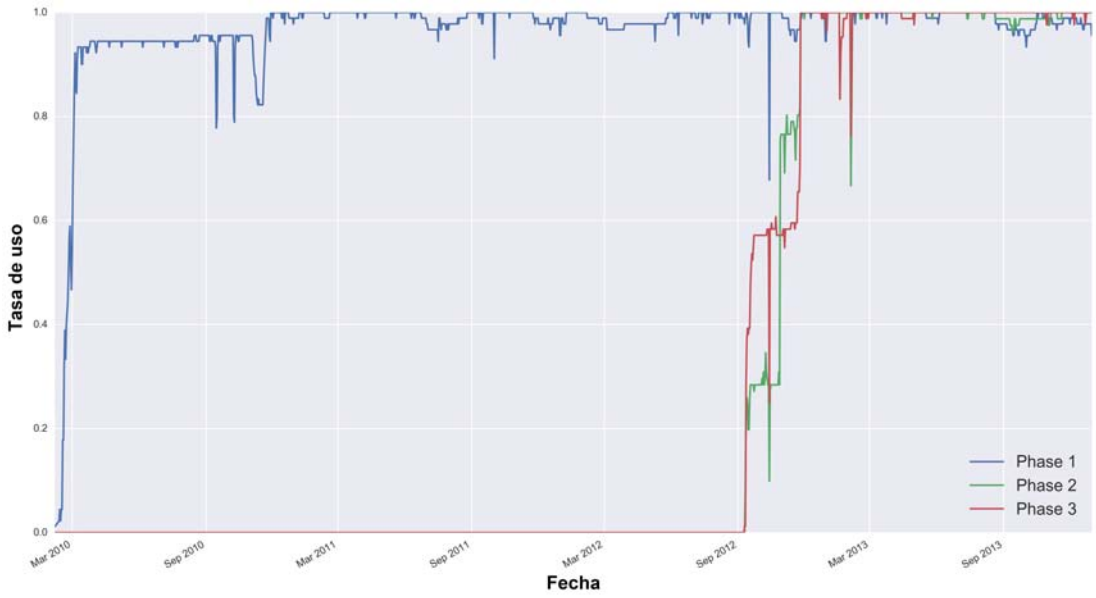

Figura 3.5: Razón de uso para las estaciones a lo largo de los días. - Cada línea representa cuando la razón de uso de una fase en función del tiempo. Rojo) Razón de uso para la fase 1. Azul) razón de uso para la fase 2. Verde) razón de uso para la fase 3.

Realizando una suma sobre las filas en X con $\Delta = 24$ horas podemos observar la tendencia en viajes que tiene el sistema en el curso del tiempo del periodo total

mostrado en la fig.(3.6) donde las líneas punteadas marcan las fechas estimadas donde se estima la incorporación de las fases 1, 2 y 3 desde la apertura de Ecobici, incorporando nuevas estaciones de servicio en nuevas zonas. Puede ser apreciado que las fluctuaciones sobre los días generados por la actividad de los fines de semana, haciendo constatar la diferencia de uso que tiene el sistema entre fines de semana y días entre semana. Por esta razón ha sido decidido el separar el conjunto de datos en dos diferentes grupos de días: Fines de semana y días entre semana [Froehlich et al.] y solo enfocarse en el conjunto de días entre semana para esta investigación.

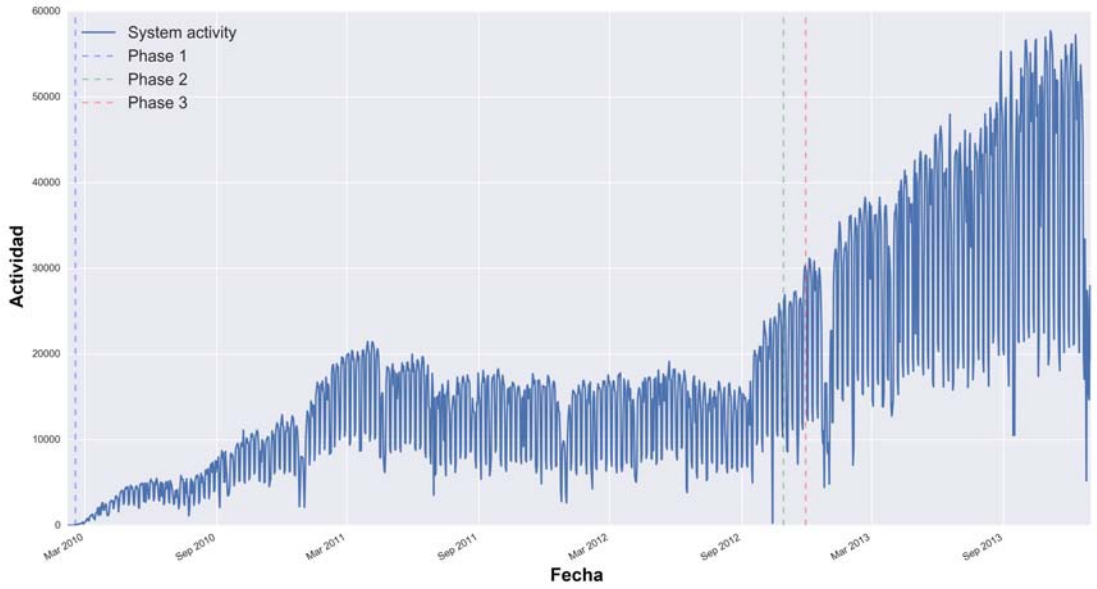

Figura 3.6: Tendencia de uso global para el sistema de Ecobici. - La línea azul corresponde al conteo total de cada día en el sistema. Línea roja punteada: Comienzo de la fase 1. Línea azul punteada: Comienzo de la fase 2. Línea verde punteada: Comienzo de la fase 3.

3.7. Obtención de perfiles para las estaciones de Ecobici utilizando clustering

Una técnica para obtener perfiles de las estaciones que ha sido aplicado en el pasado ha sido emplear técnicas de agrupación, donde una medida de similitud (o disimilitud) es dada sobre las observaciones para ser agrupadas en K diferentes grupos (mejor conocidos como clusters), cada agrupación cuenta con un punto representativo conocido como centroide, estos puntos recolectan y minimizan la distancia entre éstos y el conjunto de observaciones según la métrica utilizada. Un análisis de agrupación que se ha

realizado para este trabajo es la caracterización sobre el comportamiento de las estaciones haciendo uso de los conteos de entrada/salida que se tienen por estación a lo largo de un día promedio, utilizando una $\Delta = 1$ hora como un buen compromiso de tiempo lo suficientemente corto para capturar la actividad relevante de la estación y lo suficientemente largo para evitar una señal demasiado fluctuante [Borgnat et al.], Froehlich et al.].

3.7.1. Comportamiento promedio para cada estación

Después de ajustar el modelo de agrupación con los datos las coordenadas del centroide correspondiente puede ser interpretado como la actividad representativa por hora en un día entre semana promedio del grupo de observaciones. Las matrices separadas de entrada y salida fueron calculadas y después concatenadas formando un vector de comportamiento de actividad de 48 horas v_s. Después, en orden de reducir la longitud del vector de características de conteos por cada estación que se encuentran entre los horarios de 00:00 a las 05:00 horas para cada v_s, fueron removidos. Estos son horarios en el que el servicio de Ecobici se encuentra cerrado, dejando como resultado un vector de características de longitud 36 (Ec. (3.6)) para cada estación:

$$v_s = [x_{s1}^{in}, \cdots, x_{s18}^{in}, x_{s1}^{out}, \cdots, x_{s18}^{out}], \forall s \in S \tag{3.6}$$

Donde s representa la ID de cada estación. Un modelo de mezclas (Mixture models en inglés) es una técnica de agrupamiento suave que parte de una extensión probabilística del conocido método de las k medias (k-means) [Hastie et al. (13)]. Una de sus principales características es la propiedad de asignar una probabilidad de pertenencia a cada elemento, donde cada grupo es descrito como una densidad gaussiana, con su propia media y matriz de covarianza. Determinando la pertenencia de cada elemento a un grupo dado seleccionando el conjunto que maximice la función de verosimilitud entre la observación y el grupo k (Ec. (3.7)):

$$\hat{G}\left(v_s\right) = argmax_k \hat{p}_k\left(v_s\right) \tag{3.7}$$

Donde $\hat{G}$ representa la asociación a un grupo y $p_k(v_s)$ es la función de densidad de probabilidad para v_s evaluada en el grupo k. El modelo de mezclas fue seleccionado en vista de un desempeño superior comparado con otras técnicas de agrupación en este rubro [Vogel et al.]. La distancia Euclidiana fue seleccionada como medida de disimilitud entre las observaciones.

Una vez decidido el modelo de mixture models es necesario fijar el número de grupos k debe ser seleccionado arbitrariamente donde el número de grupos explicaría un comportamiento diferente de las estaciones en el espacio de actividad por hora. Seleccionar el número de grupos puede resultar una tarea difícil sin ayuda de herramientas que nos ayuden a evaluar los grupos generados de las distribuciones generadas para cada grupo

en este caso, con la posibilidad de conducirnos a agrupaciones espurias. En orden de obtener un criterio que ayude a seleccionar el número k de grupos más adecuado fue evaluado haciendo uso del índice de información Bayesiano (BIC en inglés) (Apéndice A - Bayesian Information Criteria (BIC)) para obtener una medida que evalúe la verosimilitud de que las agrupaciónes procedan de k grupos y penalizando al mismo tiempo la complejidad en el modelo y así evaluar haciendo uso de este índice para la selección del número de grupos k más adecuado [Pelleg et al.].

Una vez seleccionado el método de medición para la calidad de la agrupación fue realizada tomando en cuenta las fechas correspondientes a las diferentes fases permitiendo hacer un análisis más detallado del comportamiento de las estaciones en el sistema antes y después de su incorporación, presumiblemente cambiando la demanda en magnitud, comportamiento y dirección. El agrupamiento fue aplicado a los tres periodos de las fases por separado, utilizando las estaciones correspondientes disponibles durante ese periodo de tiempo.

3.7.2. Agrupación y medidas de evaluación – fase 1

La fig.(3.7) ilustra los valores aplicados a cada agrupación obtenidos por BIC, donde el gráfico muestra el número de agrupaciones en función al valor obtenido por el índice BIC; un valor mayor equivale a una mayor penalización al modelo. La leyenda corresponde a diferentes deformaciones a las que se somete la matriz de covarianza con las que se realizó cada agrupación donde:

- spherical: Corresponde a que todas las matrices de covariaza empleadas compartirán una forma esférica en la distribución de los datos.
- tied: Las matrices de covarianza se elongarán hacia una misma dirección.
- diag: Las matrices de covarianza utilizadas ajustarán una distribución con una forma más elongada.
- full: Permite a cada matríz de covarianza ajustarse como mejor posible a su conjunto de datos correspondiente.

Durante la fase 1 las estaciones no muestran un comportamiento variado, donde la agrupación con una mejor calificación muestra dos grupos con comportamientos que proceden de distintas distribuciones. La parte izquierda de la fig.(3.8) ilustra el comportamiento de cada centroide de los k grupos elegidos para realizar la agrupación, ambos mostrando un comportamiento similar con picos de actividad a las mismas horas y solo discrepando en magnitud de tráfico. En parte derecha se muestran las estaciones esparcidas por su ubicación geográfica, un grupo A ubicado en el centro y el grupo B en la zona exterior del sistema. Las estaciones que forman parte del grupo A están localizadas sobre la calle Reforma, un importante camino para quienes realizan una ruta laboral debido a que sobre esta calle se encuentran situados los más grandes e importantes establecimientos corporativos.

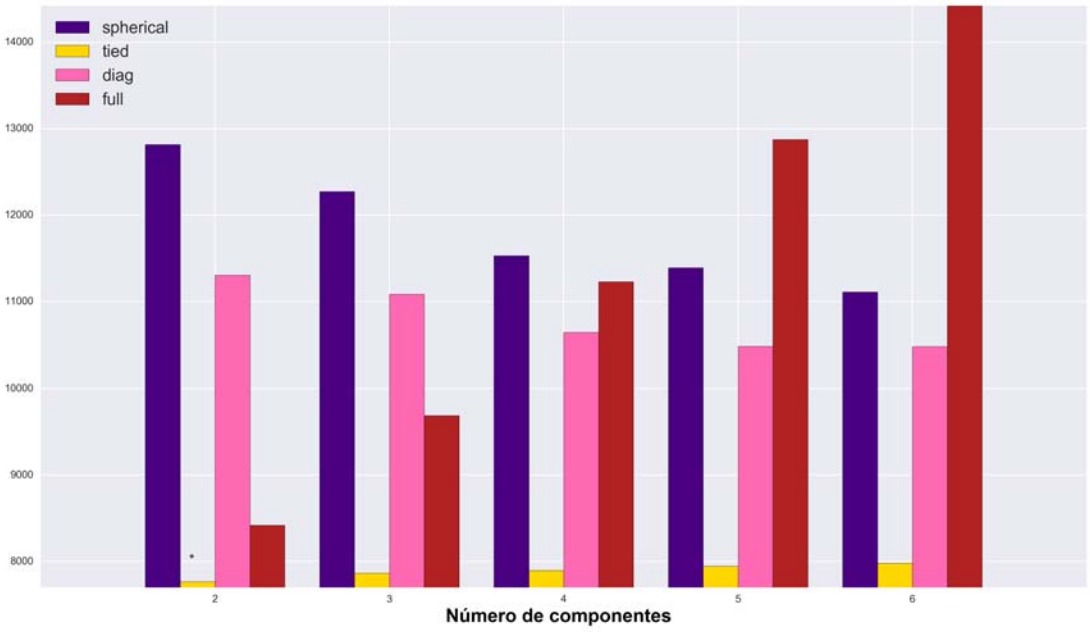

Figura 3.7: Medidas para el índice BIC utilizando diferentes ajustes para la matriz de covarianza – fase 1. - En el eje horizontal se grafica el diferente número de grupos que se probaron para realizar el agrupamiento, el eje vertical muestra el valor obtenido por el índice BIC: Un mayor valor equivale a una penalización mayor de k en el modelo. Los índices obtenidos sobre diferentes números de grupos k detecta dos estructuras con una menor medida del índice BIC, sugiriendo una agrupación más compacta que los demás.

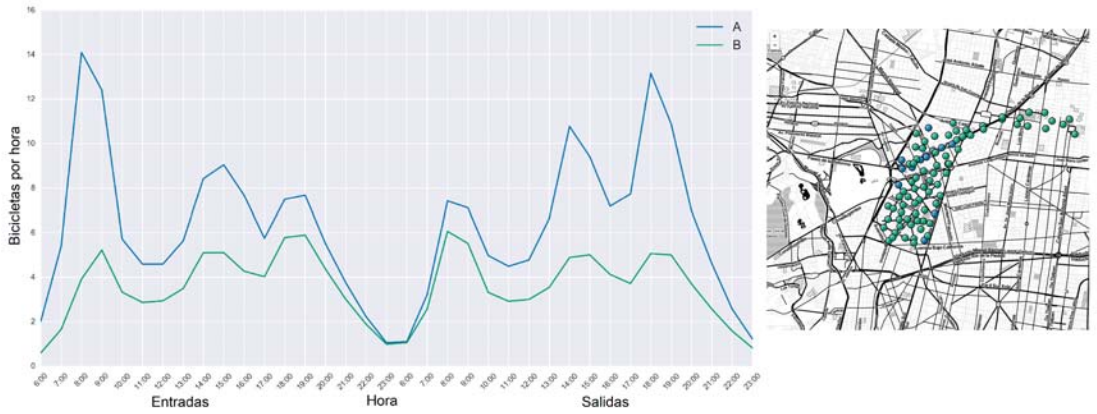

Figura 3.8: Comportamiento para los centroides de cada grupo – fase 1. - Izquierda: Cada centroide representa el comportamiento descriptivo para representar a las estaciones que lo integran. El eje vertical detalla el conteo de bicicletas por hora. Derecha: Estaciones esparcidas en la ciudad de acuerdo a su ubicación geográfica y coloreadas acorde a su grupo asignado.

Es interesante señalar que a pesar de que esta técnica no utilizó ningún atributo que hiciera referencia a su ubicación geográfica, sin embargo las estaciones parecen estar agrupadas en proximidad obedeciendo a la pertenencia de su agrupación.

3.7.3. Agrupación y medidas de evaluación – fase 2

Mientras nuevas fases incorporan más estaciones, otros comportamientos comienzan a ser notados por el criterio para seleccionar más que solo dos grupos para realizar el agrupación, el resultado del índice BIC es presentado en la fig.(3.9). Los comportamientos de los centroides de cada grupo se muestran en la parte izquierda de la fig.(3.10) y los resultados de las estaciones geográficamente ubicadas se presentan en la parte derecha.

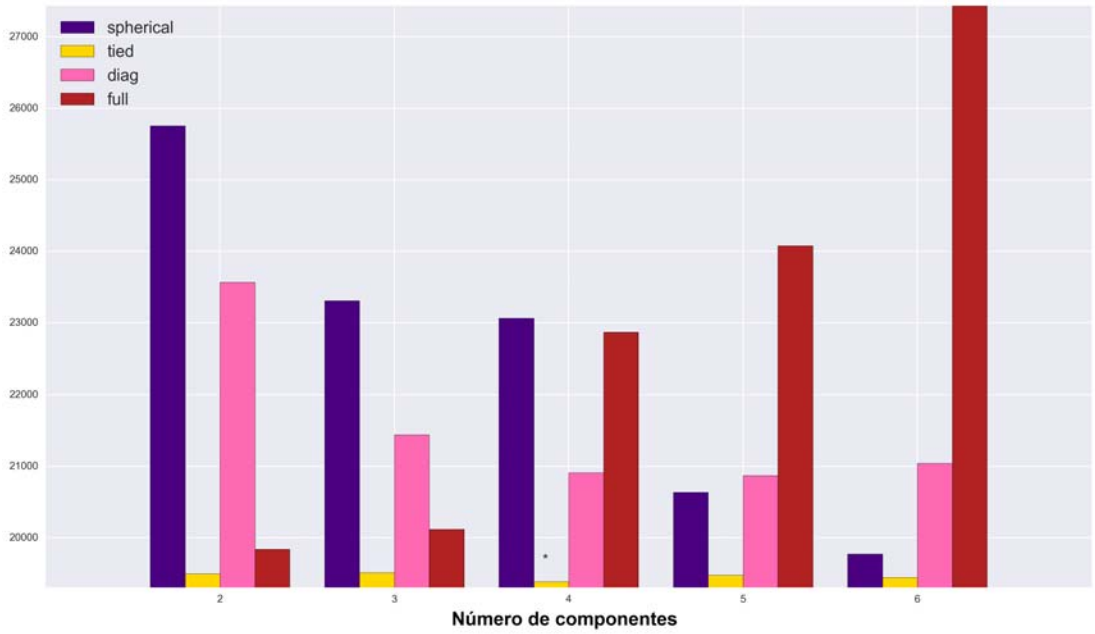

Figura 3.9: Medidas para el índice BIC utilizando diferentes deformaciones para la matriz de covarianza – fase 2. - Los índices obtenidos sobre diferentes números de grupos k muestran que 4 grupos se obtiene una menor medida del índice BIC, sugiriendo un agrupamiento más compacto que los demás.

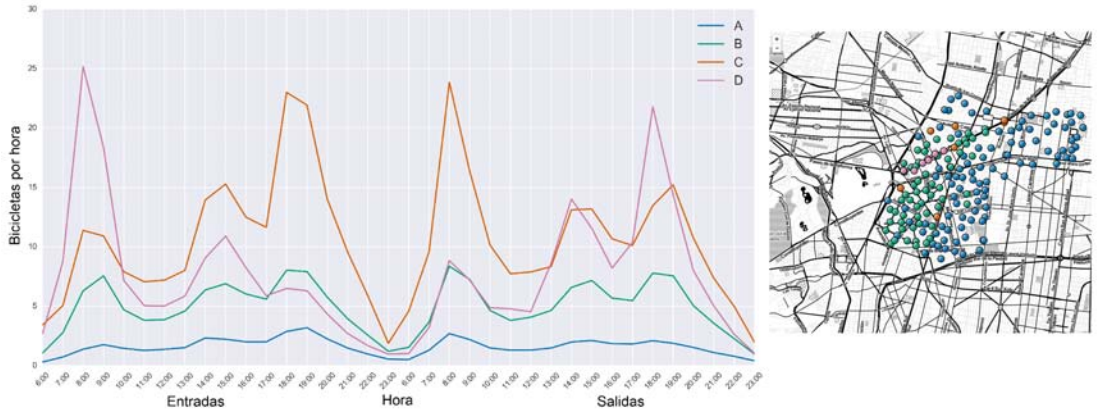

Figura 3.10: Comportamiento para los centroides de cada agrupación – fase 2. - Izquierda: El comportamiento entre el grupo C y el grupo D muestran un comportamiento espejo, donde los picos de entradas en D son los picos de salida en C mostrando estaciones fuente y estaciones destino para viajes de trabajo. Derecha: Las estaciones que se encuentran sobre reforma son cercanas a edificios corporativos.

Es de notarse que algunos de los perfiles de comportamiento para la estaciones está bien definido acorde al rol que juegan en la ciudad, como esas estaciones que se encuentran sobre la calle Reforma en el grupo D, cuyos picos de actividad de entradas en las mañanas (entre las 07:00 y 10:00 horas) y de salidas durante las noches (entre las 18:00 y las 20:00 horas) denotando el rol de una estación densamente utilizada por las personas que viajan diariamente de su hogar a su trabajo durante las mañanas y de regreso en las horas de la noche (commuters), su pico a medio día (entre la 12:00 y las 16:00 horas) puede ser causado por viajes realizados en horas del almuerzo. El segundo grupo obtenido C refleja exactamente el comportamiento opuesto, grandes picos en conteos de salida durante las mañanas y en los viajes de entrada durante las noches. Estas estaciones están localizadas cerca de puntos de transporte público, posiblemente utilizando las bicicletas del servicio para llegar más rápido a la estación de transporte público más conveniente para el suscriptor. El grupo B tiene un comportamiento promedio entre sus viajes entrada y salida, aunque es interesante que las estaciones pertenecientes a este grupo se encuentren adyacentes a estos dos primeros grupos, una posible explicación para el comportamiento de este grupo puede acuñarse a que estas estaciones pueden servir de reemplazo cuando las estaciones más populares se llenan o se vacían debido a las densas horas de tráfico en horas pico. Para el último grupo A, estas estaciones en la periferia de la ciudad están principalmente ubicadas en áreas residenciales, apuntando que este grupo de estaciones no se encuentran sobre áreas de tráfico denso y que solo son utilizados para viajes casuales.

3.7.4. Agrupación y medidas de evaluación – fase 3

En la incorporación de la fase 3 el agrupamiento resultante son mostrados en la fig.(3.11) y fig.(3.12), donde el número indicado de grupos y el comportamiento representativo de las estaciones pertenecientes a cada grupo son mostrados en la parte izquierda mientras que en la derecha se muestran todas las estaciones activas coloreadas dentro de su respectivo grupo hasta la inclusión de la fase 3. Cerca de la delegación La Roma los comportamientos de estas estaciones cerca del transporte público pertenecen al mismo grupo con un comportamiento similar en la ciudad. Las estaciones ubicadas sobre Paseo de la Reforma siguen siendo densamente utilizadas durante este periodo, mostrando que este comportamiento en la ciudad está bien establecido. Aunque, para este caso, el criterio BIC solo marcó una cantidad $k = 2$ grupos, siendo menor a la obtenida en su fase anterior. Es difícil imaginar que tras la inclusión de nuevas fases en el sistema el comportamiento de las estaciones se simplifique, por lo que se continuó realizando un análisis más en profundidad sobre lo que podría estar ocasionando este comportamiento.

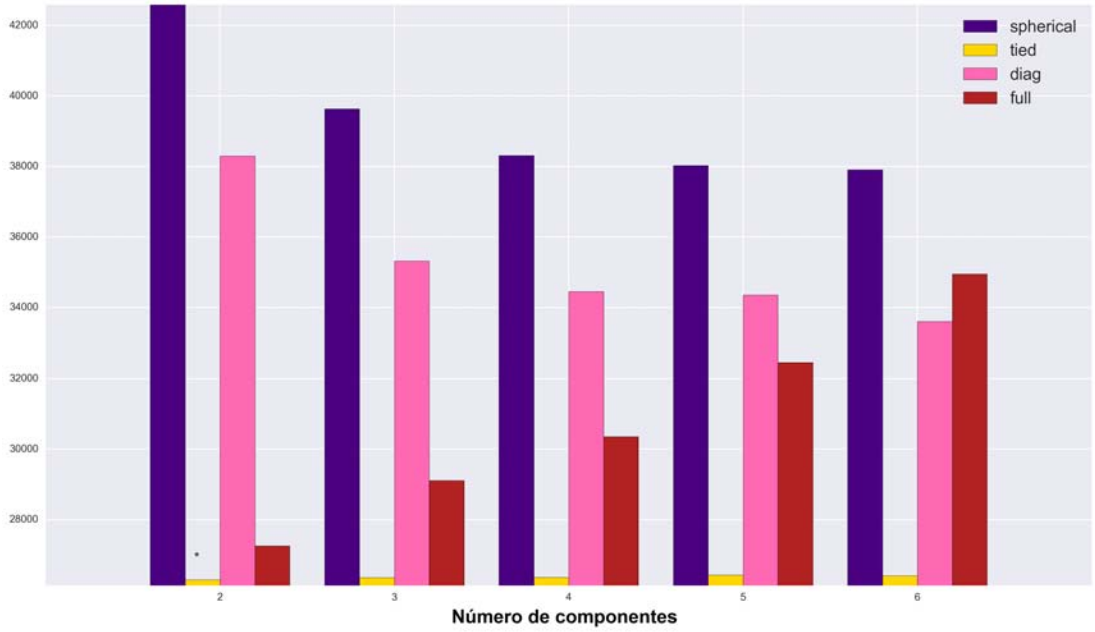

Figura 3.11: Índice BIC utilizando diferentes deformaciones para la matriz de covarianza – fase 3.

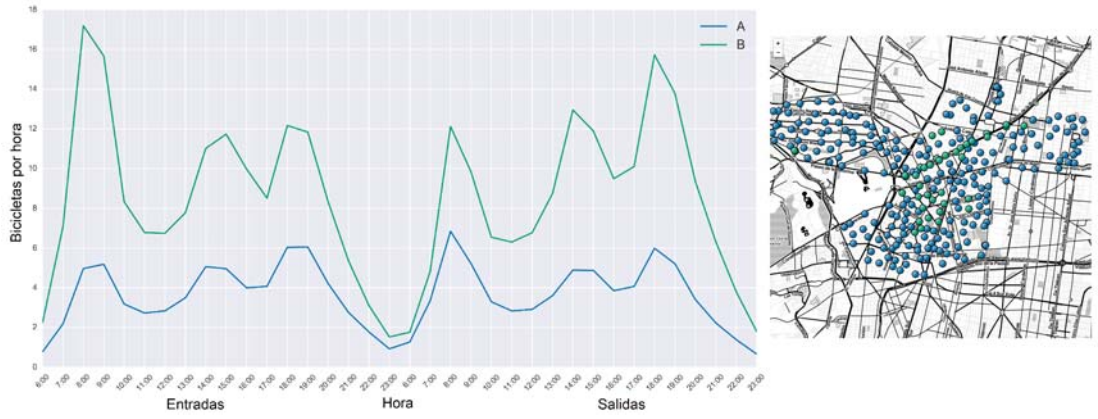

Figura 3.12: Comportamiento para los centroides de cada grupo y mapa de la ciudad con las estaciones esparcidas, coloreadas con respecto a su grupo – fase 3.

3.7.5. Reducción de dimensionalidad aplicando Análisis en Componentes Principales (PCA)

Aún con los resultados anteriormente mostrados, el realizar diferentes corridas del proceso de agrupación sobre los mismos datos puede llevar a obtener diferentes resultados en el número de grupos y el cambio en la pertenencia de una estación hacia otros, lo que sugiere una agrupación altamente inestable y no muy bien definida. Una matriz de correlación en los conteos de las estaciones por hora mostrada en la fig.(3.13) se puede observar una alta correlación entre las horas de entradas durante las mañanas (09:00 a 12:00 horas) y los conteos de salidas durante las tardes (11:00 a 17:00 horas) que sugieren una relación de incremento lineal donde si los conteos aumentan sobre las entradas en horas tempranas lo mismo se puede esperar de las salidas en las horas de la tarde. También puede apreciarse que las regiones de color blanco corresponden a las horas pico mostradas en los conteos de las estaciones, las cuales se podrían interpretar como regiones que ofrecen un comportamiento característico y podrían ser utilizadas para la creación de perfiles en las estaciones. Las características colineales presentes sugieren un problema debido a que un modelo de mezclas tiene un mal desempeño frente a estas observaciones. Para deshacerse de dicha colinealidad se tuvo que aplicar un método para la eliminación de correlación entre los componentes.

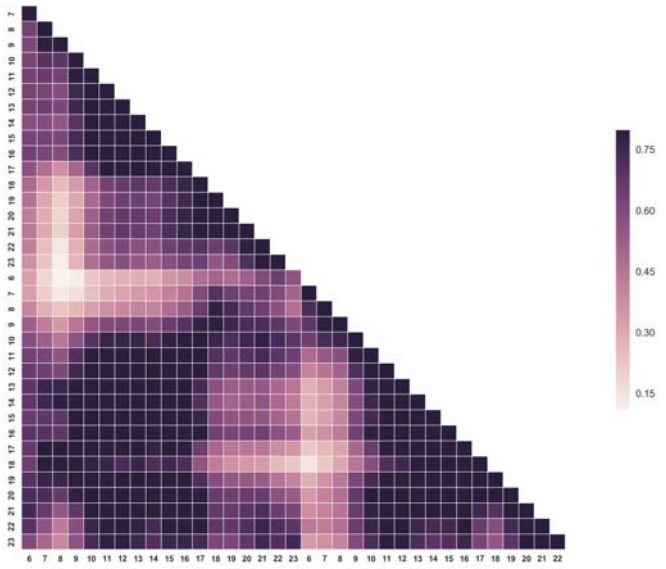

Figura 3.13: Matriz de correlación diagonal de las estaciones de Ecobici. - La intensidad en el color de este diagrama denota una mayor correlación entre conteos en diferentes horas.

El algoritmo de Análisis en Componente Principales (Principal Component Analysis o PCA por sus siglas en inglés) es una técnica para proyectar el espacio de características descrito por los datos decomponiéndolos en componentes principales (referidos aquí como PC). Cada PC es seleccionado en una dirección ortogonal que maximiza la varianza lineal. Cada PC está conformado por un par de elementos conocidos como eigenpares (vectores propios y valores propios) que aportan un grado de explicación en la varianza contenida dentro del espacio original (Apéndice A – Reducción de dimensionalidad con PCA).

PCA realiza una transformación en el espacio regresando el mismo número de PCs que de dimensiones original aunque no todos aportan el mismo grado de explicación de la variabilidad que existe entre los diferentes descriptores. En la figura(3.14) se muestran los PCs que explican hasta un 99 % la variabilidad del conjunto de datos, en esta figura los eigenpares no han sido ordenados y puede apreciarce que las entradas a las 06:00 horas son explicadas en mayor grado por el componente principal 5, 7 y 13 con una relación lineal positiva, por mencionar un ejemplo.

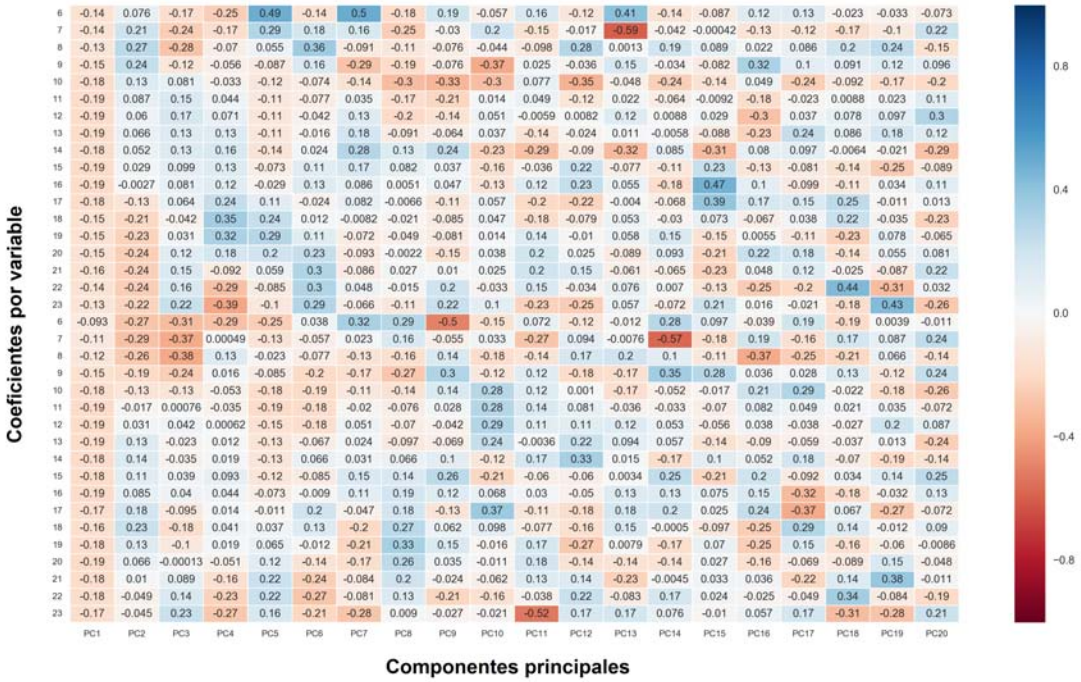

Figura 3.14: Matriz de pesos los cuales si son multiplicados por cada variable original estandarizada se obtiene su variación, mostrando el grado de correlación que existe del espacio de características por cada componente principal.

Debido a que PCA selecciona los PCs basado en la varianza lineal máxima interprentándose como los eigenpares que contienen los valores propios de mayor valor son aquellos explican en mayor grado la variabilidad en el espacio observado. Aquellos PCs con el mayor valor son seleccionados con la intención de describir una vasta mayoría de los datos y reducir la dimensionalidad para evitar la comparación y contraste con cada una de las características, conservando los PCs de mayor importancia. Realizando PCA sobre los conteos de las estaciones en Ecobici se obtuvo un sub espacio conformado por los cuatro componentes principales que explican el 95 % de la varianza en el comportamiento de las estaciones [Borgnat et al.]. En la fig.(3.15) se muestra la ganancia de varianza acumulativa obtenida por los cuatro PCs con mayor valor en sus valores propios (PC1 - PC4), explicando un 95 % de la varianza total.

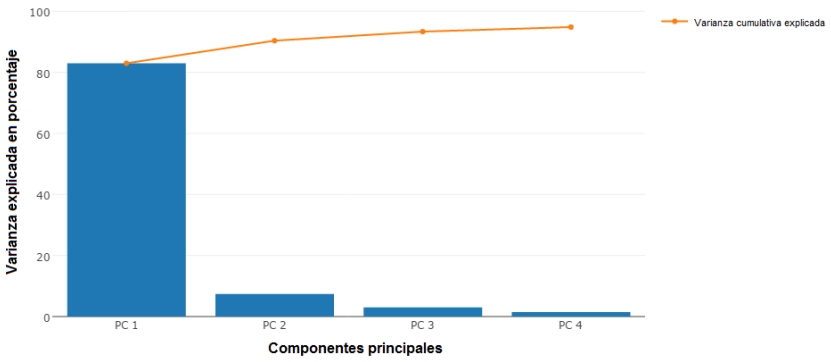

Figura 3.15: Varianza explicada por los cuatro primeros PCs. - Los cuatro componentes principales con mayor valor en sus valores propios son suficientes para explicar el 95 % de la varianza.

Tras obtener el nuevo sub espacio producto de la multiplicación del conjunto de datos original X por el sub espacio de características conformado con los cuatro PCs que explican el mayor porcentaje de la variabilidad W se obtiene un nuevo Y que representaría la proyección del espacio X en el sub espacio W.

3.7.6. Agrupación y medidas de evaluación para la fase 1 – PCA

Después de realizado el PCA sobre v_s una mayor estabilidad durante el proceso de agrupación fue obtenida, obteniendo la misma cantidad de grupos regresada por el índice BIC en cada corrida. Con anterioridad se mostró que con un sub espacio de cuatro PCs es suficiente para describir el 95 % de la variación eliminando la colinealidad en sus componentes y al mismo tiempo reducir la dimensionalidad del espacio, dando como resultando un espacio de proyección Y. A manera de exploración de datos se realiza un gráfico de dispersión contra todos los componentes principales, con el objetivo de observar si estructuras interesantes son capturadas en un mapeo de dos dimensiones de este nuevo espacio mostrado en la fig.(3.16). Cada punto representa la actividad de una estación en el espacio. Todo parece indicar que la gran mayoría de las estaciones tienden a agruparse en una determinada región del sub espacio, otros puntos parecen alejarse de esta región, ubicándose en zonas alejadas de la concentración de puntos en el espacio.

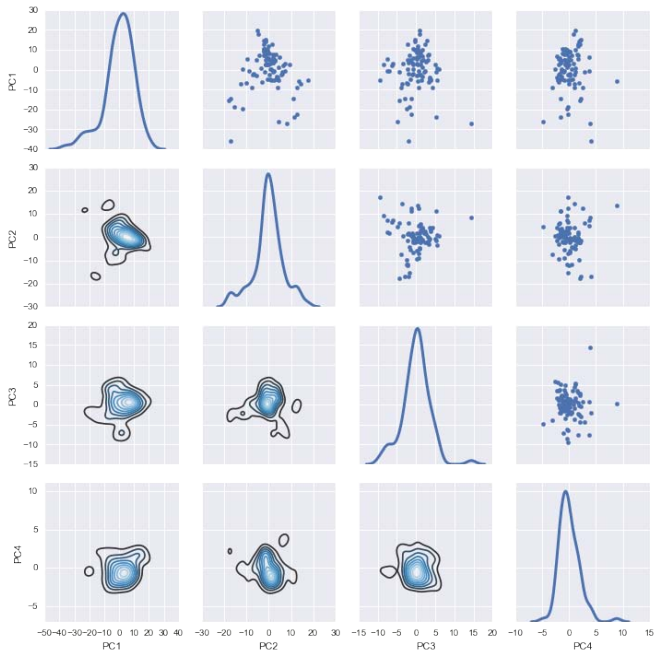

Figura 3.16: Gráfico de dispersión del nuevo espacio de características que consta de cuatro PCs de mayor importancia. - La matriz triangular superior muestra un gráfico de dispersión en dos dimensiones para los diferentes componentes principales, la matriz triangular inferior muestra las posibles distribuciones que se forman con la concentración de esos puntos.

Para la fase 1, ahora son obtenidos 3 grupos. Nuevamente esas estaciones con alto uso sobre la calle Reforma forman un grupo A por sí mismas mostrada en la fig.(3.17).

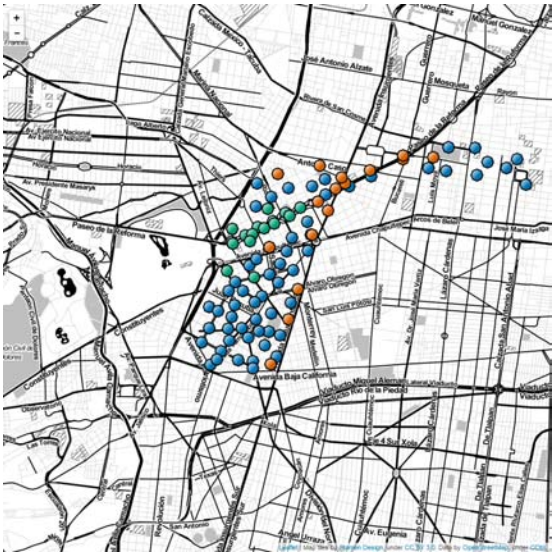

Figura 3.17: Estaciones dispersas por la ciudad y coloreadas por su pertenencia de grupo después de PCA – fase 1.

El segundo grupo B de estaciones está situado cerca de puntos de transporte público, el tercer grupo C puede indicar las áreas residenciales.

3.7.7. Agrupación y medidas de evaluación para la fase 2 – PCA

Para la fase 2, cinco grupos fueron obtenidos donde dos nuevos comportamientos son observados ilustrados en la fig.(3.18), la mayoría de las estaciones que conforman estos nuevos grupos están situados donde la fase 2 fue incorporada, también el comportamiento para el PCA de estos dos grupos es muy similar, difiriendo solamente en la magnitud.

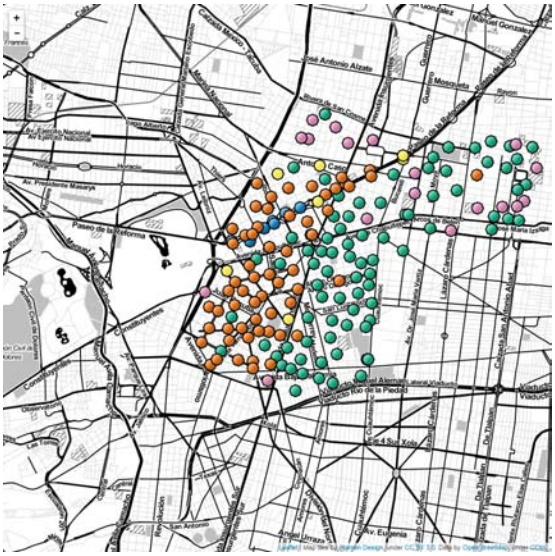

Figura 3.18: Estaciones dispersas por la ciudad y coloreadas por su pertenencia de grupo después de PCA – fase 2.

3.7.8. Agrupación y medidas de evaluación para la fase 3 – PCA

La fase 3 con PCA en la fig.(3.19) revela cinco grupos significativos: los elementos en D cuyas estaciones cuentan con un alto uso en entradas y salidas, localizadas en monumentos de la ciudad, puntos corporativos y puntos clave de transporte público. Los grupos A y B están situados en áreas residenciales con mucho menor uso.

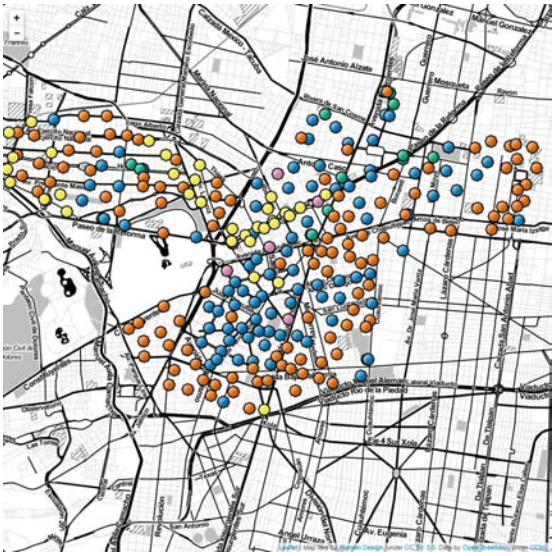

Figura 3.19: Estaciones dispersas por la ciudad y coloreadas por su pertenencia de grupo después de PCA – fase 3.

3.8. Resumen

Este capítulo consistió en la aplicación de técnicas de visualización de los datos que se manejan en este trabajo, primero realizando una serie de transformaciones sobre los registros en los cuales el enfoque es centrado alrededor de los conteos de las estaciones del sistema. También, con las justificaciones antes mencionadas, se realizó sobre los datos diferentes procedimientos eliminar actividad en los registros que en este estudio fueron considerados como información que podría contaminar los conteos realizados, eliminando todos aquellos conteos que fueran considerados como atípicos. El uso de herramientas visuales para una captura de diferentes aspectos en la actividad y observación de las tendencias que este servicio suele manifestar. También se obtuvo un estimado de las fechas en inclusión de las 3 diferentes fases empleadas en el sistema durante todo el periodo de tiempo que comprenden los registros de actividad.

También se realizó una agrupación sobre las estaciones en función de su similitud de comportamiento dado un vector de características formado por los conteos de entradas y salidas a lo largo de las horas y realizando una agrupación para todas las estaciones activas tomando en cuenta los diferentes periodos que comprenden la agregación de las diferentes fases hasta donde se tiene fecha en los registros. Las agrupaciones formadas ofrecen una explicación sobre su rol en los hábitos de los suscriptores dentro de la ciudad que con un poco de conocimiento sobre las estructuras que se encuentran en sus alrededores se puede construir una historia que cuente sobre día regular y sus actividades. Durante este análisis se encontró evidencia de la existencia de colinealidad entre los

conteos de horas utilizadas para realizar la agrupación, eliminando dicha característica haciendo uso de PCA y seleccionando un sub espacio construido a partir de aquellos componentes principales que ofrecieran una mayor explicación sobre la variación en las características. Realizando las agrupaciones nuevamente se observó un mejor comportamiento sobre las estaciones, resultando en grupos más estables y en una prueba BIC constante en la selección del número más adecuado de grupos qué utilizar. La agrupación con PCA muestra estaciones alejadas de lo que se podría interpretar como la región un uso general conformada por la mayoría de las estaciones que conforman el sistema, estos puntos sugieren que provienen de una distribución diferente.

Capítulo 4

Diversidad de rango (rank diversity)

En sistemas que se componen por múltiples elementos es común medir cuáles son los de mayor valor. Se conoce como ordenamiento (o ranking) cuando se ordenan, ya sea de manera ascendente o descendente, estos elementos cuantificando un aspecto observable. Realizar un ordenamiento a menudo es importante para categorizar los componentes más y menos sobresalientes debido a que estos pueden manejar gran cantidad de los recursos disponibles dentro del sistema. Si se realizan ordenamientos durante suficientes intervalos de tiempo puede llegar a notarse una diversidad en los diferentes elementos que ocupan un rango determinado durante el ordenamiento.

El estudio de diversidad de rango (Rank diversity en inglés) es una técnica donde obtenemos una medida de la variabilidad en el sistema formado por elementos diferentes que aparecieron en los diferentes rangos, siendo dependiente en el tiempo en lugar de ser una distribución estática.

La distribución de diversidad de rango obtenida se propone ser separada en 3 regiones diferentes de acuerda a la variabilidad en sus rangos: cabeza, cuerpo y cola (Apéndice A - Definición de las regiones de la cabeza, cuerpo y cola utilizando diversidad de rango), donde se asume que estas regiones contienen rangos con un diferente grado de diversidad entre sí.

Por esta razón diversidad de rango ha sido utilizada con anterioridad como una medida para determinar la probabilidad que tiene un elemento de cambiar su posición en el rango dentro de un sistema, por ejemplo, en el análisis sobre vocabularios se puede obtener la probabilidad que tiene una palabra de cambiar su frecuencia de uso sobre el tiempo en la literatura usando un conjunto de millones de palabras en seis idiomas Indo-Europeos [Cocho et al., Morales et al.], revelando que las palabras que ocupan lugares correspondientes a los rangos más altos en un sistema tienden también a ser los que tienen menor probabilidad a decaer en jerarquía en el rango con el tiempo (También interpretado como caer en desuso[Cocho et al.]). Este fenómeno también es observado para otros sistemas donde el desempeño de sus componentes puede ser medido, como en los deportes que muestran que los equipos con un mejor desempeño tienden a conservar sus posiciones de una manera más definida con respecto a los demás participantes, un ordenamiento generado por las ciudades clasificadas por su complejidad económica y

las 500 empresas que líderes ordenadas por la revista 'fortune'.

Como en los sistemas antes mencionados, las estaciones pueden ser agregadas, removidas o modificadas, de la misma la representación de un sistema parecido puede llevarse a cabo haciendo uso de BSS. Este trabajo hace uso de diversidad de rango para establecer una caracterización en un BSS haciendo uso de las regiones de cabeza, cuerpo y cola para corroborar aseveraciones establecidas. Para realizar un ordenamiento se utilizan las estaciones de Ecobici como lo elementos del sistema y se establece como medida para realizar el ordenamiento los conteos de entradas/salidas. Aquellas estaciones que aparecen en altos rangos en un periodo seleccionado que tienden a ocupar ese lugar sin cambiar mostrando estabilidad en su comportamiento deberían estar en una posición más confiable para realizar estimaciones.

4.1. Construcción de diversidad de rango para Ecobici

Para obtener la diversidad de rango, primeramente, se crea una matriz X usando Ec. (3.1) para los periodos comprendidos de la fase 1 y la unión de la fase 2 y 3; se consideró que un intervalo de tiempo $\Delta = 24$ horas apropiado para realizar el ordenamiento de las estaciones basándonos en el total de conteos obtenidos durante ese periodo. Si definimos matriz de rangos con frecuencia R a partir de X (matriz (4.1)) donde k_d es una secuencia resultante de ordenar las estaciones x en función de sus conteos de actividad (preservando duplicados) durante cada día d. Para el periodo considerado, R con dimensiones de D x S donde D corresponde al número total de días del periodo y S al número total de rangos en el sistema, donde k_{d1} es el ID de la estación con el mayor rango de frecuencia y k_{ds} es el ID de la estación con el menor rango de frecuencia en un día fijo d y donde cada columna puede ser interpretada como un rango k_s.

$$R = \begin{bmatrix} k_{11} & \geqslant & k_{12} & \geqslant & \cdots & \geqslant & k_{1s} \\ \vdots & & \vdots & & \ddots & & \vdots \\ k_{d1} & \geqslant & k_{d2} & \geqslant & \cdots & \geqslant & k_{ds} \end{bmatrix} \tag{4.1}$$

Después, para volver la diversidad de rango un subconjunto de elementos distintos sobre las columnas, se define un vector $\theta(k_s)$ creado a partir de la matriz R (fórmula (4.2)), donde θ es una función que regresa el conjunto de estaciones únicas q que se encuentran presentes en el rango k. Después se extrae el número de elementos diferentes obtenido por θ y se divide entre D. Esta división se realiza con la intención de normalizar la diversidad en el rango k_s (fórmula (4.3)) con respecto al periodo D utilizado (para obtener un resultado entre 0 y 1 D debe de ser igual o mayor al número de estaciones utilizadas).

$$\theta\left(k\right) = \left\{k_i\right\}_{i \in \{1 \ldots, n\}}, \text{donde } k = [k_1, \ldots, k_n] \tag{4.2}$$

$$d\left(k_s\right) = \frac{1}{D}\left|\theta\left(k_s\right)\right|, \forall q \in S \tag{4.3}$$

Se obtuvo la diversidad de rango para las fases 1, 2 y 3. Para este caso los periodos de tiempo para las fases 2 y 3 fueron unidas debido a que ambas se encuentran muy cercanas una de la otra en la fecha de incorporación al sistema que se obtuvo con el criterio antes mencionado. La fig.(4.1) muestra la estructura del sistema obtenida con la diversidad de rango correspondiente para la fase 1 donde cada barra representa el número de estaciones únicas que aparecen en ese rango k dividido entre el total de días D utilizado para ese periodo, la forma de la distribución observada en la figura presenta un comportamiento que puede ser aproximado mediante una función parabólica. Un incremento en el valor del rango puede ser interpretado como una ganancia en incertidumbre resultando en valores muy altos en solo unos cuantos rangos, diciéndonos la naturaleza caótica y poco predecible del sistema.

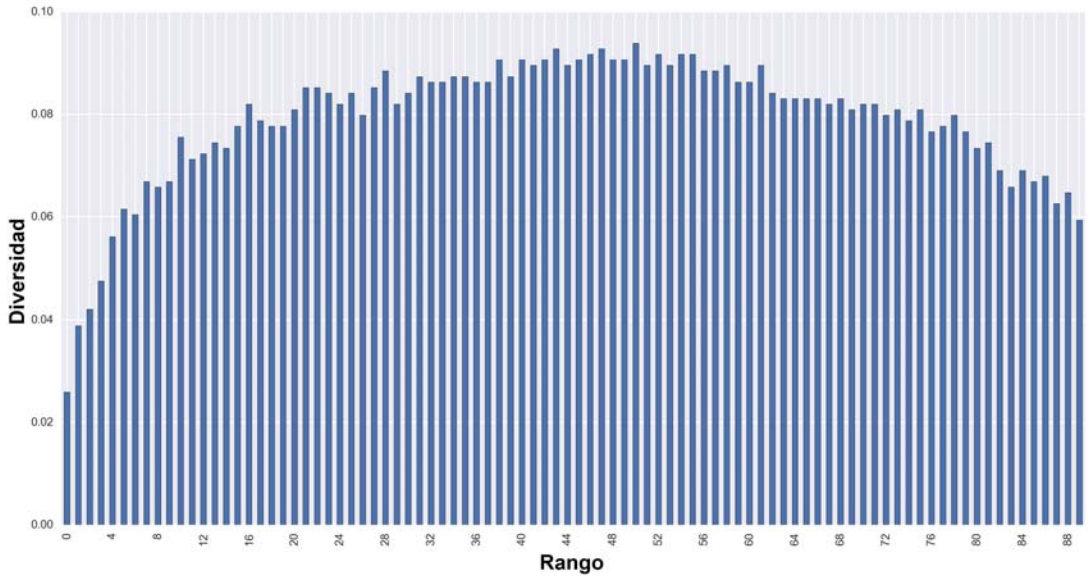

Figura 4.1: Diversidad de rango durante la fase 1. - Los valores $d(k)$ para cada rango.

4.2. Frecuencia de rango

Si tomamos en cuenta como la diversidad de rango es obtenida, algunos aspectos pueden perderse, como el número de veces que una estación aparece en los diferentes rangos. En la fig.(4.2) se muestra un enfoque diferente para ver la variabilidad en el sistema donde se expone la frecuencia que tiene una estación en cada rango, ordenadas

de manera ascendente: De manera complementaria a la diversidad de rango que muestra cuantas estaciones han estado presentes en cada rango, la frecuencia en el rango indica el número de diferentes rangos en los que ha estado presente cada estación. Este enfoque permite ver la variabilidad que las estaciones tienen en los rangos, complementando las observaciones sobre los elementos estables en el sistema. Las estaciones 1, 27, 64 y 88 tienen los valores más bajos observados, con lo que se ha dicho anteriormente esto puede interpretarse en que estas estaciones tienen una baja presencia relativa sobre los diferentes rangos comparadas con el resto de las demás estaciones, cuyos valores se ubican por alrededor de una frecuencia de 75 rangos distintos, casi el doble de frecuencia que estas primeras cuatro. Entre menor sea la frecuencia de aparición para las estaciones sobre los rangos, mayor es la confianza de que su uso en el sistema sea estable.

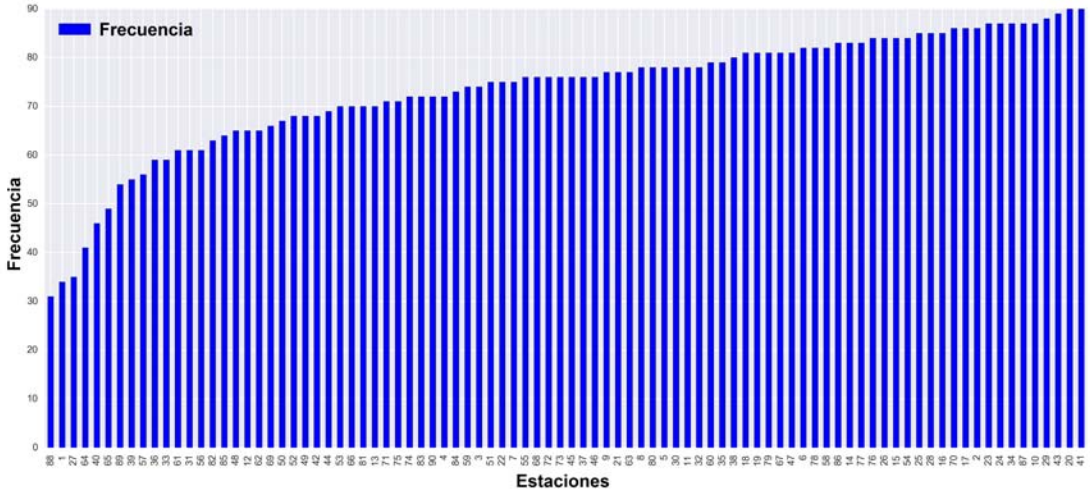

Figura 4.2: Frecuencia de presencia para las estaciones para cada rango para la fase 1. - Cada barra representa una estación, el valor está dado por cuantos rangos diferentes apareció dicha estación.

La segunda parte de la expansión de Ecobici supone bastantes efectos importantes sobre el sistema, haciendo posibles más viajes a otras localidades, desplazando algunas estaciones de la región de la cabeza hacia el cuerpo de la frecuencia de rango e incrementando el número de suscriptores y bicicletas, resultando en mayores conteos por día. La fig.(4.3) y la fig.(4.4) muestran la diversidad de rango y las primeras 100 estaciones con menor frecuencia sobre los rangos respectivamente. Para los periodos que comprenden las fases 1 y la 2-3 muestran la misma forma parabólica, apuntando a una muy diversa variación de estaciones sobre cada rango y una considerable ganancia en variabilidad adquirida en medida de que se salta a un rango mayor a uno menor. Con lo expuesto anteriormente es evidente que un modelo dependiente de datos históricos que se encarga e realizar predicciones de ordenamiento sobre las estaciones de Ecobici obtendrá un desempeño sustancialmente superior en aquellas estaciones más consistentes dentro del

sistema.

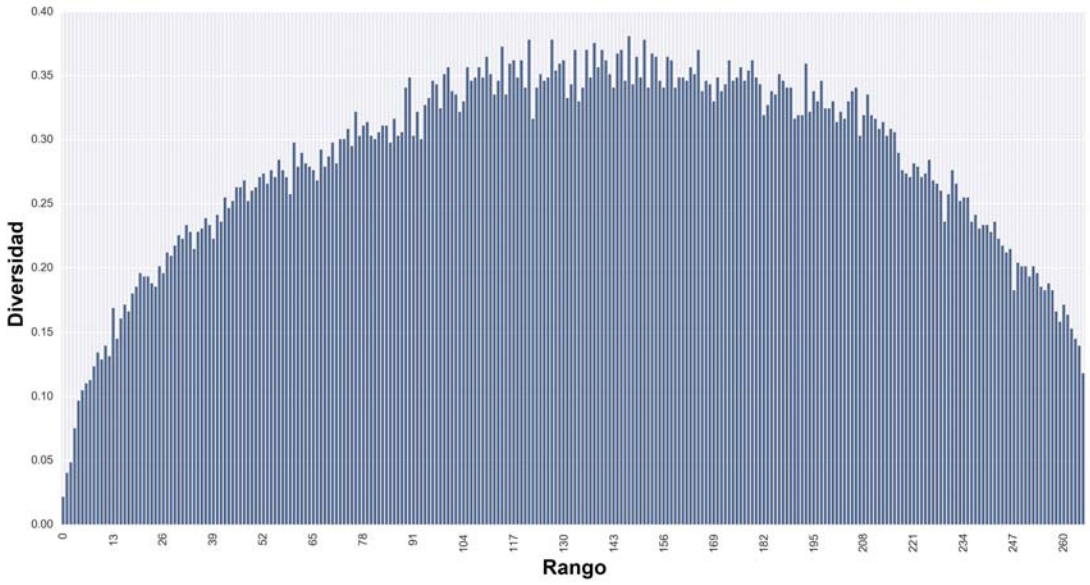

Figura 4.3: Diversidad de rango para el sistema durante la fase 2 y 3. - La aparición de diferentes estaciones en los rangos a lo largo del periodo de la fase 2 y 3.

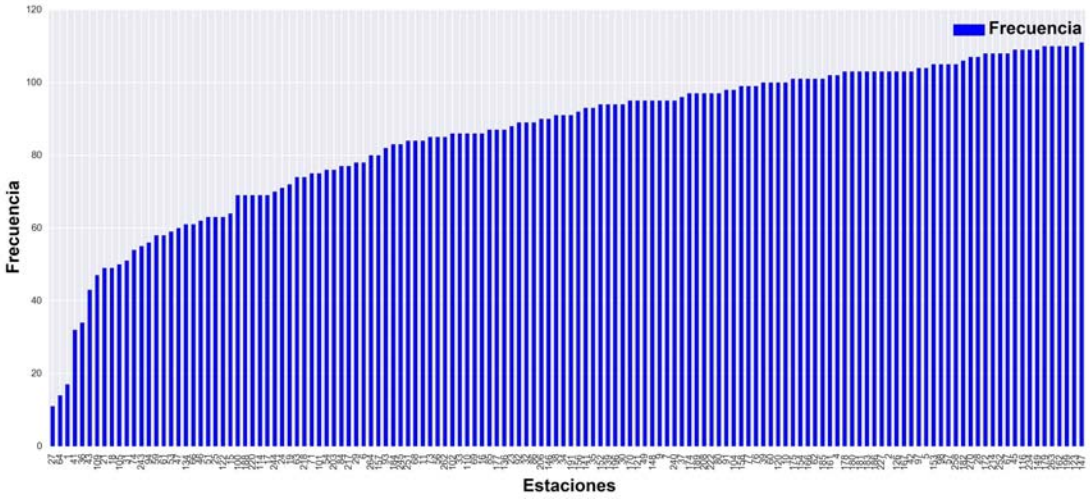

Figura 4.4: Frecuencia de presencia para las estaciones para cada rango para la fase 2 y 3. - Las primeras 100 estaciones con menor aparición sobre los rangos dentro del periodo comprendido.

No obstante puede ocurrir el caso que, por fenómenos ajenos al sistema, estaciones

que normalmente no son frecuentemente utilizadas hayan sido muy usadas durante un día en particular, igual puede surgir el caso contrario haciendo que algunas estaciones que tienen mucho uso sean desplazadas a rangos donde usualmente su lugar no corresponde, esto puede deberse a diferentes motivos en los cuales las estaciones tuvieron que ser cerradas temporalmente, orillando a los usuarios del sistema a buscar estaciones sustitutas. Con la finalidad de mitigar este fenómeno se determinó que los elementos que forman parte de cada región en la diversidad de rango deben ser mutuamente excluyentes al momento del muestreo. Los elementos que forman parte de la cola que aparecen en la región de la cabeza fueron removidos mientras que los elementos que forman parte de tanto la cabeza como de la cola que aparecen en la región del cuerpo fueron removidos de este último conjunto.

Esta aseveración establece que existe la posibilidad de que estaciones que no pertenecen a la región situada en su distribución correspondiente hayan aparecido en otras distribuciones, no obstante, en promedio el desempeño debería conservarse, como se verá en el capítulo 6.

4.3. Resumen

Se explicó acerca de diversidad de rango, de cómo esta herramienta puede medir un fenómeno emergente dentro de un largo espectro en los sistemas midiendo como la diversidad de las diferentes estaciones que conforman un rango varían así como se detalló sobre trabajos anteriores donde se ha empleado esta herramienta para capturar estructuras haciendo uso de una métrica para medir la característica de interés que, para este caso, es la actividad de entradas y salidas de las estaciones. También se definieron los términos y las fórmulas necesarias para construir una matriz de elementos ordenados con base en su actividad para obtener las diversas estaciones que se encuentran presentes en los rangos durante un periodo comprendido. Finalmente en este capítulo se introdujo otro ángulo de perspectiva diferente a la diversidad de rango: la frecuencia de rango, con la cual puede apreciarse la variabilidad vista desde las estaciones en lugar de los rangos y, aunque un ordenamiento es preferible para observar de una manera más agradable, puede percibirse de manera más rápida el aumento de frecuencia que ocurre en las estaciones y por lo consecuente su desplazamiento de región en la diversidad de rango.

Capítulo 5

Diseño experimental

5.1. Gradient Boosted Regression Trees

Los árboles de decisión basados en métodos de ensambles (ensamble methods en inglés) son técnicas de aprendizaje que consisten en partir el espacio formado por las características en sub regiones utilizando árboles de decisión, esto en conjunto de desarrollar múltiples modelos de una reducida complejidad para posteriormente combinar la predicción múltiple con el objetivo de reducir el sesgo (Apéndice A – Gradient Boosted Regression Trees). Una técnica derivada de los métodos de ensambles, conocida como Gradient Boosted Regression Trees (GBRT por sus siglas en inglés) ha sido la elección de preferencia sobre una variedad de técnicas de regresión existentes en vista de un desempeño satisfactorio para propósitos de ordenamiento donde se ha utilizado anteriormente para el la recuperación y ordenamiento de páginas web de mayor relevancia en motores de búsqueda web [P et al., Chen and K.], además estas técnicas han probado ser robustas ante problemas presentes en otras técnicas de regresión, tales como uso de variables categóricas y un mal desempeño de regresión ocasionado por la correlación existente en el vector de características.

Para este experimento la predicción es realizado tomando solamente la fase 1 de Ecobici y midiendo el error obtenido para las diferentes estaciones basados en la categoría que se asignó utilizando diversidad de rango.

5.2. Conjunto de prueba y entrenamiento

Para poder fundamentar la hipótesis aquí propuesta es necesario someter a prueba el modelo con un conjunto de datos para evaluar su desempeño de acuerdo a una métrica de error. Se realiza una separación para el conjunto de datos para comprobar que no ha sobre aprendido los datos y efectivamente corroborar que el modelo generado describa el fenómeno que se desea pronosticar fig.(5.1).

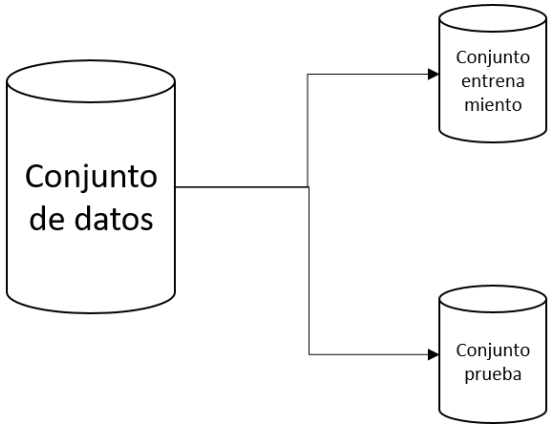

Figura 5.1: Conjunto de entrenamiento y prueba. - Se realiza una separación en el periodo total de datos en un conjunto de entrenamiento y otro de prueba construidos a partir del conjunto total disponible.

Para la obtención de los conjuntos de entrenamiento y prueba en este caso el conjunto de entrenamiento está conformado por 486 días, comprendidos desde el día 29 de septiembre del 2010 al 30 de septiembre del 2012 (tabla 5.1). El conjunto de datos de prueba está conformado desde el día 1ro de octubre al 29 de octubre el año 2012. Se omitieron los primeros ocho meses de actividad de Ecobici debido a la baja magnitud de actividad que presenta el sistema durante este periodo comparando con los conteos en las fechas finales del periodo. A partir de ahora cuando se haga referencia al conjunto de prueba y conjunto de entrenamiento se estará haciendo referencia periodos comprendidos dentro de estos dos conjuntos.

	Periodo inicial	Periodo final
Entrenamiento	29-09-2010	29-08-2012
Prueba	30-08-2012	01-09-2012

Tabla 5.1: Periodos de tiempo que comprendidos para el conjunto de prueba y el conjunto de entrenamiento

5.3. Definición de las regiones de cabeza, cuerpo y cola

Una vez obtenida la diversidad sobre cada rango, θ (fórmula (4.2)) es utilizada para obtener aquellos elementos que son más representativos de las regiones de la cabeza, cuerpo y cola para realizar comparaciones posteriores entre ellas (fig.(5.2)). Las

fig.[(4.1)-(4.4)] muestran que un alto nivel en variabilidad se puede obtener recorriendo solo unos pocos rangos, por esta razón y en orden de mantener los elementos que pertenecen a cada distribución lo más puros posibles solo se tomaron el primer y el último conjunto de elementos que forman parte de los rangos correspondientes a la cabeza y cola respectivamente mientras que el cuerpo fue tomado de un rango aleatoriamente seleccionado de los rangos centrales.

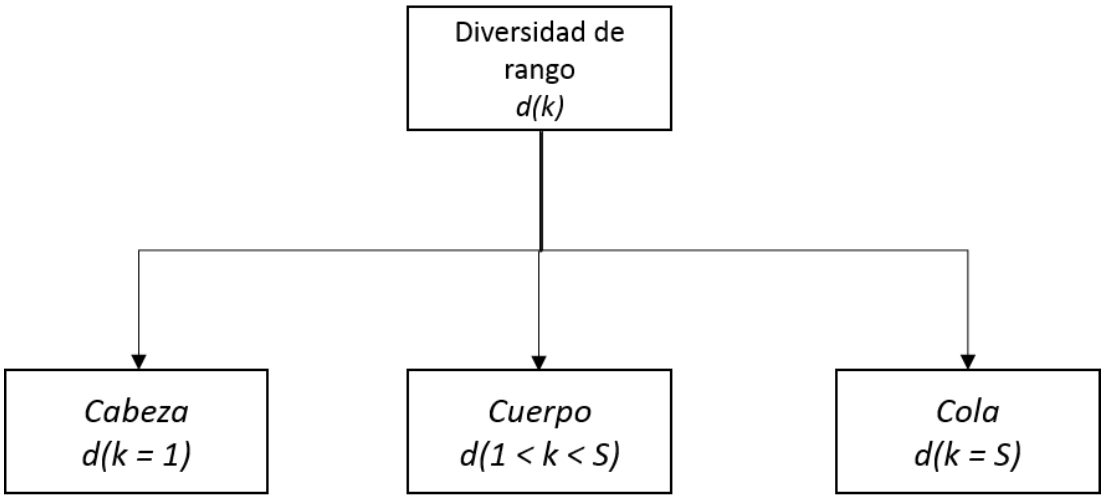

Figura 5.2: Selección de subconjuntos a partir de diversidad de rango. - Se seleccionan las d estaciones diversas que aparecen en el rango k.

5.4. Ajuste y predicción para GBRT

Una vez seleccionado el modelo y definidos conjuntos de prueba y entrenamiento se procede a explicar el proceso que se utilizó en este experimento para construir los pronósticos de ordenamiento para las estaciones. En orden de obtener un modelo que realice un ordenamiento para s estaciones dado un día d se optó por un modelo de regresión para ajustar y posteriormente predecir la actividad para cada una de las estaciones que forman el sistema. Se optó por un enfoque para predecir el rango y por predecir la actividad de entradas + salidas del sistema en sí debido a que las medida de evaluación sobre el rango serían menos susceptibles a errores ocasionados por la magnitud en los conteos, ya que al predecir el rango se está realizando una predicción sobre el lugar de la estación con respecto al sistema y no la actividad de la estación en sí, así la(s) estación(es) con más actividad pronosticada diferirán con respecto a su posición en el rango y no por el conteo en el flujo de bicicletas obtenido en d. Debido a que solo se dispone en este experimento de los registros de viajes realizados en las estaciones (trabajos similares a este han reportado tener información adicional en el día como: temperatura, precipitación fluvial, indicadores de días festivos, huelgas, etc. [Mellers

et al.]) la necesidad de crear variables para ajustar el modelo y obtener pronósticos fue necesaria. El periodo de tiempo para la acumulación en la actividad de entradas + salidas por estación que se considera adecuada para este caso sigue siendo $\Delta = 24$ horas, donde para cada día d se obtuvieron las siguientes variables independientes:

- Una media móvil construida a partir de diez observaciones en el pasado: En ocasiones las estaciones pueden sufrir averías o necesitar de mantenimiento por lo cual tengan que obligarse a ser cerradas temporalmente, también su uso puede ser afectado al existir algún evento social el cual sature la calle donde se encuentra la estación, volviendo inaccesible su acceso. El llevar una media del funcionamiento que ha tenido dicha estación en el pasado ayuda a tener una variable más robusta ante vicisitudes que afecten su desempeño en el uso habitual.

- Una variable ordinal que represente el día de la semana en el que se encuentra, bajo el argumento de que existen días que influyen en una mayor magnitud en el funcionamiento para ciertas estaciones en el sistema, cambiando así el orden en el uso de las estaciones.

La variable dependiente que se consideró para cada estación fue el conteo de actividad que la estación s tuvo durante el día d, esta cifra también es importante ya que con ella se realiza el posterior ordenamiento sobre la actividad de las estaciones. Estas variables fueron construidas tanto para el conjunto de entrenamiento como el de prueba. Posteriormente estos vectores de características fueron dados a ajustar al modelo GBRT con los parámetros antes mencionados, obteniendo un modelo que pronostica la actividad de una estación en base a su actividad los últimos 10 días y el día de la semana en el que se encuentra (nótese que el modelo aquí planteado desconoce de la estación que se trate dentro del sistema). Para cada vector de características se tiene el rango observado que dicha estación obtuvo en realidad en d.

De esta forma el modelo GBRT se ajusta con el conjunto de entrenamiento para después realizar la predicción sobre el conjunto no observado de prueba, después de realizada la predicción, para cada día d se realiza el rankeo en base a la actividad pronosticada por el modelo para todas las estaciones de ese día y después se mide la diferencia del rankeo pronosticado contra el observado dada una medida de error seleccionada. Se utilizaron dos medidas de error diferentes para evaluar la precisión del ranking: La media del error cuadrático (Root Mean Squared Error o RMSE en inglés) y ganancia acumulativa descontada normalizada (Normalized Discounted Cumulative Gain o NDCG en inglés), el uso de estas medidas es usual en la evaluación de algoritmos de ordenamiento, aunque NDCG (Apéndice A – Normalized Discounted Cumulative Gain) es una medida de evaluación que se ha demostrado más apta que ofrece un error que es más relevante con los resultados para este tipo de tareas [Kalervo and Jaana].

La medida de desempeño se realizó para los tres subconjuntos de estaciones basados en sus categorías: cabeza, cuerpo y cola obteniendo la media de desempeño de cada subconjunto que forman parte de cada categoría en el conjunto de prueba. Se realizó un sub muestreo para los elementos de cada categoría de la siguiente manera:

1. Se toman de manera aleatoria una sub población para cada una de las categorías: cabeza, cuerpo y cola, cada una consiste de diez elementos. Estas estaciones seleccionadas representa el error calculado para esas diez estaciones representa en promedio, el error respecto a la predicción del rango para su categoría.

2. Se realiza la predicción en la actividad de las estaciones para un día d dado, posteriormente las estaciones son ordenadas en función de su actividad de mayor a menor, asignando una etiqueta a cada estación según su magnitud en actividad (la estación s con mayor actividad obtiene la etiqueta 1 y la estación con menor actividad recibe la etiqueta de S, donde S es el número de estaciones activas en el sistema durante el periodo comprendido).

3. Se calcula el MSE o NDCG (según sea el caso) tomando en cuenta solo las estaciones que fueron seleccionadas en el sub muestreo para esa iteración y se realiza un promedio para cada d en el conjunto de prueba obteniendo 3 errores representativos que corresponden a cada categoría.

4. Se realizan los pasos 1-3 para 100 iteraciones, obteniendo un error promedio en cada una de las tres categorías que se desea comparar.

5.4.1. Evaluación y prueba de hipótesis con RMSE

Aplicando la serie de pasos antes mencionados, los resultados de aplicar MSE son expuestos en esta sección. En la fig.(5.3) es mostrado un diagrama de cajas (box-plot en inglés) de los errores obtenidos para las tres categorías diferentes utilizando MSE en el conjunto de entrenamiento. Se puede apreciar que la mitad de los errores obtenidos para las tres categorías se encuentran entre los valores de 8 y 10, siendo la región que corresponde a la cabeza la que se encuentra un poco más inclinada hacia tener un menor error. Por otro lado, la varianza para cada una de las categorías es bastante grande, el cuerpo la categoría que tiene menor varianza de las tres.

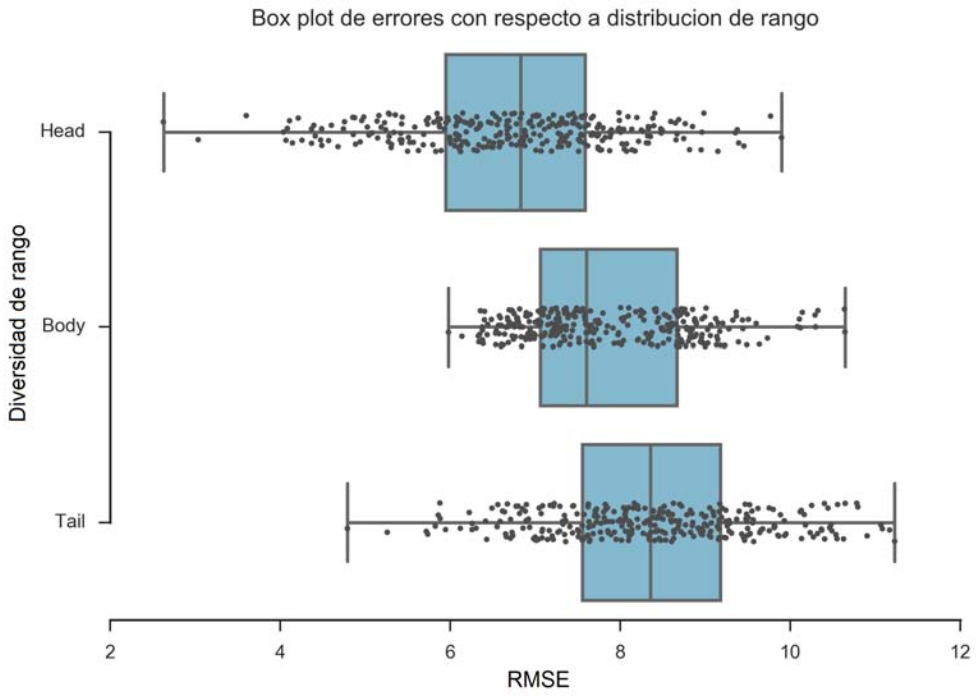

Figura 5.3: Diagrama de caja para cada una de las categorías – conjunto de entrenamiento. - Las medias MSE obtenidas para las 100 iteraciones realizadas en cada categoría de Rank diversity en el conjunto de entrenamiento.

Para el conjunto de prueba se obtuvieron los resultados mostrados en la fig.(5.4) donde se conservan los mismos comentarios realizados con el conjunto de entrenamiento a excepción de que la varianza aumenta para el cuerpo. De los datos que se obtuvo la raíz del error cuadrático las escalas que se manejan son los errores promedios en rankeo en la diferencia del rango pronosticado respecto al rango observado que en realidad las estaciones obtuvieron en d. Los resultados obtenidos fueron validados usando una prueba estadística.

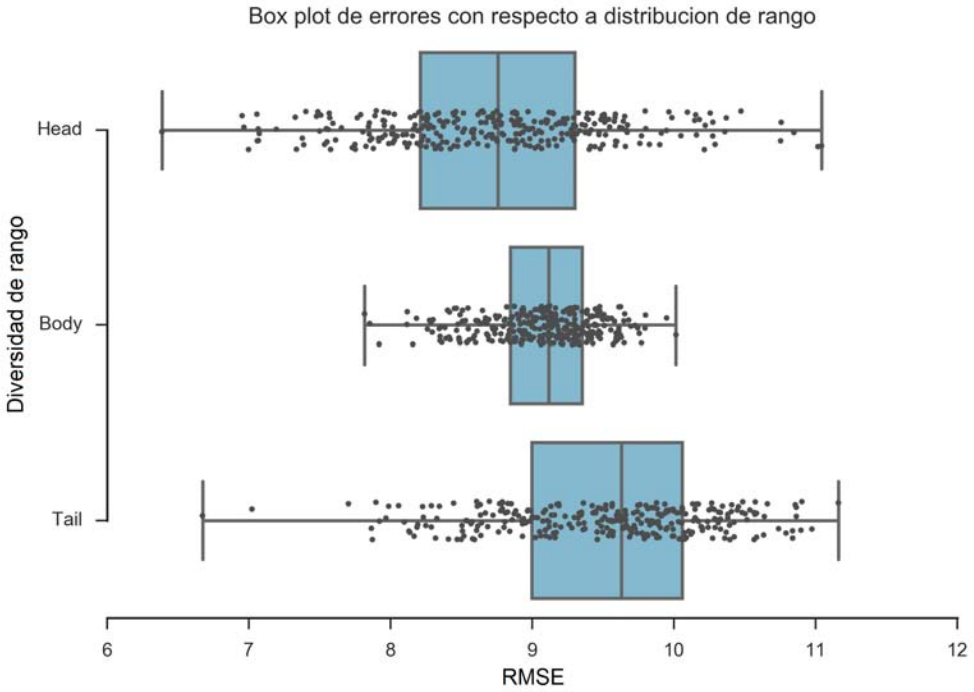

Figura 5.4: Diagrama de caja para cada una de las categorías – conjunto de prueba. - Las medias de MSE obtenidas para las 100 iteraciones realizadas en cada categoría de diversidad de rango en el conjunto de prueba.

La prueba ANOVA es un conjunto de herramientas estadísticas cuyo objetivo es rechazar o no la hipótesis nula que establece que los grupos de tratamiento que son obtenidos de una población tienen la misma media. ANOVA produce una estadística F (F-statistic en inglés) que es la razón entre la varianza que fueron calculadas entre los grupos entre la varianza que se produce dentro de los grupos (Apéndice A - ANalysis Of VAriance (ANOVA)). Si las medias obtenidas entre los diferentes grupos de la población tienen los mismos valores, la varianza entre el grupo de las medias debería de ser menor. Siguiendo lo antes dicho, un valor más alto de esta razón implica que las muestras en las poblaciones provienen de poblaciones con medias distintas.

Para este caso donde nos interesa rechazar la hipótesis nula que involucra la tasa en el error obtenido tomando en cuenta la categoría a la que las estaciones pertenecen según la diversidad de rango se realizó la prueba estadística One-way ANOVA tomando en consideración este factor. Para el caso computacional ofrece dos valores para realizar conjeturas:

- El F-test, donde un valor acercado a 1 indica que no hay evidencia suficiente que indique que alguno de los grupos efectivamente tiene una media distinta que

proceda de otra población.

- P-values, que indican la probabilidad de que los resultados que están siendo observados sean producto de una casualidad. En medida de que el P-value disminuye también lo hace la probabilidad de que dicha observación sea producto de la aleatoriedad.

Los resultados obtenidos por el One-way ANOVA se muestran en la tabla 5.2 pertenecen al f-test y al p-value obtenidos de analizar dicha prueba para el factor de la cabeza, cuerpo y cola y conocer si alguna de estas sub poblaciones tienen una distribución diferente.

f-test	90.12
p-value	2.15e-36

Tabla 5.2: F-test y p-values resultantes para el conjunto de entrenamiento

En el valor de f-test se obtuvo un valor cercano a 90.12, una cantidad por mucho mayor a la necesaria para rechazar la hipótesis nula mientras que el p-value se obtuvo un valor muy cercano a cero lo cual indica que existe un porcentaje de probabilidad muy baja de que los resultados obtenidos sean productos de mera casualidad para ser tomada en cuenta. Ambos valores revelan en conjunto de que existe una fuerte evidencia de que efectivamente hay separabilidad en las medias de los errores de las poblaciones tomando en cuenta su pertenencia en la distribución en el rango.

Una vez obtenidos resultados que efectivamente ofrecen separabilidad entre las medias de las sub poblaciones lo siguiente que se realiza es una prueba de seguimiento (follow through-test en inglés) con el objetivo de reafirmar los resultados y encontrar aquellas sub poblaciones que tienen esa separabilidad en sus medias con un intervalo de confianza estadísticamente significativo. El follow through test seleccionado es conocido como la prueba de Tukey (Tukey's test en inglés), una prueba comparativa o prueba estadística de múltiples comparaciones (Apéndice A – Tukey's test) con el objetivo de encontrar aquellas observaciones cuyas medias difieran significativamente la una de la otra. La tabla muestra las comparaciones realizadas con las medias y los intervalos de confianza de un 95 % obtenidos del Tukey's test. La hipótesis nula se ha rechazado para los tres casos considerados (Tabla 5.3) lo cual indica que existe una separabilidad significativa entre las tres diferentes sub poblaciones.

Donde cada columna en la tabla explica un factor diferente:

- group 1 es el primer grupo de tratamiento comparado.
- group 2 es el segundo grupo con el que se realiza la comparación.
- Meandiff ofrece la diferencia entre las medias observadas.
- lower es el punto terminal menor en el intervalo de confianza.

group 1	group 2	Meandiff	lower	upper	reject
Cuerpo	Cabeza	-0.3153	-0.4413	-0.1892	True
Cuerpo	Cola	0.4038	0.2777	0.5298	True
Cabeza	Cola	0.7191	0.593	0.8451	True

Tabla 5.3: Tabla de resumen con los resultados del Tukey's test sobre el conjunto de entrenamiento

- upper es el punto terminal mayor en el intervalo de confianza.
- reject es un valor booleano que informa si se ha rechazado o no la hipótesis nula sostenida con el Tukey's test

Debido a que la diferencia entre las restas en las medias son significativamente diferentes tomando en cuenta un intervalo de confianza del 95 % (las comparaciones cuyos intervalos no pasan sobre cero) podemos decir que la hipótesis nula es rechazada para los tres conjuntos. La fig.(5.5) muestra las medias y los intervalos de confianza de un 95 % de todas las medidas de error realizadas en las comparaciones múltiples de todos los pares. La línea roja vertical hace notar el extremo superior que corresponde a la categoría de estaciones situadas en la región de head. La imagen nos permite apreciar dos puntos importantes:

- Al categorizar las estaciones por su pertenencia en la diversidad de rango existe una separación estadísticamente significativa entre los errores obtenidos que se obtuvieron de cada conjunto diferente.
- El error obtenido para el conjunto de estaciones que forman parte de la cabeza se centra en una medida menor con respecto a aquellos errores obtenidos por los conjuntos del cuerpo y cabeza, aportando mayor evidencia a que efectivamente existe una inclinación hacia un mejor pronóstico de rango realizado sobre el subconjunto de estaciones que mantienen un alto grado de actividad en el sistema.

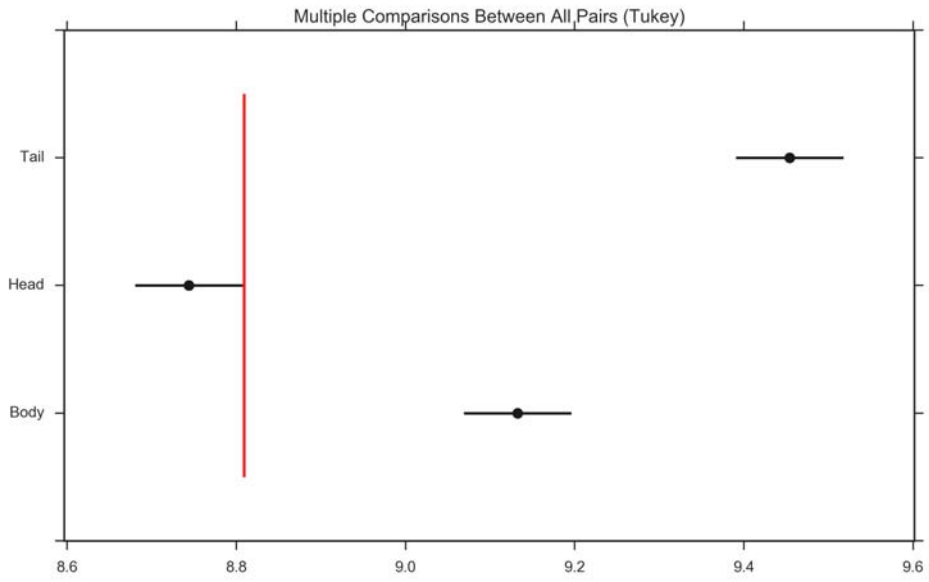

Figura 5.5: Distribución de errores usando MSE. - Medias e intervalos de confianza de las diferentes poblaciones obtenidos realizando el Tukey's test sobre los errores obtenidos utilizando MSE en el conjunto de entrenamiento.

Una vez obtenidos los resultados para el conjunto de entrenamiento se procede a realizar lo mismo utilizando el conjunto de prueba que fue deparado para corroborar la efectividad en los pronósticos realizados. Se realiza la predicción en la actividad seguida de su rankeo sobre el día d para el conjunto de datos que no ha sido observado por el modelo. Los resultados obtenidos de realizar el One-way ANOVA sobre los errores medidos con la métrica MSE son mostrados en la tabla 5.4 a continuación:

f-test	141.13
p-value	5.11e-54

Tabla 5.4: Resultados obtenidos al aplicar One-way ANOVA sobre el conjunto de datos de prueba.

Igualmente los resultados apuntan a que existe evidencia significativa de que las medias entre una o más de las poblaciones no son iguales así que se procede a realizar de nueva cuenta un Tukey's test para realizar la comparativa entre poblaciones y corroborar los resultados anteriores (Tabla 5.5).

De nueva cuenta la media en los errores para cada categoría es diferente, rechazando así la hipótesis nula establecida por el Tukey's test. El gráfico de las medias junto

group 1	group 2	Meandiff	lower	upper	reject
Cuerpo	Cabeza	-1.1708	-1.3999	-0.9417	True
Cuerpo	Cola	0.4086	0.1795	0.6377	True
Cabeza	Cola	1.5794	1.3503	1.8085	True

Tabla 5.5: Tabla de resumen con los resultados del Tukey's test sobre el conjunto prueba.

con los intervalos de confianza para cada población mostrado en la fig.(5.6) muestra nuevamente los puntos expuestos en el conjunto de prueba. Esta vez con los errores obtenidos en un grado menor que con los de entrenamiento. Para esto último pueden existir diferentes explicaciones, como por ejemplo que el conjunto de entrenamiento está formado por un amplio periodo donde lentamente se experimentó un aumento en la actividad del sistema, donde hubo un incremento de usuarios que seguramente trajo como consecuencia a un aumento en viajes para algunas estaciones, ubicando al modelo de regresión ajustado dentro de un lento periodo de transición.

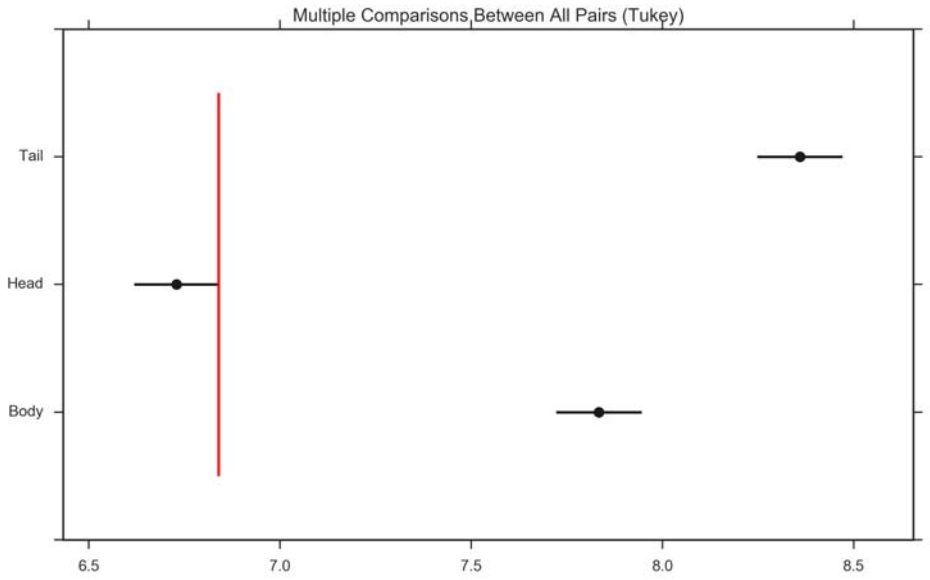

Figura 5.6: Errores obtenidos utilizando MSE en el conjunto de prueba. - Medias e intervalos de confianza de las diferentes poblaciones obtenidos realizando el Tukey's test

5.4.2. Evaluación y prueba de hipótesis con NDCG

Las evaluaciones se realizaron aplicando MSE que es una medida sencilla y rápida para medir el desempeño obtenido por el algoritmo de rankeo de las estaciones, pero existen mejores medidas de para este tipo de propósitos, como lo es NDCG, donde además de medir la calidad del ordenamiento para todas las estaciones participantes se da una mayor ponderación a aquellas estaciones que poseen un mayor grado de actividad, siendo consideradas las estaciones con más relevancia para el sistema, donde aquí relevancia se interpreta como el grado de actividad que mantiene una estación en relación al sistema.

De la misma forma que se realizó con MSE, un One-way ANOVA fue realizado sobre las medidas obtenidas de NDGC para rechazar la hipótesis nula y así fundamentar que existe una diferencia en la predictibilidad del orden de la actividad en las estaciones sobre el sistema con una medida de error distinta. Cabe hacer notar un ligero cambio que se hizo para tomar en cuenta esta evaluación: se invirtieron las etiquetas de las estaciones, ahora de manera ascendente, donde la estación con mayor actividad en el sistema obtiene la etiqueta S, donde S es el número de estaciones activas en el sistema durante ese periodo mientras aquella estación con menor actividad en el sistema obtiene la etiqueta de 1, esto se realizó debido a que a diferencia de la medida MSE donde un valor menor obtenido por esta cifra representa una mejor evaluación, lo que NDCG intenta realizar es maximizar su índice de evaluación, ponderando con las etiquetas de mayor valor a las estaciones más relevantes. Se realizan las mismas pruebas estadísticas que se utilizaron para medir el error en el apartado anterior, esta vez para NDCG y comprobar que existe separabilidad entre las medias de las poblaciones. La fig.(5.7) muestra los errores promedios obtenidos en el conjunto de entrenamiento realizando nuevamente una separación categórica sobre las estaciones según su posición en el rango utilizando NDCG, la fig.(5.8) muestra los errores obtenidos para el conjunto de prueba.

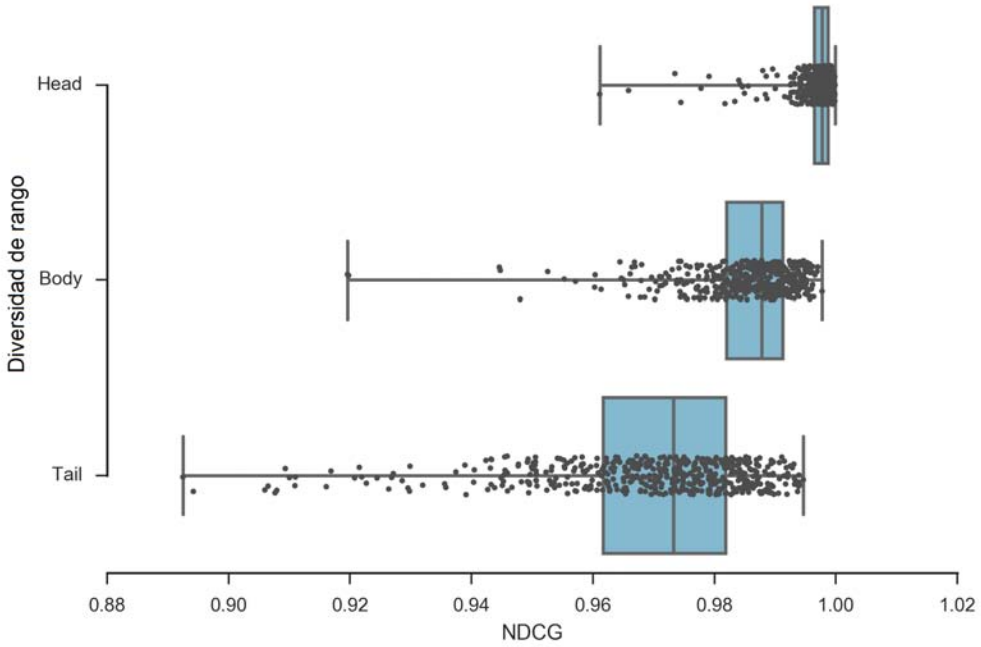

Figura 5.7: Gráfico de caja de los errores usando NDCG conjunto de entrenamiento. - Los valores más cercanos a 1 se encuentran más cercanos al orden ideal.

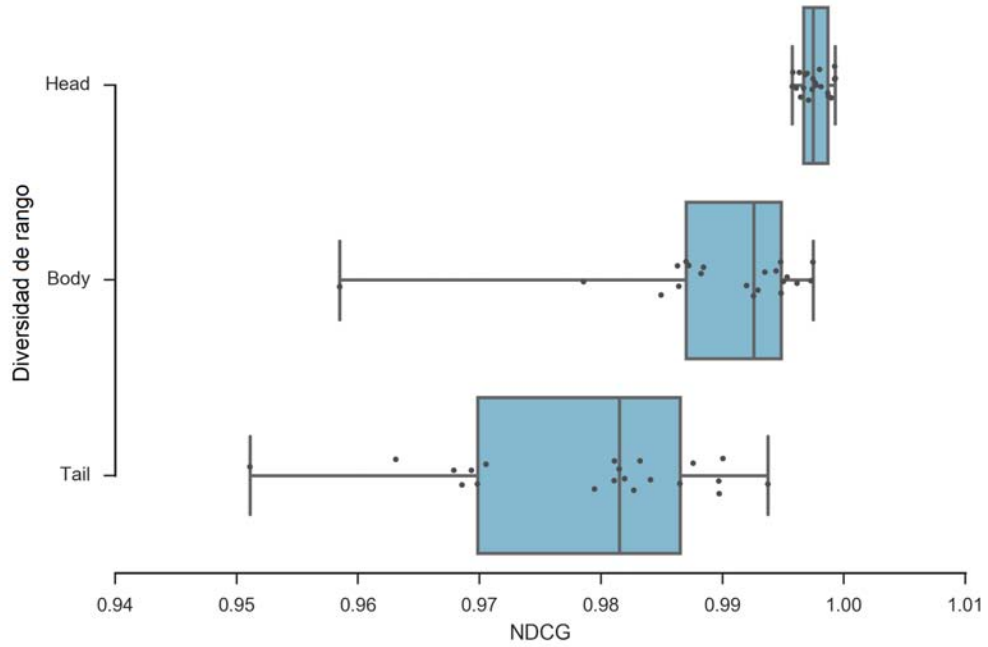

Figura 5.8: Gráfico de caja de los errores usando NDCG conjunto de prueba.

Donde las distribuciones para los conjuntos en entrenamiento y prueba reflejan una separabilidad evidente, esta medida también parece reflejar una particularidad: una menor variabilidad en el error obtenido de las estaciones que forman parte de la cabeza donde los errores en esta sección son mucho menores que el de los demás. Una vez obtenidos los errores se realizan las pruebas estadísticas correspondientes en orden de obtener diferencias significativas sobre las medias de cada sub grupo. La fig.(5.9) se muestran los resultados obtenidos con las pruebas estadística Tukey's test y la tabla 5.6 muestra los resultados obtenidos de la prueba ANOVA, ambos resultados sugieren una separabilidad entre sus categorías.

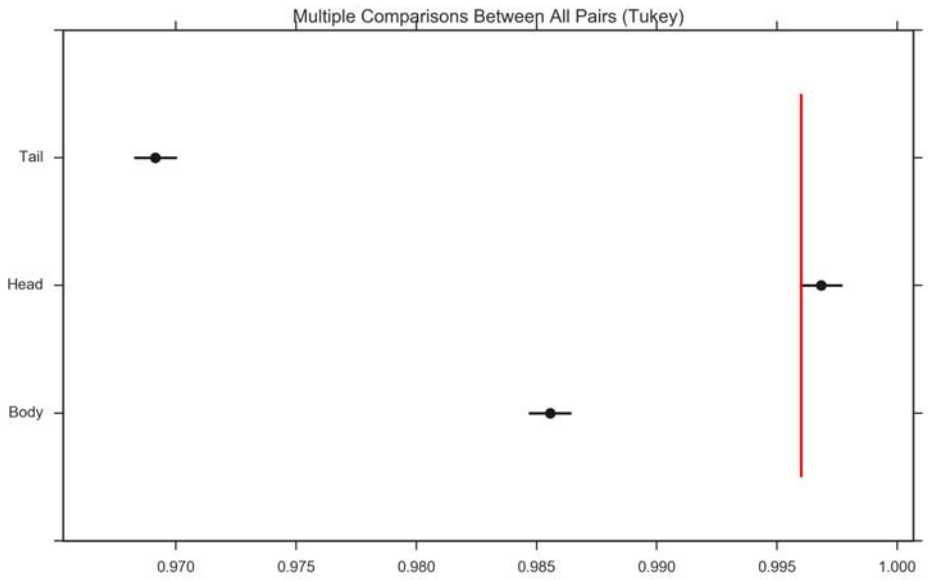

Figura 5.9: Tukey's test sobre los errores obtenidos utilizando NDCG en el conjunto de entrenamiento. - Medias e intervalos de confianza de las diferentes poblaciones obtenidos.

f-test	29.58
p-value	1.14e-09

Tabla 5.6: Resultados obtenidos al aplicar One-way ANOVA sobre el conjunto de datos de entrenamiento con NDCG.

Finalmente, en la fig.(5.10) se muestran los resultados del Tukey's test para el conjunto de prueba, la tabla 5.7 también muestra resultados significativos para dicho conjunto. Estos resultados obtenidos anteriormente muestran la separabilidad entre los errores obtenidos en los subconjuntos de las estaciones y por lo tanto también se sugiere una diferencia en el grado de predictibilidad para las secciones de las regiones de la cabeza, cuerpo y cola.

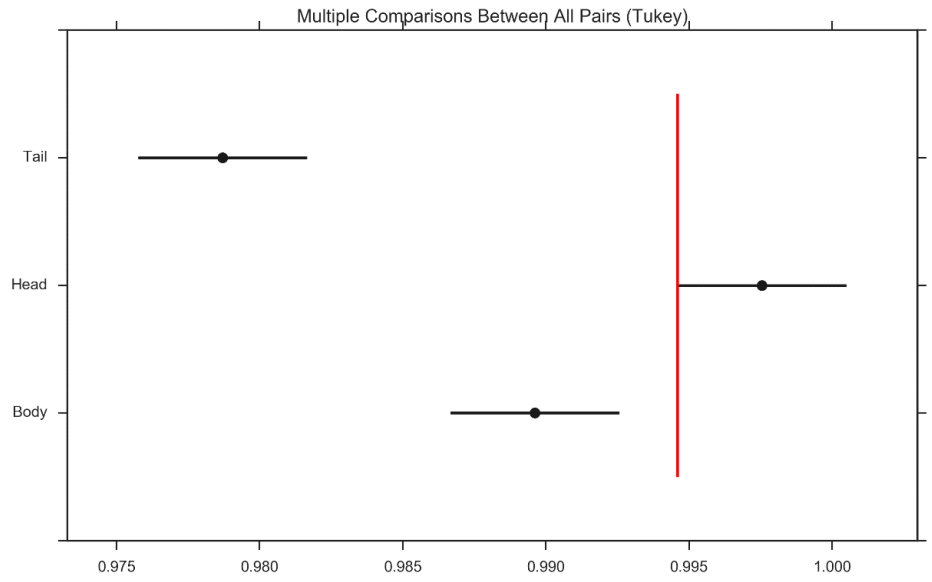

Figura 5.10: Tukey's test sobre los errores obtenidos utilizando NDCG en el conjunto de prueba. - Medias e intervalos de confianza de las diferentes poblaciones obtenidos.

f-test	677.07
p-value	1.39e-208

Tabla 5.7: Resultados obtenidos al aplicar One-way ANOVA sobre el conjunto de datos de prueba con NDCG.

5.5. Resumen

En este capítulo se puso en aplicación la unión de las herramientas necesarias para someter a prueba la hipótesis. Se realizó la separación del conjunto de datos para la fase 1 en dos: un conjunto de entrenamiento para ajustar el modelo y un conjunto de prueba para someter el modelo a evaluación ante nuevas observaciones no vistas anteriormente. Después, con ayuda del modelo de regresión GBRT ajustado y evaluado y realizando la predicción de actividad esperada sobre un día d se realizó el rankeo de las estaciones para cada día contenido en ambos conjuntos, para después obtener sus

respectivos errores de evaluación utilizando las predicciones en el rango obtenido de dichas estaciones sobre el sistema para un d dado utilizando las métricas de error MSE y NDCG.

La evaluación se realizó con los rangos obtenidos de las estaciones y no sobre la actividad de las mismas con el propósito de tomar como enfoque la predictibilidad de los rangos en el sistema, que es un punto de vista más acorde con lo que explica la diversidad de rango y dejando de lado la precisión en la cantidad de los conteos. Después de obtenidos los errores de los rangos pronosticados de cada estación contra los rangos observados para cada d, lo siguiente fue realizar una prueba estadística con la intención de comprobar que las medias de los errores obtenidos para cada subconjunto provienen de poblaciones distintas tomando en cuenta su pertenencia en la región de diversidad de rango. Al obtener evidencia significativa de que efectivamente existe una separabilidad en las medias de los errores, rechazando la hipótesis nula de ANOVA, se realiza en seguida una prueba de seguimiento conocida como Tukey's test para encontrar aquellas medias entre las poblaciones que tengan una separabilidad estadísticamente significativa entre sí.

El análisis sobre los resultados obtenidos de estas pruebas estadísticas sugiere que realizar una separación por las categorías obtenidas por la diversidad de rango efectivamente influye en la calidad de predicción sobre el orden de uso que las estaciones obtendrán en el sistema, siendo las estaciones con etiquetas correspondientes a la distribución de la cabeza aquellas que obtienen medidas de error menores con respecto a las demás y corroborando que, a pesar de que la distribución observada en la diversidad de rango sobre el sistema de Ecobici tiene una ganancia en varianza pronunciada sobre la mayoría de los rangos diferentes, aún puede decirse que sus propiedades son apreciables en este sistema.

Capítulo 6

Conclusiones y discusión

6.1. Resumen

Utilizando herramientas de análisis en los conteos generados de las estaciones y sometiéndolos a diversas transformaciones se logró obtener un entendimiento más extenso sobre las dinámicas de los viajes generados por los usuarios de Ecobici teniendo en cuenta aspectos existentes dentro de la ciudad.

Los agrupamientos de actividad para las estaciones realizados con mixture models ofrecen comportamientos característicos de viajes distribuidos en diferentes regiones, generando perfiles para el diferente tipo de servicios que ofrece cada estación a lo largo del día. El agrupamiento en conteos por hora que se realizó permite ver las horas de más alta demanda en las estaciones y en base en esas observaciones es posible contar una historia sobre cómo los suscriptores utilizan estas estaciones como intermediarias para desplazarse de las zonas residenciales a sus trabajos y viceversa. También mediante este análisis puede obtenerse información de el tipo de actividad que surge en áreas nuevas donde implementan expansiones del servicio, encontrando similitudes en diferentes regiones y quizá una manera alternativa para determinar el valor de la tierra de zonas urbanas.

Realizando las transformaciones necesarias en los conteos de las estaciones se optó por un sistema de ordenamiento para los rangos que considerara la magnitud en los flujos de entradas + salidas de bicicletas en las estaciones pudiendo hacer uso de la técnica de diversidad de rango para observar el comportamiento del sistema desde esta perspectiva y se pudo llegar a la conclusión de que, efectivamente, dicho comportamiento está presente en Ecobici y que éste puede cambiar en medida se introducen modificaciones en el sistema, como ha pasado al agregar estaciones en diferentes zonas de la ciudad. En este trabajo también se introdujo una extensión de diversidad de rango que aquí se ha llamado frecuencia de rango, la cual ayuda a observar la frecuencia con la que las estaciones aparecen en diferentes rangos a lo largo del periodo utilizado. El uso de estas dos herramientas en conjunto puede ser de utilidad al momento de determinar cambios en las rutinas de los suscriptores observando el comportamiento de las estaciones en el

sistema.

Posteriormente se clasificó cada estación por el lugar que ésta ocupa dentro de las regiones obtenidas por la diversidad de rango. Los rangos con mayores valores según el criterio de actividad mostraron una menor variabilidad en medida que poseían mayor valor. Con el uso de algoritmos de machine learning para el aprendizaje y predicción de la magnitud de uso en sus estaciones se realizaron del desempeño del pronóstico en el ordenamiento de las estaciones dentro del sistema, los resultados obtenidos tras realizar las pruebas estadísticas correspondientes en las magnitudes de error de los resultados mostraron que existe un mayor grado de predictibilidad para la región comprendida de la cabeza en diversidad de rango, comparada con las demás regiones. Los resultados satisfactorios aquí se prestan para diferentes interpretaciones y medidas de uso que se le pueden dar a este conjunto de herramientas:

Los algoritmos más confiables que actualmente se utilizan para realizar predicciones suelen ser de un origen no paramétrico, trazando regiones de decisión adecuadas a los datos que fueron ajustados. Estas técnicas no pueden ajustarse al cambio posterior que llegue a ocurrir cambiando el comportamiento del sistema y conllevando a una predicción de menor calidad a medida que el modelo se va desactualizando. Diversidad de rango puede servir como una herramienta de monitoreo para esos cambios (estacionales, modificaciones en el sistema) y reaccionar de acuerdo a las observaciones.

Por último cabe mencionar un experimento que se realizó pero el cual no obtuvo resultados satisfactorios: la predicción en la actividad de las estaciones realizando pronósticos de conteos en las entradas/salidas hacia el futuro. La expectación sobre este trabajo mantenía que un mayor nivel de predictibilidad se concentraba sobre aquellas estaciones que están situadas sobre la región de la cabeza bajo el argumento de tener un rol bien establecido en la dinámica de la ciudad. Para este experimento se utilizó un modelo de regresión para series de tiempo estacionales conocido como SARIMA (Seasonal Auto Regressive Integrated Moving Average) ajustando un modelo para cada una de las estaciones y realizando una separación en el conjunto de datos para entrenamiento y prueba. Los resultados obtenidos de este experimento al realizar la predicción sobre el conjunto de prueba no reflejaron un mejor desempeño entre las regiones de la diversidad de rango, donde el MSE de las estaciones en la cabeza no reflejó menor error que las estaciones que conforman el cuerpo y cola.

6.2. Trabajo futuro

Para este trabajo solo se tomó como objeto de estudio la actividad en las estaciones, este es un buen comienzo para comprender su funcionamiento pero ignora el importante hecho de que este sistema esta compuesto de interacciones que se asemejan a una red completamente conexa, en el cual se puede generar un viaje de una estación a cualquier otra (inclusive hacia la estación de origen). Para un trabajo futuro se propone extender el enfoque planteado aquí a un modelo de grafos dirigidos con la intención revelar un comportamiento más descriptivo de las actividades que ocurren, como lo sería obtener

las densidades en las rutas generadas de una estación origen a una destino.

Otro aspecto que podría ser explotado con la información que aporta rank diversity es el desplazamiento que ocurre entre los rangos realizando una comparación sobre las múltiples diversidades de rango obtenidas en diferentes intervalos de tiempo, con el cual podrían ser utilizadas para obtener de manera automatizada cambios que se generen en la estructura del sistema que no pueden observarse a simple vista para capturar la distribución móvil (drifting distribution) que sucede con el tiempo para tomar medidas al respecto.

Del experimento de regresión con series de tiempo SARIMA no se contempló lo siguiente: Las estaciones que cuentan con una mayor magnitud debido al elevado flujo de generado por viajes tienden a obtener un mayor error que las estaciones con un flujo moderado o bajo. Las estaciones que mantienen un comportamiento menos predecible suelen tener una magnitud menor en su actividad y en cierta medida esto puede actuar como una malinterpretación para las diferentes magnitudes de error que se manejan así que, a pesar de que se observó un comportamiento más predecible en las estaciones de la cabeza al realizar este experimento existe la posibilidad de que, aunque los errores que se cometan durante la regresión para las estaciones de la cabeza sean menores, éstos sean más graves que un error de mayor frecuencia pero menor magnitud. Para un posterior estudio que contemple estas diferencias se propone normalizar la magnitud en los conteos de todas las estaciones. Esto con la finalidad de obtener resultados normalizados en la predicción de actividad para estaciones que conforman la región de la cabeza en comparación con las estaciones que conforman las regiones del cuerpo y la cola.

Como Ecobici ha liberado el registro de las actividades realizadas por los suscriptores también se ha realizado en diferentes ciudades. Estas ciudades muchas veces cuentan con una cultura diversa, temperaturas a lo largo del año que influyen más en el uso de los BSS e incluso existen ciudades con un relieve no uniforme, donde todos estos factores en conjunto pueden llegar a afectar los viajes realizados en otros aspectos no observados aquí. Como trabajo futuro para este caso se propone el estudio de otros BSS situados en otras ciudades, comparar las similitudes y diferencias que existen entre ellas, observar si el comportamiento de diversidad de rango también está presente en estos otros sistemas y comparar haciendo un análisis que comprenda diferentes periodos separados, como se realizó para Ecobici, si la influencia de estos factores afecta la estructura de la diversidad de rango en el sistema.

Para realizar predicciones más certeras en necesario tener en consideración la mayor cantidad de factores que afecten los viajes realizados por un BSS. La adquisición de nuevos descriptores para agregarlos al modelo de predicción puede ayudar a explicar comportamientos que de otro modo nos resultaría imposible apreciar. Como trabajo futuro se propone obtener y agregar descriptores ajenos al sistema pero que ofrezcan explicación en el uso del mismo, descriptores tales como: Precipitación, temperatura, indicadores de días festivos, indicadores de contingencia ambiental y la recolección e implementación del estado de las estaciones que ofrece la aplicación para smartphone de Ecobici, con el cual sería posible realizar un análisis con respecto a la demanda de

lugares y bicicletas por hora que surge para cada estación dentro del sistema.

Apéndice A

Código/Manuales/Publicaciones

A.1. Algoritmos complementarios

A.1.1. Bayesian Information Criteria

El BIC (También conocido como criterio de Schwarz) es una función incremental donde se asume un error de varianza σ_e^2 y una función incremental k. Cuando hace una selección de entre distintos modelos con diferentes componentes que expliquen las distribuciones generadas asumiendo una distribución gaussiana. La variable dependiente para este caso es la variación sin explicación y el número de variables explicatorias aumenta el valor del BIC, penalizando fuertemente modelos complejos. Así, un BIC con menor valor implica menores variables explicatorias, un mejor ajuste, o las dos en conjunto. Dentro de las propiedades de BIC que son dignas de mencionar se encuentran:

- Es independiente del componente a priori.
- Penaliza la complejidad del modelo, donde complejidad se entiende por número de parámetros utilizados por el modelo para explicar la varianza de los datos.
- Es aproximadamente igual a la longitud mínima de descripción, pero con símbolo negativo.
- Puede ser utilizada para seleccionar el número de clusters de acuerdo a la complejidad intrínseca presente en un conjunto de datos particular.

Bajo el presupuesto de que los errores del modelo son independientes e idénticamente distribuidos (i.i.d) de acuerdo a una distribución normal y que la condición de frontera que es la derivativa de la función de verosimilitud a escala logarítmica (log likelihood en inglés). Se tiene entonces el siguiente modelo BIC:

$$BIC = -2 \cdot ln(\hat{L}) + k \cdot ln(n) \tag{A.1}$$

Donde:

- $\hat{L}$: Es el valor maximizado del modelo M, es decir, $\hat{L} = P(x|\hat{\theta}, M)$, donde $\hat{\theta}$ son los parámetros a maximizar en la función de verosimilitud.
- x: Corresponde a los datos observados.
- θ: Los parámetros del modelo M.
- n: El número de puntos x equivalente al tamaño de la muestra.
- k: Corresponde al número de parámetros utilizados para explicar x, este parámetro describe que tan simple es el modelo.

La varianza que explica la distribución para cada parámetro del modelo en este caso es tomada como la simple distancia euclidiana entre un punto y la media de su distribución:

$$\hat{\sigma_e^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_i)^2 \tag{A.2}$$

La función para maximizar la probabilidad está dada por:

$$\hat{P}(x_i) = \frac{n_i}{n} \cdot \frac{1}{\sqrt{2\hat{\sigma_i}^M}} exp\left(-\frac{1}{2\hat{\sigma_i}^2}\|x_i - \mu_i\|^2\right) \tag{A.3}$$

Y la función de verosimilitud que hay que maximizar es:

$$L(x) = log\prod_i P(x_i) = \sum_i \left(log\frac{1}{\sqrt{2\pi\sigma_i^M}} - \frac{1}{2\sigma_i^2}\|x_i - \mu_i\|^2 + log\frac{n_i}{n}\right) \tag{A.4}$$

Dado que se está utilizando un modelo de mezclas, tenemos que:

$$p(x_i \mid y) \propto \sum_i^{k_y} N(x_i; \mu_i; \sigma_i) \tag{A.5}$$

Con cantidades fijas $1 \leq k \leq n$. Enfocándose solamente en el conjunto x los puntos que pertenecen al centroide k y conectando los estimados de máxima verosimilitud. Cuando se ajusta un modelo de mezclas lo que se intenta hacer es maximizar la función de verosimilitud de los datos. El criterio BIC puede ser utilizado para seleccionar el número de componentes en un modelo de mezclas de Gaussianas de una manera eficiente [Pelleg et al.] y, en teoría, recobra el verdadero número de componentes en el régimen asintótico.

A.1.2. Reducción de dimensionalidad aplicando PCA

Principal Component Analysis es un algoritmo en el que a menudo se utiliza con el objetivo para reducir las dimensiones de un conjunto de datos d-dimensional proyectándolo en un sub-espacio k-dimensional, donde $k < d$ en orden de incrementar la eficiencia computacional mientras se mantiene la mayoría de la información [Abdi and Williams.]. Para lograr esto se calculan los eigenvectores (los componentes principales) de un conjunto de datos para recolectarlos en una matriz de proyección, cada uno de esos eigenvectores está asociado con un eigenvalor que puede ser interpretado como la 'magnitud' asociada al eigenvector correspondiente. Si algunos eigenvalores tienen un valor de magnitud significativamente mayor que los otros con los que se hizo la reducción vía PCA en un sub-espacio más pequeño dimensionalmente es una común práctica eliminar los eigenpares "menos informativos" con la finalidad de reducir la dimensionalidad. Para lograr un acercamiento por PCA se realizan los siguientes pasos:

- Se estandarizan los datos.
- Se obtienen los eigenvalores y los eigenvectores de la matriz de covarianza o la matriz de correlación o se realiza descomposición de vectores singulares (SVD)
- Se ordenan los eigenvalores en orden descendente y se escogen los k eigenvectores que correspondan a los k eigenvalores donde k es el número de dimensiones del nuevo sub-espacio de características ($k < d$).
- Se construye una matriz de proyección W de los k eigenvectores seleccionados.
- Se transforma el conjunto de datos original X vía W para obtener un sub-espacio de características k-dimensional Y.

A.1.2.1. Estandarización

Como una buena práctica suele realizarse una estandarización sobre los datos se con el propósito de que la magnitud de los valores no afecte los resultados obtenidos por la técnica. Cuando se estandarizan datos previos a realizar PCA en la matriz de covarianza depende de las escalas en las medidas de las características originales. Debido a que PCA cede un sub-espacio de características que maximiza la varianza sobre sus ejes, tiene sentido el estandarizar los datos, en especial si las medidas se encuentran en diferentes escalas. Se continúa con la estandarización de los datos en una escala unitaria con media $\mu = 0$ y varianza $\sigma = 1$, que es el requerimiento para el desempeño óptimo en muchos algoritmos de machine learning.

A.1.2.2. Ordenamiento y selección de los eigenpares

La meta para la práctica de PCA es la de reducir la dimensionalidad del espacio original de características proyectándolo en un sub-espacio más pequeño, donde

los eigenvectores formarán los ejes. Los eigenvectores solo definen las direcciones de los nuevos ejes, así que todos tienen una magnitud de 1. En orden de decidir cuáles eigenvectores pueden ser eliminados sin perder demasiada información para la construcción de un sub-espacio con menor dimensionalidad, se necesita inspeccionar sus correspondientes eigenvalores asociados: los eigenvectores con los eigenvalores más pequeños indican menos información explicada con respecto a la distribución de los datos; Para perder la menor cantidad de varianza explicada sobre los datos aquellos eigenpares con valores menores son los que suelen ser eliminados. En orden de realizar lo anterior, un acercamiento común es el ordenamiento de los eigenvalores del más alto al más bajo en orden se seleccionar los k eigenvectores.

A.1.2.3. Varianza explicada

Con los eigenpares ordenados, lo siguiente que hay que determinar es cuántos componentes principales son los que van a ser utilizados para este nuevo sub-espacio de características. Una medida útil es la llamada 'varianza explicada', que tiene que ser calculada de los eigenvalores. La varianza explicada nos dice cuánta información (varianza) puede ser atribuida para cada uno de los componentes.

A.1.2.4. Matriz de proyección

La construcción de la matriz de proyección que será utilizada para reducir la información de los datos a un nuevo sub-espacio de características. La matriz de proyección puede ser entendida como una matriz de los k eigenvectores con mayor valor. Para reducir el espacio de características 36-dimensional a un sub-espacio de características 4-dimensional se seleccionan los 4 eigenvectores con los eigenvalores más altos para construir una matriz de eigenvectores W con dimensiones d x k.

A.1.2.5. Proyección en el nuevo espacio de características

Se utiliza la matriz de proyección 36 x 4-dimensional W para transformar nuestras observaciones a un nuevo sub-espacio con la ecuación:

$$Y = X \cdot W \tag{A.6}$$

Donde W es una matriz de 90 x 4 de nuestras observaciones transformadas y X es la matriz de observaciones original de los datos.

A.2. Técnicas en complejidad

A.2.1. Definición de las regiones de la cabeza, cuerpo y cola utilizando diversidad de rango

Como se ha venido diciendo, la diversidad de rango es una técnica enfocada a medir el cambio de los rangos en un sistema utilizando una dependencia en el tiempo. Existen diferentes distribuciones de la diversidad de rango de acuerdo a su comportamiento: la cabeza consiste en los elementos que casi no cambian su rango en el tiempo, el cuerpo son elementos de uso general en el sistema, los que tiene mayor variabilidad según se ha observado y la cola son elementos de uso específico que varían mucho su rango en el sistema en el tiempo. La diversidad d es una función del rango k donde $d(k)$ mide cuantos elementos diferentes aparecen para un rango k dado durante el tiempo considerado ($\Delta = 2$ años para este caso). Por ejemplo, para el periodo comprendido de la fase 1, $d(1)$ en el sistema de Ecobici el resultado es de 16/486, donde 16 diferentes estaciones ocuparon el rango 1 para todos los días considerados. Como ya se ha establecido antes, este comportamiento $d(k)$ asemeja una curva sigmoidal donde la sigmoide es una distribución acumulativa Gaussiana:

$$\Phi_{\mu,\sigma}(k_0) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{k} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \tag{A.7}$$

Que es una función de k. El valor de μ es obtenido identificando al valor k_0 de menor valor para el cual:

$$d(k_0) = \frac{max_i d(k_i)}{2} \tag{A.8}$$

Mientras el valor σ que representa la desviación estándar ajustada donde $d(k)$ se acerca a sus valores extremos, cuyos los intervalos de confianza están dados por:

$$k_{\pm} = \mu \pm 2\sigma \tag{A.9}$$

Luego así, los segmentos de nuestro interés se obtienen de la siguiente manera:

- Cabeza: $k \leq k_{-}$
- Cuerpo: $k_{-} \leq k \leq k_{+}$
- Cola: $k_{+} \leq k$

A.3. Medidas de errores para evaluación

A.3.1. Normalized Discounted Cumulative Gain (NDCG)

El $NDCG$ mide el desempeño de un sistema de recomendación basado en el grado de relevancia de las entidades recomendadas. La medida varía de 0.0 a 1.0 con 1.0 representando el rango ideal de las entidades. Esta métrica es comúnmente usada en retorno de información y evalúa el desempeño de motores en páginas de búsqueda. $NDCG$ es una variación de DCG. La premisa de DCG es que las entidades con alta relevancia que aparezcan en un rango bajo deben ser penalizados, donde la ganancia acumulada descontada en un rango particular k es definido como:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2 (i + 1)} \tag{A.10}$$

Donde K representa el máximo número de entidades que pueden ser recomendadas. $IDGC_k$ es el máximo posible (el orden ideal) de DCG para un conjunto de secuencias dado. Así, $NDCG$ puede ser obtenido mediante:

$$NDGC_k = \frac{DCG_k}{IDCG_k} \tag{A.11}$$

$NDCG$ puede ser promediado de entre repetidas iteraciones de rankeo para obtener una medida acerca del desempeño promedio en un algoritmo. Nótese que en un algoritmo de rankeo perfecto el DCG_k será el mismo que el $IDCG_k$ produciendo un $NDCG$ de 1.0. Todos los cálculos de $NDCG$ son entonces valores relativos en un intervalo de 0.0 a 1.0 y son comparables en cada ordenamiento.

A.4. Pruebas estadísticas

A.4.1. ANalysis Of VAriance (ANOVA)

Las pruebas F (o $f - test$ por sus siglas en inglés) son usadas para la comparación de factores de la desviación total. El propósito de realizar la prueba $ANOVA$ sobre un conjunto de grupos consiste en rechazar o no la hipótesis nula que establece que la media estadística para todas las muestras de los tratamientos son iguales $H_0 : \mu_1 = \mu_2 \cdots = \mu_n$, utilizando el método $ANOVA$ de un solo factor ($Single - way$ o $one - way$ en inglés), cuya significancia estadística es probada comparando las pruebas estadísticas F ($F - test$ statistics por sus siglas en inglés):

$$F = \frac{\text{varianza entre los tratamientos}}{\text{varianza dentro de los tratamientos}} \tag{A.12}$$

Donde los tratamientos corresponden a los elementos que la suposición establece son las diferentes poblaciones, la varianza entre los tratamientos representa la distancia de la media en la media muestral total, la varianza dentro de los tratamientos es la variación dentro de cada muestra. Para calcular tanto el numerador como el denominador se utiliza la siguiente fórmula:

$$F = \frac{MS_{tratamientos}}{MSerror} = \frac{\frac{SStratamientos}{l-1}}{\frac{SSerror}{nt-l}} \tag{A.13}$$

Donde MS es la media al cuadrado, l = número de tratamientos y nt = total de casos para la distribución F con l–1, nt–l grados de libertad. El valor esperado de F es $1 + n\sigma^2_{Tratamientos}/\sigma^2_{Error}$ (Donde n es el tamaño de la muestra del tratamiento) el cual es 1 si no existe una prueba convincente para el tratamiento. En medida de que F incrementa hacia valores mayores a 1, las evidencias de inconsistencia con respecto a la hipótesis nula incrementan. Dos métodos experimentales de una F incremental aumentan el tamaño de la muestra y reducen la varianza del error a controles experimentales ajustados. Existen dos métodos concluyentes para una prueba de hipótesis $ANOVA$, ambos que producen el mismo resultado:

- En el método del libro se comparan los valores observados de F con un valor crítico de F tablas determinadas. El valor crítico F es una función de grados de libertad como numerador y como el nivel de significancia a un nivel (α). Si $F \leq F_{critica}$ entonces la hipótesis nula es rechazada.
- Métodos de computadora calculan la probabilidad ($p - value$) de que el valor F sea mayor o igual que el valor observado. La hipótesis nula es rechazada si la probabilidad es menor o igual que un nivel de significancia (α).

La prueba F de $ANOVA$ es conocida por ser cercana al óptimo en el sentido de minimizar errores de tipo falso negativo de una tasa fija de errores falsos positivos.

A.4.2. Tukey's test

Tukey's test es una prueba de seguimiento (follow throught test en inglés) la cual comúnmente es utilizada después de haber obtenido evidencia de que existe una población con una media diferente a la muestra total en los grupos de tratamiento con $ANOVA$. Esta prueba consiste en comparar las medias de todos los elementos de tratamiento con las medias de todos los demás tratamientos, es decir, una serie de múltiples comparaciones simultaneas realizadas en pares donde la hipótesis nula y la hipótesis alternativa están definidas como las siguientes:

$$H_0 : \mu_i = \mu_j \tag{A.14}$$

$$H_1 : \mu_i \neq \mu_j, \text{ i y j son dos poblaciones diferentes} \tag{A.15}$$

Estas comparaciones verifican si la diferencia entre dos medias $\mu_i - \mu_j$ es mayor a un error estándar esperado (o valor crítico), si la diferencia es mayor que el valor crítico la hipótesis nula es rechazada. El coeficiente de confianza para el conjunto, cuando es muestreado en tamaños iguales, es exactamente 1–α. Para todas las muestras de tamaño desigual. El coeficiente de confianza es mayor a 1–α.

El Tukey's test funciona bajo tres importantes aseveraciones:

1. Las observaciones que están siendo probadas son independientes dentro y entre los grupos.

2. Los grupos asociados con cada media en la prueba se encuentran distribuidas de manera normal.

3. La varianza que existe es igual dentro de los grupos y entre los grupos asociados con cada media en la prueba.

El Tukey's test está basado en una fórmula muy similar a la del t-test. De hecho, el Tukey's test es esencialmente una t-test, a excepción de que este método corrige los errores realizados cuando se realizan múltiples comparaciones, la probabilidad de hacer un error de tipo I en al menos una de las comparaciones. Dado que el Tukey's test corrige eso, es considerada una prueba más adecuada para realizar múltiples comparaciones que lo que sería un número variado de t-test diferentes.

La fórmula para el Tukey's test es la siguiente:

$$q_s = \frac{Y_A - Y_B}{SE} \tag{A.16}$$

Donde Y_A es el valor más grande de las dos medias comparadas, Y_B es la media de menor valor entre las medias comparadas y SE es el error estándar de los datos en cuestión y el valor de q_s puede entonces ser comparado con una el valor de q de un rango de distribución de student. Si el valor de q_s es mayor a una $q_{critica}$ el valor obtenido de la distribución, se concluye que las dos medias son significativamente diferentes. Como la hipótesis nula para el Tukey's test establece que todas las medias que están siendo comparadas son de la misma población ($\mu_1 = \mu_2 = \ldots = \mu_n$) las medias deberían estar normalmente distribuidas. Esto da la asunción de normalidad del Tukey's test. Los límites para el intervalo de confianza para todas las comparaciones por pares con un coeficiente de confianza a al menos 1–α que es:

$$\hat{y_i} - \hat{y_j} \pm \frac{q_\propto; k; N-k}{\sqrt{2}} \hat{\sigma_\varepsilon} \sqrt{\frac{2}{n}} i, j; 1, \cdots, k, i \neq j \tag{A.17}$$

Es importante darse cuenta que el punto estimador y la varianza estimada son lo mismo para una comparación por pares. La única diferencia entre los límites de confianza para comparaciones múltiples y esos para una sola comparación es el múltiple de la desviación estándar estimada.

A.5. Técnicas de machine learing

A.5.1. Gradient Boosted Regression Trees (GBRT)

GBRT es un modelo de ensambles basado en la partición del espacio de características en regiones utilizando árboles de decisión que frecuentemente son de muy baja profundidad generados de manera secuencial y reduciendo el error de la información obtenida de sus modelos pasados, donde un nuevo árbol de decisión es ajustado a los residuos del árbol ajustado anterior (13, chap. 8). Después, este modelo es agregado en orden de adaptar los residuos, reduciendo de manera progresiva el error sobre cada decisión, utilizando:

$$F(X) = \sum_{m=1}^{M} C_m \cdot 1(X \in R_m) \tag{A.18}$$

Donde R representa una región de decisión sobre el espacio de características. *GBRT* es un acercamiento que puede ser aplicado a muchos modelos de aprendizaje estadístico para regresión o clasificación. *GBRT* es un modelo de múltiples árboles de decisión donde cada árbol es desarrollado utilizando información de árboles previamente desarrollados: cada árbol es ajustado a una versión modificada del conjunto de datos original. Considerando el enfoque de árbol de regresión que se utilizará aquí, *GBRT* involucra combinar un gran número de árboles de decisión $f^1, ..., f^b$. Algoritmo para GBRT:

1. Determinar $\hat{f}(x) = 0$ y $r_i = y_i$ para toda i en el conjunto de entrenamiento.
2. Para $b = 1, 2, \ldots, B$, repetir:
   - Ajustar un modelo $\hat{f}^b$ con d divisiones ($d + 1$ nodos terminales) al conjunto de entrenamiento (X, r).
   - Actualizar $\hat{f}^b$ agregando una versión reducida del árbol:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{A.19}$$

   - Actualiza los residuos,

3. Salida del modelo,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{A.20}$$

Ajustando pequeños árboles a los residuos, se mejora $\hat{f}$ lentamente en áreas donde no se desempeña bien. El parámetro de reducción λ reduce el proceso aún más, permitiendo a más árboles de diferentes formas atacar los residuos. $GBRT$ tiene 3 parámetros:

1. El número de árboles B. El modelo $GBRT$ puede sobre aprender si B es demasiado grande, aunque este sobre aprendizaje tiende a ocurrir lentamente.

2. El parámetro de reducción λ, un pequeño número positivo. Éste controla la tasa a la que el modelo $GBRT$ aprende. Los valores típicos son 0.01 o 0.001 y la decisión correcta puede depender del problema. Valores demasiado pequeños de λ pueden requerir utilizar valores de B muy grandes en orden de obtener un buen desempeño.

3. El número de particiones de cada árbol d, que controla la complejidad del ensamble $GBRT$. A menudo una $d = 1$ funciona bien, en cuyo caso cada árbol es un tocón (*stump* en inglés), consistiendo de una sola partición. En este caso, el modelo $GBRT$ es ajustado un modelo aditivo, dado que cada término involucra solo una sola variable. De una manera más general d es la profundidad de interacción y controla las interacciones del modelo $GBRT$, ya que d puede involucrar a lo más d variables.

Los parámetros seleccionados para el modelo GBRT fueron los siguientes:

- Número de árboles $B = 1000$
- Razón de reducción $\lambda = 0.01$
- Profundidad de cada árbol $d = 1$

Bibliografía

[1] Abdi, H. and Williams., L. J. (2010). *Principal component analysis.* Springer. 67

[2] Borgnat, P., Abry, P., and Flandrin., P. (2011). Shared bicycles in a city: a signal processing and data analysis perspective. *Advances in Complex Systems.*, 14:415–438. 8, 20, 29

[3] CDMX, H. (2014). Datos liberados hack cdmx. `http://datos.labplc.mx/datasets/`. 12

[4] Chemla, D., Meunier, F., and Calvo, R. W. (2013). Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization*, 10:120–146. 2

[5] Chen, M. A. and K., Z. W. (2011). Web-search ranking with initialized gradient boosted regression trees. *Journal of Machine Learning Research.*, 14:77–89. 43

[6] Cocho, G., Flores, J., Gershenson, C., Pineda, C., and Sanchez, S. (2015). Rank diversity of languages: Generic behavior in computational linguistics. *PLOS ONE*, 10:1–12. 4, 37

[7] Dias, G. M., Bellalta, B., and Oechsner, S. (2015). Predicting occupancy trends in barcelona's bicycle service stations using open data. *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, 21:439–445. 9

[8] Etienne, C. and Latifa., O. (2014). Model-based count series clustering for bike-sharing system usage mining, a case study with the vélib' system of paris. *ACM Trans. Intell. Syst. Technol.*, 5(3):39:1–39:21. 8

[9] for local goverment, I. (2015). In focus: The last mile and transit ridership. http://www.ca-ilg.org/resource/being-less-forthright-about-agencys-decision-conditioning-project-death-example. 11

[10] Froehlich, J., Neumann, J., and Oliver., N. (2009). Sensing and predicting the pulse of the city through shared bicycling. *IJCAI International Joint Conference on Artificial Intelligence.*, 2:1420–1426. 7, 14, 19, 20

[11] Goodman, Anna, Cheshire, and James (2014). Inequalities in the london bicycle sharing system revisited: Impacts of extending the scheme to poorer areas but then doubling prices. *Journal of Transport Geography*, 41:272–279.

[12] Google (2014). Google api de elevaciones. https://developers.google.com/maps/documentation/elevation/intro. 16

[13] Hastie, T., Tibshirani, R., and Friedman., J. (2009). *An Introduction to Statistical Learning with Applications in R*. Springer. 20, 73

[14] Kalervo, J. and Jaana, K. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446. 46

[15] Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., and Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: Applied*, 21:1–14. 45

[16] Morales, J. A., Sánchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., Zizumbo, J., and Iñiguez, G. (2016). Universal temporal features of rankings in competitive sports and games. *J. Alg.*, 111:427–430. 37

[17] oficial, E. (2014). Datos oficiales de ecobici. https://www.ecobici.df.gob.mx/es/informacion-del-servicio/open-data. 12

[18] P, L., C, B., and Q., W. (2008). Learning to rank using classification and gradient boosting. *Microsoft*. 43

[19] Pelleg, Dan, Moore, and W., A. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Linear Algebra Appl.*, 173. 21, 66

[20] Rixey, R. (2013). Station-level forecasting of bike sharing ridership: Station network effects in three u.s. systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2387:46–55. 8

[21] Vogel, P., Greyser, T., and Mattfeld., D. C. (2011). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences.*, 20:514–523. 2, 20