PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MÉTODOS DE APRENDIZAJE ESTADÍSTICO PARA PREDECIR VELOCIDADES DE BUSES Y ANALIZAR EL IMPACTO DE LAS VARIABLES EXPLICATIVAS

JAN BERCZELY PRADA

Tesis para optar al grado de
Magister en Ciencias de la Ingeniería

Profesor Supervisor:
RICARDO GIESEN ENCINA

Santiago de Chile, (Mayo, 2017)

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MÉTODOS DE APRENDIZAJE ESTADÍSTICO PARA PREDECIR VELOCIDADES DE BUSES Y ANALIZAR EL IMPACTO DE LAS VARIABLES EXPLICATIVAS

JAN BERCZELY

Tesis presentada a la Comisión integrada por los profesores:

RICARDO GIESEN ENCINA

FELIPE DELGADO BREINBAUER

MIGUEL ANDRES FIGLIOZZI

SERGIO GUTIÉRREZ

Para completar las exigencias del grado de
Magister en Ciencias de la Ingeniería

Santiago de Chile, (Mayo, 2017)

Dedicada a mis padres Mariela y Gabriel, mi hermano Alexis, y en especial a mi hermano Maki.

AGRADECIMIENTOS

Me gustaría agradecer especialmente a mis padres por el apoyo incondicional en todos los proyectos que he decidido embarcarme en mi vida. Agradezco también a mi hermano Alexis por ser un excelente hermano, y a mi hermano Maki, por dejarme con tan buenos recuerdos y enseñanza de vida.

También quiero agradecer a la familia extendida la cual me ayudó durante todo este tiempo e hizo que fuera más liviano trabajar en la tesis. Anto, Mirna, Mari y Geri.

A su vez, me gustaría agradecer a todo el departamento de Ingeniería de Transporte y Logística, los cuales han generado un ambiente no solo de trabajo, sino que también de pertenencia y buenos momentos. En especial al profesor Ricardo Giesen, que con su conocimiento y experiencia me guio a lo largo de este proceso e hizo que fuera posible afrontar este desafío. Sin dejar de lado al resto de los profesores y alumnos de postgrado, por tener siempre una excelente disposición y buena onda.

Agradezco también a Pedro Lizana y TransitUC, que sin ellos no hubiera podido recopilar la información necesaria para realizar la tesis. A RoutingUC y Niko Julio, por guiarme en un comienzo y también proveerme de datos, y a Centro de Desarrollo Urbano Sustentable (CEDEUS), FONDAP 15110020.

Finalmente quiero mencionar a mis amigos del colegio, hockey y universidad, por acompañarme durante todos estos años y generar lazos que quedaran de por vida.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El creciente interés por generar Sistemas de Transporte Inteligente (ITS por sus siglas en inglés), ha incentivado la investigación en métodos para predecir velocidades o tiempos de viaje de los buses, factores que tienen un impacto positivo tanto para el usuario como para el operario. Al usuario no solo lo ayuda a planificar mejor sus rutas, sino también a disminuir la ansiedad sicológica por desconocer las horas de pasada y llegada de buses. Por otro lado, al operario lo ayuda a tomar decisiones estratégicas, tales como optimizar el número de buses y la frecuencia de pasada por los paraderos.

Los modelos de aprendizaje estadístico han logrado gran popularidad por sus buenos resultados, y por su capacidad para adaptarse fácilmente a los requerimientos del modelador. Es por ello que la literatura ha reportado trabajos relevantes en la predicción de velocidades y tiempos de viaje de los buses. Sin embargo, no hay consenso respecto a qué modelo es mejor, ya que estos dependen del tipo de datos con que se trabaja, y del lugar en que estos han sido recopilados.

Este trabajo tiene dos objetivos. El primero es comparar el desempeño de tres modelos de aprendizaje estadístico (Regresión Lineal Múltiple, Máquinas de Soporte Vectorial y Redes Neuronales), y contrastarlos con dos modelos base o *benchmark* (de información histórica y en tiempo real). Para ello, se utilizan como caso de estudio tres servicios de buses de la ciudad de Santiago de Chile. El segundo objetivo busca determinar aquellas variables explicativas que son más o menos significativas al incluirlas en los modelos. Para tales efectos se trabajó con las variables velocidad de buses reportadas por GPS, características de la demanda de usuarios (subidas, bajadas y carga), infraestructura, y factores de entorno.

En los tres servicios de buses, el modelo que tuvo mejores resultados en términos de la raíz del error cuadrático medio, es el de Redes Neuronales, seguido por el modelo de Regresión Lineal Múltiple, y luego por la Máquinas de Soporte Vectorial. En todos los casos, los modelos de aprendizaje estadístico superaron los modelos *benchmark*, con un desempeño que varía entre un 10% y un 25% en la disminución del error.

Respecto a las variables explicativas involucradas, se encuentra que la utilización de la variable de velocidad tiene un impacto relevante, mientras que el resto de las variables analizadas solo disminuyen los errores en un 2%. Del análisis efectuado, pareciera que las variables de velocidad llevan implícitas, en su valor, el efecto de las otras variables de características de demanda de usuarios, infraestructura y factores de entorno.

Palabras clave: *Predicción de velocidad, predicción de tiempos, predicción de velocidades de buses, variables que influyen en la velocidad, sistemas de transporte inteligente, ITS, regresión lineal, máquinas de soporte vectorial, redes neuronales, LR, SVM, ANN.*

ABSTRACT

The increasing interest in generating Intelligent Transport Systems (ITS) has stimulated the research on methods of prediction of speeds and travel times of buses, which have a positive impact for both the users and the operators. In case of users, it helps them to plan their routes, and to reduce the anxiety of waiting without knowing the time of arrival of their bus. In case of operators, it helps them to optimize the number of buses and frequencies at the bus stops.

Statistical learning models have seized high popularity because of their good results and ability to easily adapt to the requirements of the modeler. As a consequence, several studies have been reported in predicting speed/time of travel of buses. However, there is no consensus in which model is best, since that depends on the type of data with which they work, and the place it has been collected.

This paper has two objectives. The first one is to compare the performance of three models of statistical learning (Multiple Linear Regression, Support Vector Machines and Neural Networks), and to compare them with two benchmark models (historic and real time information). To do this, three bus services of the city of Santiago de Chile are used as a study case. The second objective seeks to determine which explanatory variables are more or less significant in the outcome of the model. Our analysis includes variables such as bus speed reported by GPS, user (passengers get on and offs, bus load), infrastructure and environmental factors.

In the three bus services analyzed, the model that showed the best results, in terms of the root mean square error, is Neural Networks, followed by Multiple Linear Regression and then by the Support Vector Machines. In all cases, machine learning models exceeded benchmark models, with a decrease in the error that varies between 10% and 25%

Regarding the explanatory variables involved, we concluded that only speed variables have a relevant impact, whilst the addition of the other analyzed variables only had a minor effect (less than 2% reduction of errors). It seems that speed variables have implicit, in their value, the effect of the other variables, such as user factors, infrastructure and environmental factors.

Keywords: Predicting speed, predicting time, bus speed prediction, variables that influence the speed, intelligent transport systems, ITS, linear regression, support vector machines, artificial neural network, LR, SVM, ANN.

1. INTRODUCCIÓN

En los últimos años, el desarrollo e incorporación de Tecnologías de Información y Comunicación (TIC) ha facilitado la recolección, difusión y uso de la información. Esto ha generado un creciente interés en los Sistemas de Transporte Inteligente (ITS, por sus siglas en ingles), los cuales hacen uso de las TIC para mejorar la eficiencia y conveniencia de los sistemas de transporte, donde, el aumento de la congestión y del volumen de tráfico ha generado un creciente interés en el modelamiento del tráfico (Du, Peeta, & Kim, 2012). Estos modelos de tráfico son la base para dos sistemas dentro de ITS: Sistema de Gestión Avanzado de Tráfico (ATMS) y Sistemas de Información Avanzados de Viajero (ATIS). ATMS es generalmente utilizado por ingenieros y administradores para mejorar la movilidad y obtener una red de tráfico más eficiente y segura. En cambio, ATIS está destinado a proporcionar a los pasajeros la información de tráfico y herramientas necesarias para permitir una mejor toma de decisiones (Mori et al., 2015).

La predicción de tiempos (velocidades) de viaje en el transporte público es uno de los factores claves para generar un Sistema de Transporte Inteligente. Su utilidad para el ATMS radica en una mejor planificación estratégica de recorridos y número de buses, en la posibilidad de mantener una regularidad en los intervalos de pasada, en la opción de generar Sistemas de Semáforos Prioritarios (TSP) para el transporte (Xiong et al., 2015), entre otras muchas aplicaciones que se podrían generar hoy en día en una red interconectada, con el fin de generar una red lo más eficiente posible.

Ahora, en ATIS, la predicción de tiempos de viaje tiene un impacto en las personas tanto en la toma de decisiones como en la parte psicológica de estas. En la toma de decisiones, saber a qué hora llega el bus o cuanto demora un recorrido ayuda a planificar el viaje y reducir tiempos de espera en los paraderos. Por otra parte, Peng et al (2002) investigaron la percepción de los usuarios del transporte hacia el valor de la información precisa y encontraron que, si bien los incrementos esperados de

utilidad de los usuarios del transporte al tener esta información son modestos, los pasajeros ponen un enorme valor en saber cuándo llegará el próximo bus. Es decir, la reducción de la ansiedad a la espera de los pasajeros es una de las medidas más significativas de los beneficios para viajeros (Kotagiri & Pulugurtha, 2016).

Esta tesis trata justamente de la utilización de distintos modelos de aprendizaje estadístico para predecir velocidades de viaje de los buses. Mediante la incorporación de una serie de variables de entrada, que serán vistas más adelante, se predice la velocidad a futuro que va a tener un bus en un cierto tramo.

Existen muchos modelos de aprendizaje estadístico para predecir velocidades de buses, y no hay un consenso en relación al mejor método. En esta tesis se trabaja con tres modelos que han sido ampliamente utilizados en este ámbito: (i) Regresión Lineal Múltiple (MLR), (ii) Máquinas de Soporte Vectorial (SVM) y (iii) Redes Neuronales (ANN). Se eligió SVM y ANN ya que son métodos que han reportado muy buenos resultados en la predicción de velocidades (tiempos) de viaje (Jeong & Rilett, 2004; Yu, Lam & Tam, 2011 y Zheng, Zhang & Feng, 2012), por nombrar solo algunos. Sin embargo, son denominados modelos de caja negra, ya que se les provee de variables de entrada y el modelo entrega un valor de salida, sin tenerse muy en claro qué ocurre entremedio ni cómo afectan las distintas variables de entrada en las predicciones. Es por esto que se incorporó MLR, ya que si bien no ha reportado resultados tan buenos como los otros dos (Jeong & Rilett, 2004; Yu et al., 2011) es un modelo rápido de calibrar y fácil de interpretar, donde se puede ver claramente la contribución de cada una de las variables de entrada en las predicciones.

El objetivo de esta tesis es: (i) predecir velocidades de buses entre paraderos mediante el uso de distintos algoritmos de Aprendizaje Estadístico, conocido también como *Machine Learning*, y (ii) ver como impactan en la precisión de las predicciones la adición y sustracción de distintas variables explicativas.

Como se mencionó anteriormente, no hay consenso de qué modelos son mejores, ya que las características de la red influyen mucho, lo que genera que un modelo que ha mostrado buenos resultados en una ciudad, no necesariamente lo haga en otra. Es por esto que el primer objetivo, es decir, testear distintos modelos, es de gran importancia si se quisiera utilizar predicciones de tiempos de viaje para mejorar ATMS y ATIS en la ciudad de Santiago.

El segundo objetivo, en cambio, es más novedoso, ya que si bien se han propuesto muchos modelos, poco se ha hablado de las variables de entrada y de cómo estás afectan las predicciones. Julio, Giesen y Lizana (2016) proponen una serie de modelos para predecir velocidades de buses en la ciudad de Santiago, sin embargo, solo utilizan como variables de entrada datos de la posición de buses reportados cada 30 segundos por GPS (*Global Positioning System*) de cada bus para realizar las predicciones. Es por esto que en esta tesis se utilizan no solo dichos datos, sino que también una serie de variables más que podrían ser interesantes para entender cuanto aporta incluir variables adicionales y cuáles son las que generan un mayor impacto en las predicciones.

El tiempo de viaje o velocidad de un bus se ve afectado por una serie de factores que han sido validados en investigaciones anteriores. En términos generales, estas pueden ser agrupadas en cinco categorías: (i) características de la demanda, tales como subidas y bajadas de pasajeros, entre otras; (ii) factores de infraestructura, tales como número de semáforos y paradas, longitud de segmentos, pagos al interior del bus o en paraderos, etc.; (iii) factores de entorno, como clima, patrones de tráfico, entre otros; (iv) factores de comportamiento del bus, como programación, características del conductor; y (v) factores de operación y administración, como tiempos de viaje previsto, por qué puertas se permite subir a los pasajeros o control de paradas con retención de buses (Xinghao et al, 2013).

Como esta tesis tiene como uno de sus objetivos medir cómo impactan en las predicciones de velocidades la adición y sustracción de variables predictivas, se agregó a las variables de velocidad utilizadas por Julio et al (2016), variables de factores de demanda, de infraestructura y de entorno, las cuales son utilizadas tanto en forma histórica como en tiempo real.

Para realizar las predicciones se utilizaron una serie de datos del sistema de transporte Transantiago en la ciudad de Santiago, Chile. Para obtener las velocidades entre paraderos se utilizaron pulsos de GPS cada 30 segundos de cada bus correspondiente a tres servicios analizados. Por otro lado, las subidas pagadas en cada paradero fueron registradas mediante el sistema de pago con tarjeta BIP. Las subidas no pagadas, en cambio, son estimadas mediante la evasión histórica en los distintos paraderos y una regresión de Poisson desarrollada por Guarda (2015). Por último, las bajadas se estimaron usando matrices Origen-Destino desarrolladas por Munizaga y Palma (2012).

La tesis se divide en 7 capítulos. En el Capítulo 2 se hace un resumen de los algoritmos de aprendizaje estadístico que se utilizaran en la tesis, para luego presentar los trabajos reportados en la literatura en la predicción de velocidades o tiempos de viaje de buses. En el Capítulo 3 se define el problema y se explica por qué se predice tiempos y no velocidades de viaje. El Capítulo 4 presenta el caso de estudio, donde se explican los tres servicios de buses empleados para analizar el desempeño de los modelos. A su vez, se define el set de variables predictivas disponibles para realizar las predicciones. El Capítulo 5, se muestra la calibración de los modelos y se presentan dos modelos base o *benchmark*, que son el punto de comparación para medir el desempeño de los algoritmos de aprendizaje estadístico. En el Capítulo 6 se muestran los resultados obtenidos por cada uno de los modelos en los tres servicios de buses. A su vez, se presenta un análisis comparativo entre los servicios y una explicación del significado en términos del tiempo de viaje de los errores generados en las predicciones. En este capítulo también se analiza el

impacto que tienen en las predicciones las distintas variables explicativas. Finalmente, en el Capítulo 7, se discuten las principales conclusiones obtenidas del trabajo y se presentan futuros pasos a considerar luego de esta investigación.

2. REVISIÓN BIBLIOGRÁFICA

En este capítulo se hace una pequeña descripción de los modelos de aprendizaje estadístico utilizados en esta tesis. De esa forma, se espera que el lector se familiarice con ellos y entienda a grandes rasgos el funcionamiento de cada uno. A su vez, se presentan los trabajos reportados en la literatura, relacionados con la predicción de velocidades o tiempos de viaje de buses en distintas partes del planeta. Donde se espera dejar claro que se ha hecho hasta el momento y en qué se diferencia este trabajo de lo reportado anteriormente.

2.1. Métodos de aprendizaje estadístico utilizados en este trabajo

Los métodos de aprendizaje estadístico, son métodos que van desde simple cálculo de medias hasta modelos complejos como Redes Neuronales Artificiales. Existen un sin fin de modelos que se pueden clasificar en modelos de aprendizaje supervisado o no. Los primeros son modelos que explican el comportamiento de una variable dependiente en función de un set de variables explicativas, encontrándose modelos de clasificación (predicción de clases) y de regresión (predicción de un valor). Los segundos, en cambio, buscan relaciones entre las variables predictivas y no disponen de una variable dependiente.

Como se verá en la Sección 2.2, se ha utilizado modelos de aprendizaje supervisado tanto de regresión como de clasificación para predecir velocidades o tiempos de viaje de buses. Los primeros han dado mejores resultados, debido probablemente a que los modelos de clasificación tienen que discretizar la variable a predecir (velocidad o tiempo) en distintos rangos, perdiéndose de esa forma información que los modelos de regresión sí disponen. Es por esto, que en esta tesis solo se trabaja con algoritmos de regresión, de los cuales se eligieron tres: Regresión Lineal, Redes Neuronales Artificiales y Maquinas de Soporte Vectorial. Se eligieron estos tres por obtener buenos resultados en la literatura y tener características distintivas, las cuales se explicaran en las secciones siguientes, que los hacen interesantes para un posterior análisis de los resultados.

En las secciones a continuación se explican los tres modelos de regresión empleados. Igualmente para una mayor revisión de los modelos de aprendizaje estadístico se recomienda leer Hastie et al (2001).

2.1.1. Regresión Lineal Múltiple (MLR)

Los modelos de Regresión Lineal, son probablemente los modelos de aprendizaje estadístico más utilizados en el mundo. Esto se debe a una serie de razones: son modelos fáciles de interpretar, rápidos de calibrar, tienen demanda computacional baja, son flexibles, pueden mapear relaciones lineales y no-lineales entre variables, entre una serie de características más (Hastie, Tibshirani, & Friedman, 2001).

Un MLR se puede resumir en la Ecuación 2.1. Donde la variable dependiente Y se puede explicar como un intercepto, más una combinación lineal de variables independientes o explicativas $X^T=(X_1, X_2, .. , X_k)$, y una perturbación aleatoria ε.

$$Y = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \varepsilon \qquad (2.1)$$

Si bien el modelo es lineal, no necesariamente lo son las variables X_k, ya que estas pueden ser transformaciones no lineales como logarítmicas, raíces, polinomios, numéricas o *dummy*, o incluso interacciones entre otras variables ($X_3 = X_1 \cdot X_2$). Esto genera que, a pesar de ser lineal, el modelo pueda expandirse para explicar relaciones no lineales sin perder la simpleza ni la interpretabilidad.

Por otro lado, los valores β_k son los valores que el algoritmo estima utilizando el método de mínimos cuadrados, donde lo que se busca es minimizar la suma de los cuadrados de los residuos, dado por la Ecuación 2.2. Para entender mejor este hecho en la Figura 2-1, se pueden ver las observaciones reales (puntos rojos) y el plano de predicción de la variable Y creado por el método de regresión lineal, dado una combinación lineal de las variables X_1 y X_2, donde los coeficientes β_1 y β_2 fueron encontrados mediante el método de mínimos cuadrados y de esa forma se crea el plano $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \qquad (2.2)$$

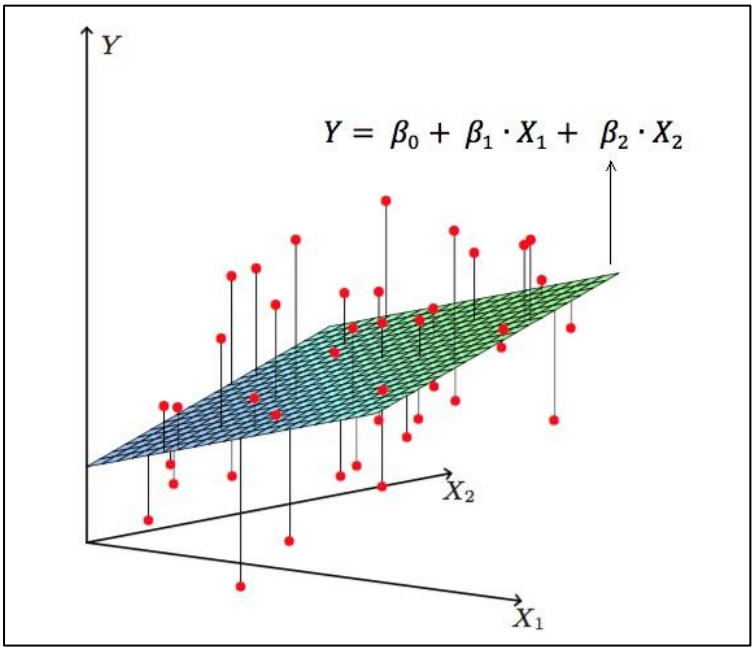


Figura 2 - 1. Ajuste lineal utilizando método de mínimos cuadrados con $X \epsilon R^2$. Se busca la función lineal de X que minimice la suma de los residuos al cuadrado de Y. Fuente: (Hastie et al, 2001)

Para una explicación más detallada de los modelos de Regresión Lineal Múltiple se puede consultar Hastie et al (2001) y James et al (2014).

2.1.2. Máquinas de Soporte Vectorial (SVM)

Las máquinas de soporte vectorial son un tipo de algoritmo de aprendizaje estadístico el cual es capaz de mapear relaciones no lineales entre variables predictivas y las a predecir. Este método fue desarrollado en la década de los 60 (Vapnik y Lerner, 1963; Vapnik y Chervonenkis, 1964), y se ha seguido desarrollando hasta la actualidad, donde sus buenos resultados obtenidos le han dado gran popularidad dentro de los algoritmos de aprendizaje estadístico.

En un comienzo, SVM fue desarrollado como un método de clasificación, donde la idea es separar distintas clases mediante hiperplanos. Sin embargo, las clases no son siempre separables por hiperplanos, como se puede ver en la Figura 2-2, por lo que SVM hace uso de funciones de *Kernels*, las cuales son funciones que logran pasar variables de entrada a otra dimensión (p.ej. pasar de un hiperplano a un plano, o linealizar variables que tienen relaciones no lineales), estas pueden ser funciones lineales, radiales, polinómicas, entre otras (Julio , 2015) . En la Figura 2-2, se observa este hecho, donde el uso de una función de *Kernels*, transformó una separación de clases no lineal a una lineal.

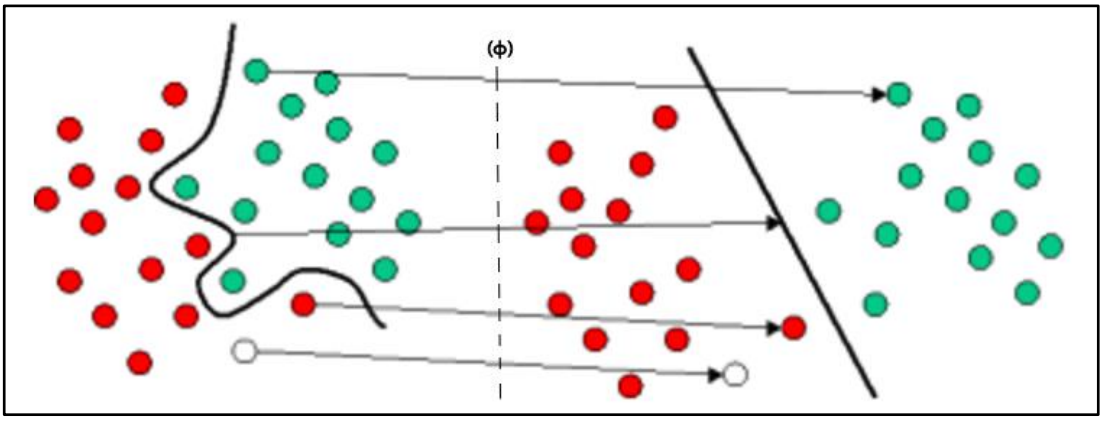


Figura 2- 2. Se muestra como mediante el uso de funciones de Kernel (ϕ) se linealiza el espacio. Fuente: Statistica [imagen] (2016)

Ahora bien, no siempre es posible aislar cada clase, o al hacerlo, podrían surgir problemas de especificación. Es por esto, que SVM es también conocido como clasificador suave de margen, en otras palabras, la separación de las clases puede ser violada dentro de un cierto margen por algunos puntos correspondientes a otras clases, como lo muestra la Figura 2-3. Este hecho genera que las SVM sean más robustas en presencia de datos sucios y a observaciones individuales (James, Witten y Hastie, 2014).

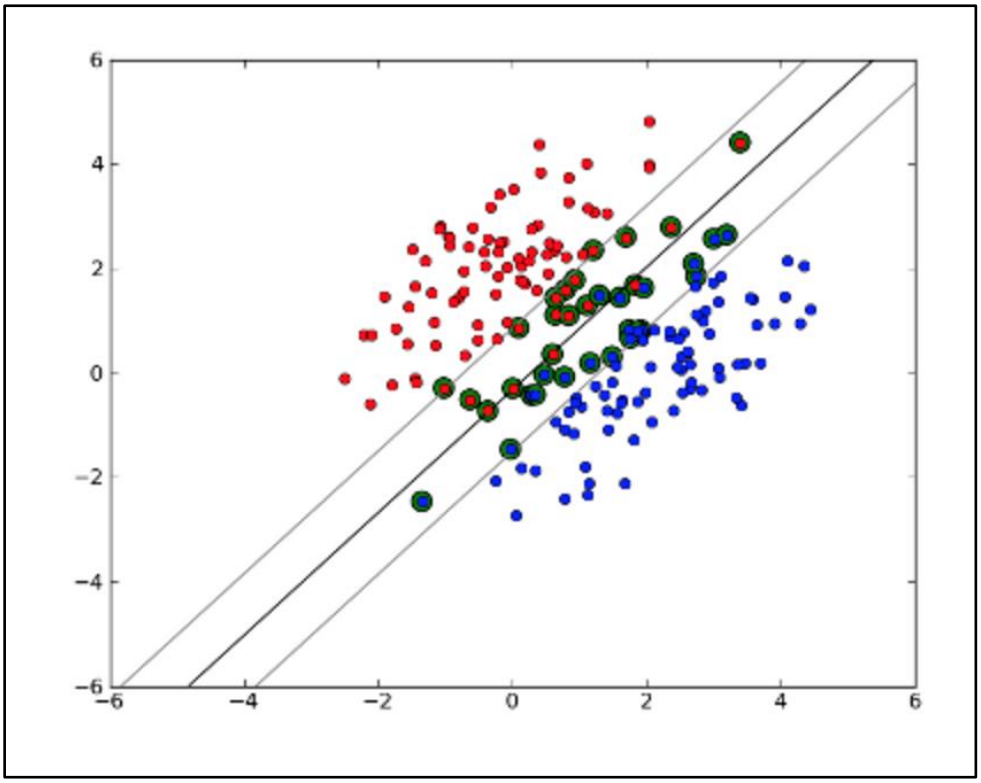


Figura 2- 3: Uso de márgenes suaves debido a clases no separables. Fuente: Soft Margin Linear SVM [imagen] (2012)

Hasta ahora se ha presentado las Máquinas de Soporte Vectorial como un algoritmo de clasificación, sin embargo, se adaptó para que pudiera utilizarse en regresión y ya en la década de los noventa se estaban obteniendo excelentes resultados (Vapnik, 1995). SVM adaptadas para realizar regresiones se les conoce como regresiones de soporte vectorial (SVR o ε-SV), donde en vez de separar clases, la idea es realizar una regresión en un espacio de variables continuas. En la Figura 2-4 se ilustra este hecho, donde la línea sólida es la curva de regresión y las líneas punteadas, separadas a una distancia ε de la línea sólida, son los márgenes, denominados márgenes suaves, donde los puntos al interior de esta se llaman vectores de soporte y no penalizan a la función objetivo. Sin embargo, los puntos al exterior de esta, si penalizan a la función objetivo con un valor igual a ξ, que vendría siendo la distancia entre la línea solida menos el margen ε.

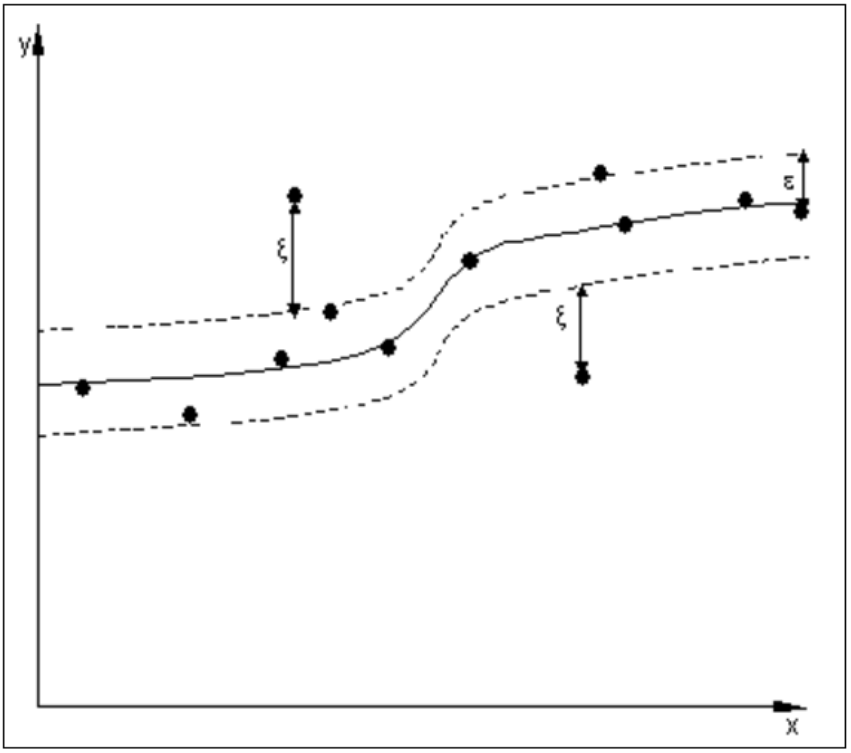


Figura 2 - 4. Los puntos que están dentro del margen de épsilon (entremedio de las líneas punteadas) son llamados vectores de soporte y no afectan la función objetivo, no así los que están fuera de este margen, como los dos puntos que se ven la imagen. Fuente: (Suport Vector Machine Regression, 2016)

Lo que busca el modelo es encontrar la curva de regresión que minimice la función objetivo, y por ende, trata de ajustar la curva que abarca mayor cantidad de puntos dentro de sus márgenes, es decir mayor cantidad de vectores de soporte. Sin embargo, al igual que otros modelos, ajustar una curva a todos los datos no necesariamente da mejores resultados debido a un fenómeno que se lo conoce como sobreajuste, en donde el modelo explica muy bien los datos de entrenamiento, pero no necesariamente los datos de validación; dicho concepto se le conoce también como *trade-off* entre sesgo y varianza. Es por esto que en SVR se ajusta el valor de ε, el cual controla el ancho de los márgenes suaves, y por ende el nivel de sobreajuste y a su vez la penalización C por dejar a puntos fuera de dichos márgenes. En donde la distancia ξ por C vendría siendo la penalización a la función objetivo de cada punto fuera de los márgenes, por lo que determina la forma de la curva de regresión.

Para una mayor comprensión de los modelos de máquinas de soporte vectorial, se puede revisar Hastie et al. (2001) y especificamente para ε-SVR se recomienda Smola y Schölkopf (2004).

2.1.3. Redes Neuronales Artificiales (ANN)

Entre los modelos de aprendizaje estadístico, Redes Neuronales Artificiales (ANN por sus siglas en inglés) es uno de los más explorados gracias a su poderosa capacidad de capturar relaciones no lineales y encontrar patrones en los datos.

ANN es un poderoso algoritmo no lineal inspirado en teorías de cómo funciona el cerebro. El primer modelo fue creado por McCulloch y Pitts (1943), y desde ahí se han desarrollado numerosas variaciones de ANN, que con el mismo trasfondo difieren un poco en su estructura y funcionamiento.

En esta tesis se utiliza una de las variaciones más utilizados de ANN, conocido como *Radial Basis Function Neural Network* (RBFNN), el cual es una red de alimentación directa con una función de activación de base radial, la cual puede aproximar cualquier función continua con exactitud personalizada (Wang, Zuo, & Fu, 2014). RBFNN está compuesta de tres capas: una capa de entrada, una oculta y una de salida, donde la capa de entrada está compuesta por las variables explicativas, la capa oculta, está compuesta por neuronas y es básicamente donde el algoritmo genera las transformaciones de los datos y la capa de salida es simplemente la predicción (ver Figura 2-5). Otras variaciones de ANN están compuestos de más capas ocultas, como *Back Propagation Neural Network*, sin embargo, RBFNN es considerado más efectivo y exacto; puede aproximar casi cualquier relación entre variables de entrada y salida. A su vez, es más rápido para converger y es capaz de evadir problemas de extremo local (Wang, Zuo, & Fu, 2014).

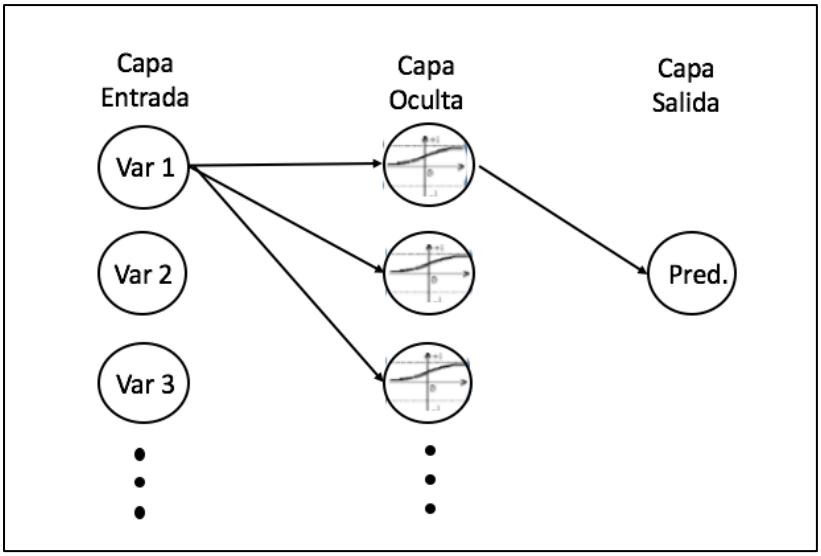


Figura 2- 5. Estructura de un modelo de Red Neuronal.

Como muestra la Figura 2-5, las conexiones entre capas son siempre hacía adelante, donde la primera capa (capa de variables de entrada) simplemente le entrega su valor a cada una de las neuronas en la capa oculta. Estas neuronas en la capa oculta reciben los valores entregados por la capa de entrada, luego los procesan con una función de base radial, para finalmente entregar un valor a la capa de salida (ver Figura 2-6). La capa de salida es una capa de una sola neurona, la cual recibe los valores entregados por cada una de las neuronas en la capa oculta y genera una combinación lineal con ellos, entregando un resultado que vendría siendo la predicción final del modelo (Hastie, Tibshirani, & Friedman, 2001).

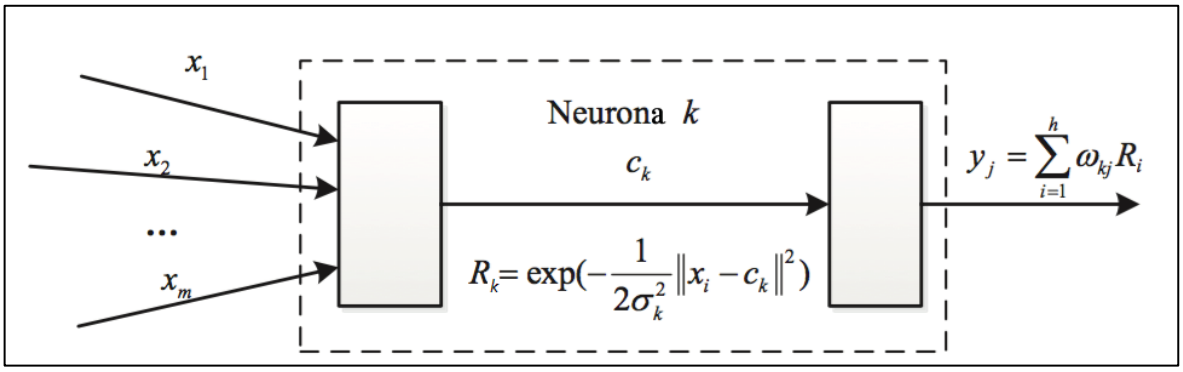


Figura 2- 6. Estructura de una neurona en una Red Neuronal

En la Figura 2-6, I= (x_1, x_2, … , x_m) es el vector de variables de entrada, y O= (y_1, y_2, … , y_n) el de variables de salidas. En regresión, O tiene solo una dimensión (y_1) y esta vendría siendo la predicción. Por otro lado, h es la cantidad de neuronas en la capa oculta, donde i=1,2,..,m, y k= 1,2,..,h. A su vez, $\|x_i - c_k\|$ es la norma Euclidiana, donde x_i es el valor del nodo de entrada i, c_k es el valor céntrico del nodo base k en la capa oculta, y σ_k es la varianza de la función de Gauss. Por último w_{kj} es el valor de los pesos entre la capa oculta k y la variable de salida j. La función de activación genera el cambio a una distancia euclidiana en vez de una lineal.

Para mayor detalles de cómo funcionan las Redes Neuronales Artificiales y específicamente RBFNN, se recomienda leer Hastie et al (2001), Lu, Sundurarajan y Saratchandra (1998); y Ripley(1996).

2.2. Trabajos sobre predicción de velocidades o tiempos de viaje de buses reportados en la literatura

Con el crecimiento de la tecnología y la mayor facilidad para obtener datos, los modelos para predecir velocidades (tiempos de viaje) de los buses, han ido en un constante aumento. A su vez, el aumento del poder computacional ha ayudado a que modelos de aprendizaje estadístico tengan cada vez un rol más importante a la hora de realizar predicciones. Es por esto que, desde el principio del siglo, se han reportado varios trabajos en este ámbito, probando distintos tipos de modelos con diferentes sets de datos. Sin embargo, hasta ahora no hay un consenso de qué modelo es mejor. Una razón para ello es la diferencia en los sets de datos, donde algunos trabajos ocupan datos de GPS, otros de espiras, videos, recolección manual, etc. Por lo que modelos de aprendizaje estadístico, los cuales son sensibles al tipo de datos, no necesariamente tienen los mismos resultados entre distintos estudios.

La Tabla 2-1 muestra un resumen de los principales trabajos reportados hasta ahora en el área de predicción de velocidades o tiempos de viaje de buses. A su vez, en las columnas, se especifica el tipo de datos, el o los modelos ocupados y lo que se busca

predecir. En negrita se destaca cual es el modelo que predice mejor para el set de datos correspondiente.

La columna de Fuente de Información muestra con qué tipo de datos trabaja el autor. Donde GPS vendría siendo *Global Positioning System*, sistema el cual muestra la posición geográfica de un bus en un determinado momento. RFID, en cambio, es un identificador de radio frecuencia, el cual almacena y recupera información remota. Donde cada vehículo tiene un identificador único y con eso se registra información relevante. Por último, APC viene de *Automatic Passenger Counting*, sistema diseñado para contar pasajeros viajando en un determinado momento en un bus.

La columna Métodos de aprendizaje estadístico muestra los métodos utilizados por cada autor. Donde ANN vendría siendo Redes Neuronales Artificiales; SVR, Máquinas de Soporte de Regresión; y MLR, Regresión Lineal Múltiple. En "Otros", se especifica si se utilizó alguno otro modelo como K-Nearest Neighbord (KNN), Filtros de Kalman (Kalman), o modelos base con información histórica o en tiempo real

Tabla 2 - 1. Resumen trabajos sobre predicción de velocidades o tiempos de viaje de buses reportados en la literatura

Trabajo	**Fuente información**	***Input***	**Método de aprendizaje estadístico**				***Output***
			ANN	**SVM**	**MLR**	**Otros**	
Jeong & Rilett (2004)	GPS	Ajuste a itinerario, tiempo abordaje	**X**		x	Histórico	Tiempo llegada
Mazlouni et al. (2010)	GPS+Espira+Clima	Velocidades	**X**				Tiempo de viaje
Coffey et al. (2011)	GPS	Tiempos de viaje				**KNN**	Tiempo de llegada
Mazloumi et al. (2011)	GPS + Espiras	Tiempos de viaje, flujos	**X**				Tiempo de viaje + variabilidad
Yu et al. (2011)	Recolección manual	Tiempos viaje	x	**X**	x	KNN	Tiempo llegada
Xingao et al. (2013)	GPS + RFID	Velocidad histórica y en tiempo real			**X**	Histórico	Tiempo de viaje
Gurmu et al. (2014)	GPS	Tiempos de viaje	**X**			Histórico	Tiempo de llegada
Kumar et al. (2014)	GPS	Velocidades	**X**			Kalman	Tiempo de llegada
Wang et al. (2014)	GPS	Velocidad histórica y en tiempo real	**X**		x		Tiempo de llegada
Zhong et al. (2015)	Recolección manual	Velocidades históricas y en tiempo real según segmento y horario	x	**X**		Series de tiempo	Tiempo de llegada
Julio et al. (2016)	GPS	Velocidad histórica y en tiempo real para un segmento y horario	**X**	x		Reds Bayesianas	Velocidad en segmento siguiente
Rahman et al. (2016)	GPS + APC + Clima	Velocidades				**KNN** y Tiempo Real	Tiempo de llegada

Jeong & Rillet (2004) es de los primeros en comparar distintos modelos de aprendizaje estadístico, al querer predecir tiempos de llegada a paraderos, usando como variables de entrada la hora de llegada al paradero anterior (datos obtenido por el GPS), adhesión al itinerario, y tiempos de parada y subida de pasajeros. Para las predicciones ocupan modelos ANN y regresión lineal, y los comparan con un modelo simple de velocidades históricos. Concluyen que las predicciones utilizando ANN superan en todos los casos a los otros modelos.

Mazlouni et al (2010) y Mazlouni et al (2011), utilizan datos de tráfico, demanda de pasajeros y clima para predecir tiempos de viaje en segmentos y su variabilidad. A diferencia de los otros papers ellos predicen un intervalo con 95% de confianza de tiempo de viaje entre paraderos. Utilizan dos modelos de Redes Neuronales, donde uno predice tiempos y el otro la varianza. Ellos concluyen que hay que utilizar datos de tráfico agrupados en intervalos de 15 minutos previos a la salida de los buses de los paraderos, con eso tienen datos recientes y suficientes para estimar el estado de tráfico. Por último, retiran la variable clima del modelo, ya que no tiene mayor impacto en las predicciones, sin embargo, al no tener suficientes datos de esta no pueden concluir que no afecte a los tiempos de viaje.

Coffey et al (2011) analiza un servicio en la ciudad de Dublín. Lo que se busca es predecir tiempos de llegada en función de datos históricos de GPS. Solo se propone el modelo KNN, donde se muestra cómo varía la elección del parámetro *k*, es decir la cantidad de vecinos a considerar en el modelo, según la distancia al paradero que se quiere predecir la llegada. Ellos concluyen que predicciones de distancias cortas y largas tienen mucha variabilidad e incertidumbre, por lo que se debería predecir un intervalo en vez de un tiempo exacto. No así con distancias medianas, donde las predicciones mostraron ser más certeras.

Yu et al (2011), propone también, una serie de modelos de aprendizaje estadístico para predecir tiempos de viaje entre dos puntos con múltiples rutas posibles. Las variables de entrada provienen de recopilación de tiempos de viaje de buses obtenidas mediante

encuestas, y las utilizan para calibrar modelos de ANN, SVM, MLR y KNN. Encuentran que SVM predice mejor que los otros modelos, y que la utilización de múltiples rutas en vez de rutas simples, ayuda a las predicciones.

Xingao et al (2013) ocupan datos de GPS e identificación de radio frecuencia (RFID), tanto de buses como de taxis, para predecir tiempos de viaje de buses. Utilizan un modelo de Regresión Lineal Múltiple, donde las variables de input son datos tanto en tiempo real como histórico de las velocidades de buses y taxis. Le agregan velocidades de taxis para ayudar a generar mejores predicciones, ya que aumentan la cantidad de datos recopilados en todo momento, y a su vez las velocidades de estos se relacionan linealmente con la de los buses. En el paper se concluye que MLR predice mejor los tiempos de viaje que ocupar tiempos promedios históricos.

Gurmu et al (2014) ocupan datos de GPS de buses para predecir tiempos de viaje desde la ubicación actual del bus a cualquier paradero aguas abajo. Para realizar estas predicciones ellos utilizan dos métodos: ANN y promedio históricos de velocidades, donde se concluye que, tanto en precisión de las predicciones como en robustez, ANN es mejor modelo. Además, concluyen que para tiempos cortos o muy largas, ANN no predice tan bien en relación de promedios históricos, sin embargo para tiempos medianos (alrededor de 30 min de viaje) el modelo tiene errores MAPE de solo un 10%.

Kumar et al (2014) ocupan datos históricos de GPS para predecir tiempos de viaje en un servicio de bus en India. Las predicciones las hacen con dos métodos: Redes Neuronales y Filtros de Kalman (KFT). Concluyen que en presencia de muchos datos ANN es mejor para realizar predicciones que KFT. Sin embargo, cuando se tienen pocos datos esta última es muy útil.

Wang et al (2014) ocupan información histórica así como en tiempo real de GPS de buses incluyendo las subidas y las bajadas de pasajeros, para predecir hora de llegada a los paraderos. Para realizar dichas predicciones, ellos utilizan Regresión Lineal Múltiple (MLR) y Redes Neuronales (ANN). Con este último modelo ellos proponen un ajuste que consiste en introducir un método orientado en línea, con el cual, ocupando filtros de

Kalman, logran filtrar de mejor manera la información en tiempo real de la posición del bus para predecir velocidades y por ende tiempos, al siguiente paradero. Del paper se concluye que ANN ajustado es el mejor modelo para predecir tiempos de llegada, con errores promedio mucho menor a MLR y ANN no ajustado.

Zhong et al (2015) calibran modelos SVM, con información histórica y en tiempo real de datos obtenidos mediante encuestas, para poder predecir tiempos de llegada a los paraderos. El modelo SVMAA consiste en primero eliminar los datos *outliers* y luego calibrar un SVM, más un modelo adaptativo. Los *outliers* se eliminan para no considerar en las predicciones datos tomados en situaciones anormales o directamente mal tomados, y a su vez, el modelo adaptativo se incorpora para poder utilizar información en tiempo real de la posición del bus y distancia hacia la siguiente parada. En el paper se hacen comparaciones con otros modelos como ANN, modelo de series de tiempo y SVM, concluyendo que en la mayoría de las predicciones SVMAA supera al resto. ANN lo hace bastante bien superando en algunas predicciones a SVMAA. Sin embargo, en promedio esta tiene un MAPE un poco superior.

Julio et al (2016) buscan predecir velocidades de viaje de buses del Transantiago en la ciudad de Santiago, Chile. Para eso utilizan modelos ANN, SVM y redes bayesianas, donde como variables de input ocupan datos de GPS para obtener velocidades tanto históricas como en tiempo real. El paper solo utiliza variables de velocidad, lo que lo diferencia de esta tesis ya que además de ocupar dichas variables se suman otras, como de clima, infraestructura y comportamiento de pasajeros. A su vez, en esta tesis se trabaja también con Regresión Lineal, el cual es un modelo simple, rápido de calibrar y fácil de interpretar, lo que lo hace muy útil a la hora de querer medir el impacto de las distintas variables en las predicciones.

Rahman et al (2016) buscan cuál es el mejor intervalo de tiempo para obtener la posición del GPS del bus a la hora de querer realizar predicciones. A su vez, estiman a qué distancia conviene utilizar modelos simples de velocidades históricas, versus modelos basados en velocidades en tiempo real para predecir horas de llegada. Encuentra que 30

segundos de intervalo de tiempo entre pulsos de GPS es el tiempo adecuado, ya que intervalos menores no llegan a resultados significativamente mejores.

De los distintos trabajos reportados en la literatura, se pueden obtener dos conclusiones que serán útiles para proseguir con este trabajo. En primer lugar, no se puede determinar qué modelo de aprendizaje estadístico predice de mejor forma, por lo que es importante establecer qué modelo es mejor para Santiago de Chile. Esto se puede deber a que este tipo de modelos dependen plenamente del set de datos que se tenga. Por lo que al tenerse múltiples formas de obtener datos (GPS, espiras, recolección manual, entre otras), hace que los datos no sean comparables entre un trabajo y otro. A su vez, las investigaciones son reportadas en ciudades a lo largo de todo el mundo, lo que genera aún más diferencias entre los set de datos, ya que podrían haber factores determinantes en una ciudad y no en otra, como podría ser la lluvia, donde en una ciudad con un buen sistema de drenaje esta no tendría mayor implicancia en las predicciones, pero en una ciudad con un mal sistema de drenaje sí.

La segunda conclusión, tiene que ver con el segundo objetivo de este trabajo, el cual vendría siendo el impacto de las distintas variables explicativas a la hora de realizar predicciones. La mayoría de los trabajos reportados en la literatura solo utilizan variables explicativas de velocidad, y en caso de ocupar más, no se habla del impacto que estas tienen sobre las predicciones. Por lo que va a ser un aspecto importante en la literatura la medición del impacto en las predicciones al añadir o sustraer distintas variables explicativas. De esa forma se dejará un precedente de las variables más importantes a la hora de realizar predicciones de velocidades de buses, así como las que no tienen un mayor impacto en estas.

3. DEFINICIÓN DEL PROBLEMA

Este capítulo es una introducción al caso de estudio, a las metodologías y modelos utilizados. Se define el problema que se aborda en la tesis, y se da una explicación de porqué se predice velocidades y no tiempos de viaje.

3.1. Definición del problema

Como se mencionó en los capítulos anteriores, predecir velocidades de viaje (tiempos) de los buses puede ser muy beneficioso tanto para los operarios del transporte, como para los usuarios. Para los operarios puede significar mejorar la calidad de servicio del transporte público, monitoreando la ejecución de los itinerarios, mejorando la frecuencia efectiva de pasada de los buses y evaluando la eficiencia operacional (Wang, Zuo, & Fu, 2014). Para los usuarios, en cambio, la predicción de velocidades de buses les puede ayudar a planificar mejor sus rutas, disminuir los tiempos de espera en los paraderos, y eliminar la ansiedad por estar esperando. Esto no solo lo hace más atractivo como modo de transporte, sino que también permite aumentar la satisfacción y disposición al pago (Dziekan y Kottenhoff, 2007).

Santiago de Chile dispone de una red de transporte público llamada Transantiago, la cual tiene alrededor de 6.500 buses en servicio (Cruz, 2015). A su vez, el transporte público en Chile representa buena parte de los viajes realizados, donde solo en bus se realizan más de 3,5 millones de viajes diarios de un total de 18 millones equivalente a un 20% aproximadamente (Encuesta Origen-Destino de Viajes de Santiago, 2015). Es por esto que predecir velocidades de buses puede tener un fuerte impacto en la red de transporte de la ciudad, mejorando tiempos, eficiencia y calidad del transporte.

Los buses del Transantiago presentan un dispositivo GPS que emite una señal cada 30 segundos con la posición exacta de cada bus. Esto implica recibir 120 pulsos por hora por bus. Como cantidad de información no pareciera ser mucho, sin embargo, cuando se quieren generar modelos para predecir velocidades en todos los recorridos

de la red, y para lo cual se requiere de semanas de recolección de datos, la cantidad de datos pasa a ser un problema a la hora de calibrar modelos complejos como Máquinas de Soporte Vectorial o Redes Neuronales Artificiales. No así con métodos más simples como Regresión Lineal Múltiple, que no solo es más fácil de entender, sino que también es muy rápido de calibrar, por lo cual es un método valioso si se quisiera implementar de forma rápida y a gran escala.

Por otro lado, hoy en día se dispone de una gran cantidad de variables explicativas, tanto en tiempo real como histórico, entre ellas velocidades de buses, infraestructura, comportamientos de pasajeros, etc., las cuáles serán vistas con mayor detalle en las secciones siguientes. A pesar de que se dispone de estas variables, a nivel mundial poco se ha hablado de cómo influye cada una de ellas en las predicciones. A nivel chileno, Julio et al (2016) utilizan las velocidades entregadas por GPS para realizar predicciones, sin introducir ninguna variable más. Es por esto que esta tesis abarca la inclusión de estas nuevas variables explicativas y busca dejar un precedente en la importancia de cada una de ellas a la hora de realizar las predicciones.

3.2. Porqué predecir velocidad y no tiempo de viaje

En el Capítulo 2 se pudo ver que en la literatura, si bien, se predice velocidades de viaje de los buses, generalmente se hace con los tiempos. No hay un consenso de que es mejor, sino que depende de la base de datos que se tenga y del output con que se quiera trabajar.

En nuestro caso, la base de datos disponible no tiene las horas exacta de llegada o salida de los buses a los paraderos, sino que se tiene una hora estimada de pasada basada en interpolaciones de las posiciones reportadas por los pulsos de GPS cada 30 segundos. Sí la posición reportada por el pulso del GPS calza con el paradero, entonces se le asigna esa hora como la hora de pasada por este. En caso que no calce, se hace una interpolación en base al tiempo transcurrido desde el último pulso (30 segundos) y la distancia recorrida desde que se pasó el último paradero. En el paradero un bus puede estar varios segundos, sin embargo, la base de datos solo

incluye una estimación de en qué momento se estuvo, pero no se puede determinar si esta estimación es justo antes de salir, apenas llegó o mientras estuvo. Esta razón hace que sea más atractivo trabajar con predicciones de velocidades. En la Ecuación 3.1, se puede ver como se construye la variable de velocidad, donde *p* es el paradero. Por su construcción la velocidad tiene implícita la distancia, lo cual hace posible y entendible comparar dos velocidad independiente del tramo que se esté prediciendo. Sin embargo, el tiempo por sí solo no tiene mucho sentido, ya que no es posible comparar tiempos sin disponer de las distancias.

$$Velocidad\ tramo_{[p,p+1]} = \frac{Distancia_{[p\ ,p+1]}}{(Hora\ pasada_{p+1} - Hora\ pasada_{p})} \qquad (3.1)$$

Además, la medida principal de desempeño que se utiliza para calibrar y elegir los modelos, castiga al cuadrado los errores, por lo que creemos que se van a cometer errores menores al predecir valores altos, ya sea de velocidades o tiempos. Como las posiciones de los buses son reportadas cada 30 segundos por los GPS, en tiempos de viaje alto se puede ir ajustando la predicción en función de la realidad y lo esperado en el tramo que lleva. Sin embargo, para tiempos de viaje pequeños, se tienen menos instancias para ajustar los tiempos predichos, por lo que creemos que al predecir velocidades se van a cometer errores menores en velocidades altas y por ende tiempos de viaje pequeño, a que si se predijera tiempos de viaje directamente. Es por esta razón que en esta tesis se trabaja con velocidades de viaje de los buses y no tiempos. Además tiene la ventaja de que la comprensión del análisis se hace más fácil para el lector cuando se trabaja con velocidades debido a que se tiene puntos de comparación con las experiencias cotidianas en el transporte.

4. CASO DE ESTUDIO Y VARIABLES EXPLICATIVAS INVOLUCRADAS

Como se mencionó anteriormente, el lugar físico de donde se obtienen los datos puede ser determinante a la hora de comparar resultados obtenidos por distintos modelos de aprendizaje estadístico. Por lo tanto, obtener resultados y conclusiones de un servicio de buses específico podría no ser aplicable al resto. Es por esto que en esta tesis, para evaluar los resultados y robustez de forma heterogénea, se prueban los modelos con tres servicios de buses de distintas características.

En este capítulo se explican la elección de estos tres servicios y se muestran las características generales de cada uno de ellos. También se presenta el espectro de variables disponibles para realizar las predicciones, así como la elección del set final de estas. Se introduce un concepto importante para el lector, el *headway* (por su nombre en inglés), que hace referencia al intervalo de tiempo entre la pasada de dos buses consecutivos de un mismo recorrido por un punto.

4.1. Servicios analizados: Elección y características generales

Las tres líneas analizadas en esta tesis son la 212, 203 y C04. Se eligieron estas líneas por sus distintas características en términos de tipo de vías utilizadas, tamaño de los buses y frecuencias de pasada, las cuales las hacen ser muy diferente entre ellas, al mismo tiempo que abarcan el espectro casi completo de tipos de líneas del Transantiago. A su vez, para reducir la cantidad de datos totales y de esa forma poder calibrar más fácil los modelos, se elige solo servicios ida o vuelta, donde los servicios finales analizados, detallados en las secciones a continuación, son el 212 retorno, 203 retorno y C04 ida.

4.1.1. 212R

Este es un clásico servicio troncal del Transantiago. Durante su recorrido pasa por vías comunes, vías exclusivas y corredores segregados. En la Figura 4-1 a continuación se puede observar el recorrido que realiza en la ciudad de Santiago.

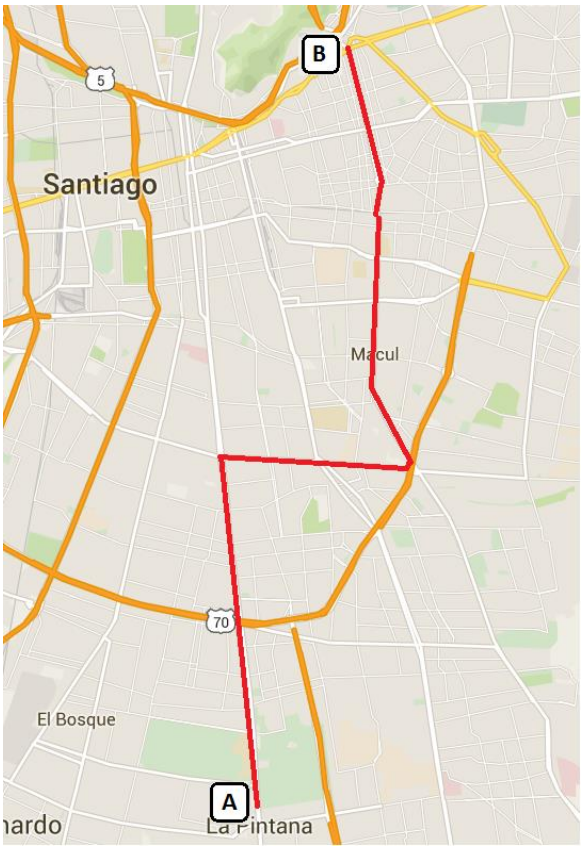


Figura 4 - 1. Recorrido servicio 212 retorno.

En la imagen se observa que este servicio parte en la comuna de *La Pintana* (punto A, en la Figura 4-1), luego pasa por 6 comunas más para terminar en la comuna de *Providencia* (punto B). La distancia total del recorrido es de 28 km, donde 6,7 km los realiza por corredores segregados, 13,8 km por vías exclusivas y los restantes 7,5 km por vías comunes. Los paraderos están separados en promedio cada 350 metros. A su vez, la frecuencia promedio es de 7 buses por hora, equivalente a un bus cada 8,5 minutos, sin embargo, en horas punta alcanza frecuencias de hasta 12 buses por hora. Por último, la flota de este servicio está compuesta por buses de mediano y gran tamaño, con capacidades que oscilan entre los 90 y 160 pasajeros por bus.

En la Figura 4-2, se puede ver un mapa de calor de las velocidades promedio del recorrido, donde el color verde representa velocidades altas y el rojo las bajas. Este servicio opera entre las 5 am y la media noche. Los valores de las velocidades

promedio van desde 8 km/hr hasta 41 km/hr, sin embargo, individualmente el rango de velocidades de los buses es desde 2,5 km/hr hasta 69 km/hr, con un promedio de 22,3 km/hr en el recorrido. A su vez, se observa una congestión que viene dada más por el tramo que por la hora del día, donde, el tramo a mitad del recorrido tiene velocidad promedio muy baja prácticamente todo el día. Esto se debe a que es un trayecto corto con muchos virajes, lo que genera que en ese tramo hayan las menores velocidades promedio de todo el recorrido.

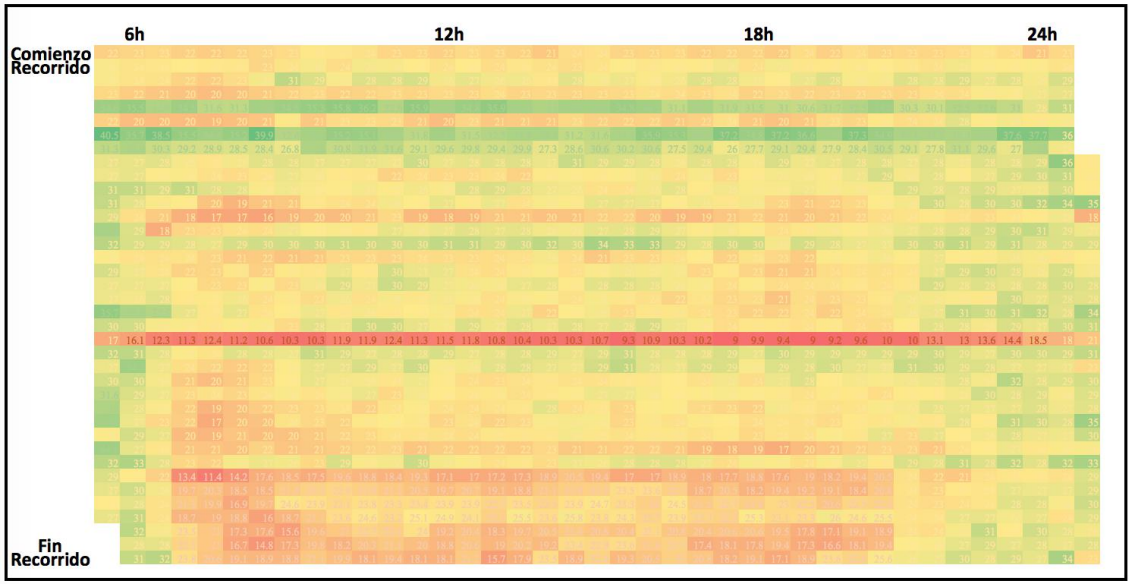


Figura 4 - 2. Mapa de calor de velocidades históricas del servicio 212R

4.1.2. 203R

El servicio 203R es un servicio troncal largo, con un total de 30 km de extensión y paraderos cada 315 metros en promedio. De la extensión total del recorrido 6 km los realiza en corredores segregados, 6,8 km en vías exclusivas y los restantes 17,2 km en vías comunes. En la Figura 4-3 se puede observar el mapa del recorrido que realiza, donde este pasa por 8 comunas, partiendo en la *Pintana* y terminando en *Huechuraba*. Este es uno de los servicios con mayor frecuencia del Transantiago, donde los buses pasan en promedio cada 6 minutos, equivalente a 10 buses por hora. Sin embargo, en horas punta su frecuencia sobrepasa los 20 buses por hora,

equivalente a *headways* promedio de menos de 3 minutos. Al igual que el recorrido 212R, la flota está compuesta de buses de tamaño mediano y grande, con capacidades de 90 hasta 160 pasajeros.

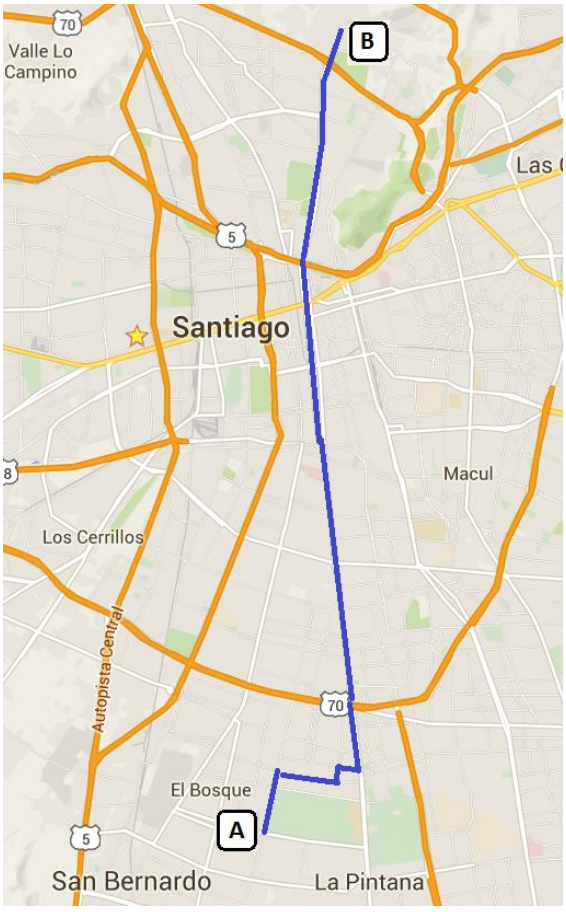


Figura 4 - 3. Recorrido servicio 203 retorno

En la Figura 4-4 se puede observar el mapa de calor de las velocidades promedio a lo largo del día y de su recorrido. Los tramos tienen velocidades en promedio entre los 9 km/hr hasta 38 km/hr; sin embargo, individualmente los buses alcanzan velocidades desde 3 km/hr hasta 69 km/hr, con un promedio total de 21,9 km/hr. A su vez, de la figura se observa que en general las velocidades promedio no varían mucho durante el día, donde ciertos tramos de la red presentan una congestión

prácticamente constante, versus otros que tienen velocidades promedio altas independiente de la hora.

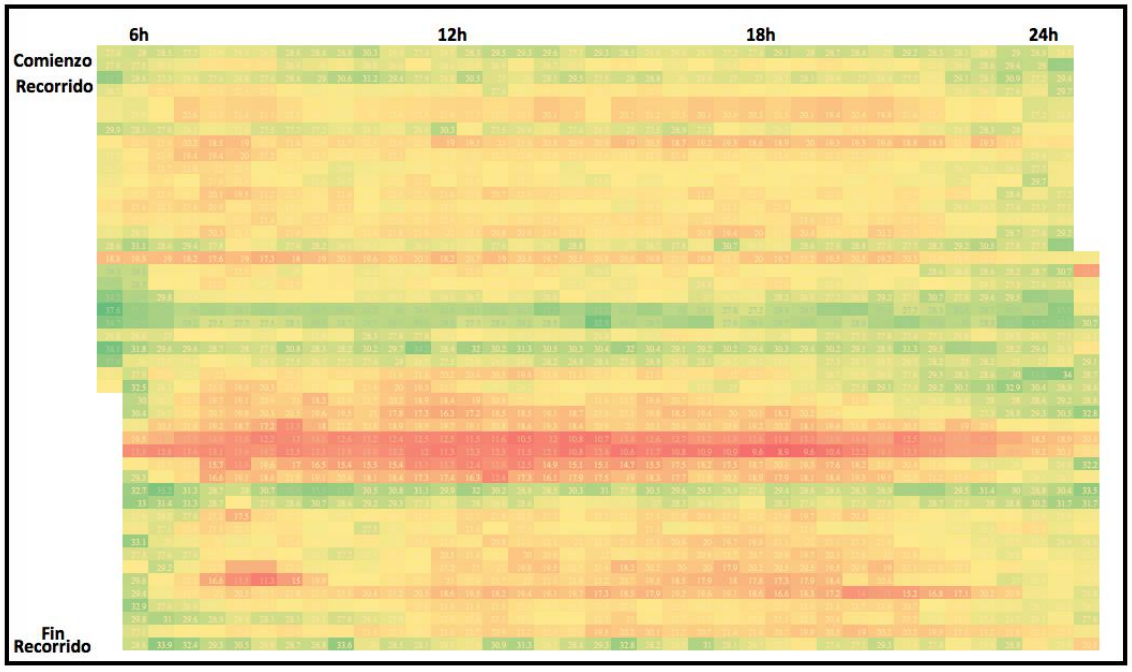


Figura 4 - 4. Mapa de calor de velocidades de servicio 203R

4.1.3. C04I

A diferencia de los otros dos recorridos, la C04 es una línea que recorre solo la zona oriente de Santiago, con un recorrido de 12 km de largo realizado solo en vías comunes. En promedio el distanciamiento de los paraderos es de 270 metros, menor a los otros recorridos. En la Figura 4-5, se puede ver un mapa con el recorrido del bus, que parte en la comuna de *Las Condes* y termina en *Providencia*.

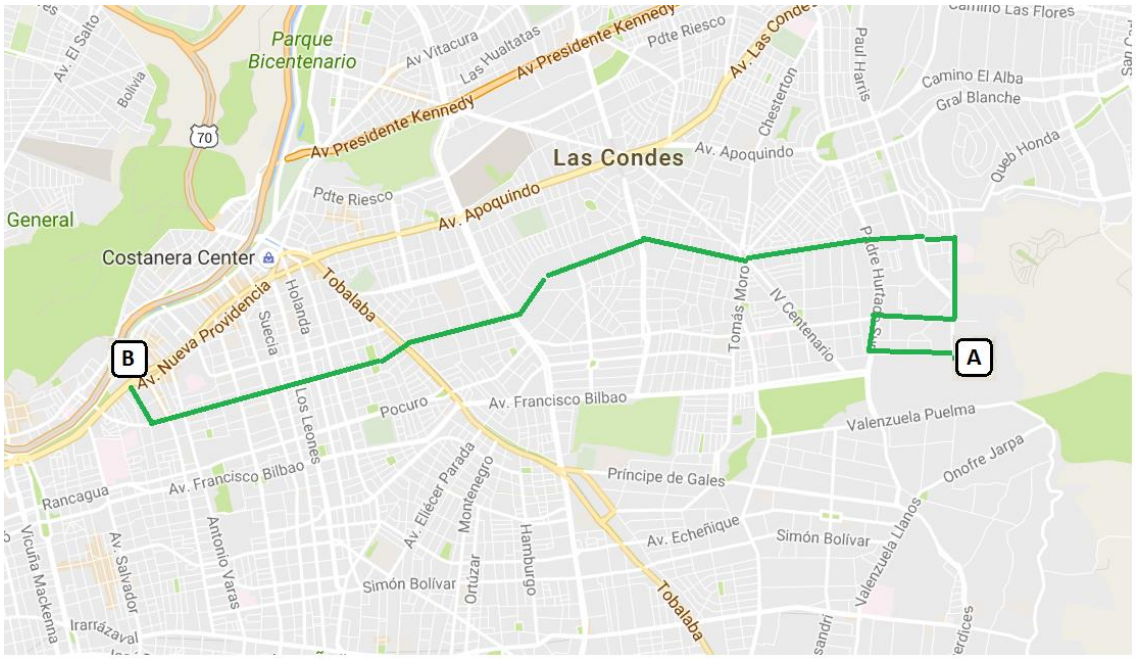


Figura 4 - 5. Recorrido servicio C04 ida

Este recorrido tiene una frecuencia mucho menor a los otros dos, con 3,5 buses por hora en promedio, equivalente a un bus cada 17 minutos. A su vez, la flota está compuesta por buses de tamaño pequeño y mediano, con capacidades que van desde los 60 a 90 pasajeros por bus.

En la Figura 4-6 se puede observar una mayor diferencia de velocidades promedio en función de la hora, donde claramente se distingue una hora punta mañana, otra punta medio día y por último una punta tarde, donde la última presenta las menores velocidades promedió y por ende mayor congestión. A su vez, se observa que hay tramos más congestionados que otros, sobre todo en tramos al principio del recorrido, otro en la mitad y algunos al final de este. Donde las velocidades de los buses en los tramos del recorrido van desde los 4 km/hr hasta los 60 km/hr con un promedio de 20,8 km/hr.

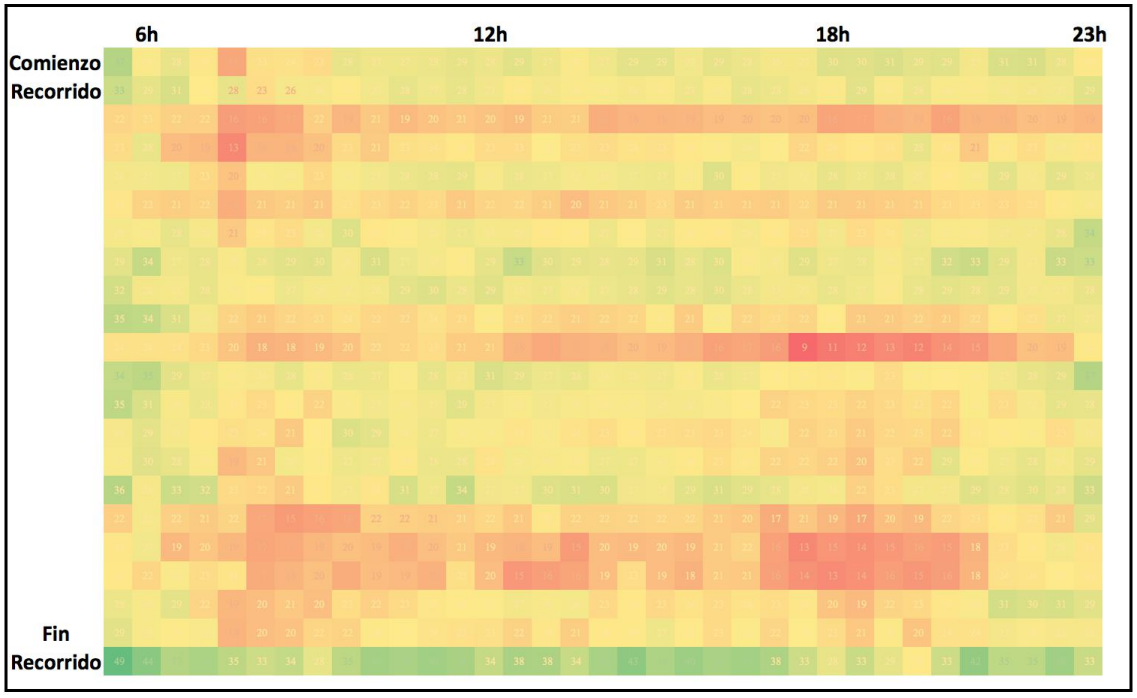


Figura 4 - 6. Mapa de calor de velocidades del servicio C04I

4.1.4. Comparación entre servicios

En la Tabla 4-1 se puede observar las características de las velocidades de los recorridos analizados, donde vemos que la velocidad promedio del recorrido 212R es la mayor, lo que probablemente se explica por la mayor proporción de su recorrido en corredores segregados. A su vez, el recorrido C04I es el que tiene la menor velocidad promedio. Este no pasa por corredores segregados, ni vías exclusivas de buses, lo que podría explicar este hecho. En cuanto a la desviación estándar y al coeficiente de variación, en los tres recorridos dieron resultados parecidos. Sin embargo, el recorrido 212R da resultados ligeramente menores de coeficiente de variación, por lo que se tiende a pensar que este va a dar resultados un poco mejores con los métodos de aprendizaje estadístico.

Tabla 4-1: Estadísticas de velocidad del set de datos final según el servicio

	212R	**203R**	**C04I**
Media (km/hr)	22,3	21,9	20,8
Desviación Estándar (km/hr)	8,2	8,3	7,9
Coeficiente de Variación	36,8%	38,0%	38,1%

En la Tabla 4-2 se puede observar características de los tiempos entre buses; es decir los *headways*, entre los distintos recorridos analizados. Estos van a ser importantes en la elección de algunas variables explicativas de las cuales se habla en las secciones siguientes. De la tabla se observa que los servicios 212R y 203R tienen en promedio *headways* pequeños, cercanos a 7 minutos, con una desviación estándar de más o menos del mismo valor. No así el recorrido C04I, el cual tiene un *headway* promedio mucho mayor, con un valor de 17,7 minutos y desviación estándar parecida a los otros, igual a 7,4 minutos.

Tabla 4 - 2: Estadístico de los *headways* en los distintos servicio

	212R	**203R**	**C04I**
Media (min)	7,3	5,6	17,7
Desviación Estándar (min)	7,1	6,2	7,4

4.2. Set de datos

Los modelos de aprendizaje estadístico son modelos basados en los datos, es decir no son obtenidos desde las teorías ni relaciones del tráfico, sino que por los datos mismos. Lo positivo de este hecho es que no es necesario ser un experto en teoría de tráfico, y lo negativo es que se requiere una gran cantidad y calidad de datos, los que no siempre están disponibles y que además ligan los resultados a un cierto lugar de estudió (Van Lint, 2004). Es por ello que en las secciones siguientes se presentará la

base de datos, con los distintos tipos de variables, así como con el diagrama espacio-tiempo que es fundamental para entender las variables de velocidad.

La base de datos con que se trabajó contiene cuatro tipos de variables: (i) posiciones reportadas por GPS, (ii) factores de demanda, (iii) infraestructura, y (iv) factores de entorno. De las posiciones reportadas por los GPS de los buses se obtienen las velocidades tanto históricas como en tiempo real. En las variables de comportamiento de demanda, encontramos las subidas pagadas y no pagadas, las bajadas, y la carga del bus. Por otro lado, se trabajó con dos variables de infraestructura: la distancia entre paraderos, y la cualidad de la calle, en otras palabras, si es corredor segregado o no el tramo a analizar. Por último, se trabajó con dos variables de entorno: la cantidad de milímetros de agua caída en las últimas horas, y si era de día o no. En las secciones siguientes se abordan con mayor detalle cada una de ellas y como se incluyeron estas en los modelos de aprendizaje estadístico.

4.2.1. Diagrama X-t

El diagrama espacio-tiempo, es uno de los diagramas fundamentales para entender las relaciones y estados de tráfico, y va a ser fundamental para entender la creación y posterior elección de las variables explicativas de velocidad. En la Figura 4-7 se muestra un diagrama espacio-tiempo. En ella se puede ver la trayectoria a lo largo del recorrido de cada uno de los buses que operan en ese servicio en un día.

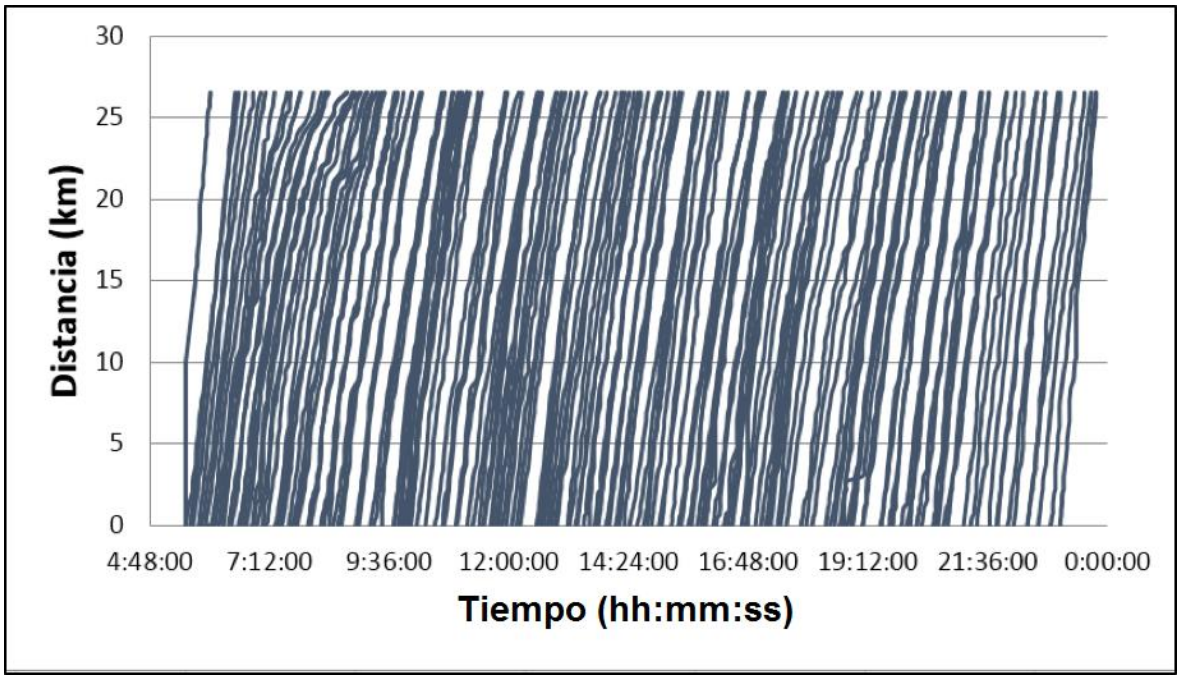


Figura 4 - 7. Diagrama espacio-tiempo del recorrido 212

En esta figura cada línea representa la ubicación espacio-temporal de un bus, mientras que la pendiente de esta representa la velocidad. Cuando dos o más líneas se cruzan o se juntan implica que dos o más buses se cruzaron o juntaron, y a su vez, la separación horizontal entre ellas representa el *headway* entre dos buses, en otras palabras; el tiempo que le lleva un bus al otro.

4.2.2. Posibles variables predictivas

A continuación, se explican las posibles variables explicativas que se crearon a partir de la base de datos. Son posibles ya que no todas mostraron ser significativas y algunas hasta empeoraron las predicciones.

a) Variables de Velocidad y la Grilla

La denominada “Grilla” utilizada por Julio et al (2016) para crear variables explicativas de velocidad para su posterior introducción a los metodos de aprendizaje estadístico, se deriva directamente del diagrama espacio-tiempo. Se cuadricula el digrama de manera de obtener rectángulos de un cierto alto (distancia) y un cierto

largo (tiempo). Si se le hace un acercamiento, se puede apreciar algo semejante a la Figura 4-8, donde en ella se puede observar cómo por cada rectángulo pasan distintas trayectorias de buses.

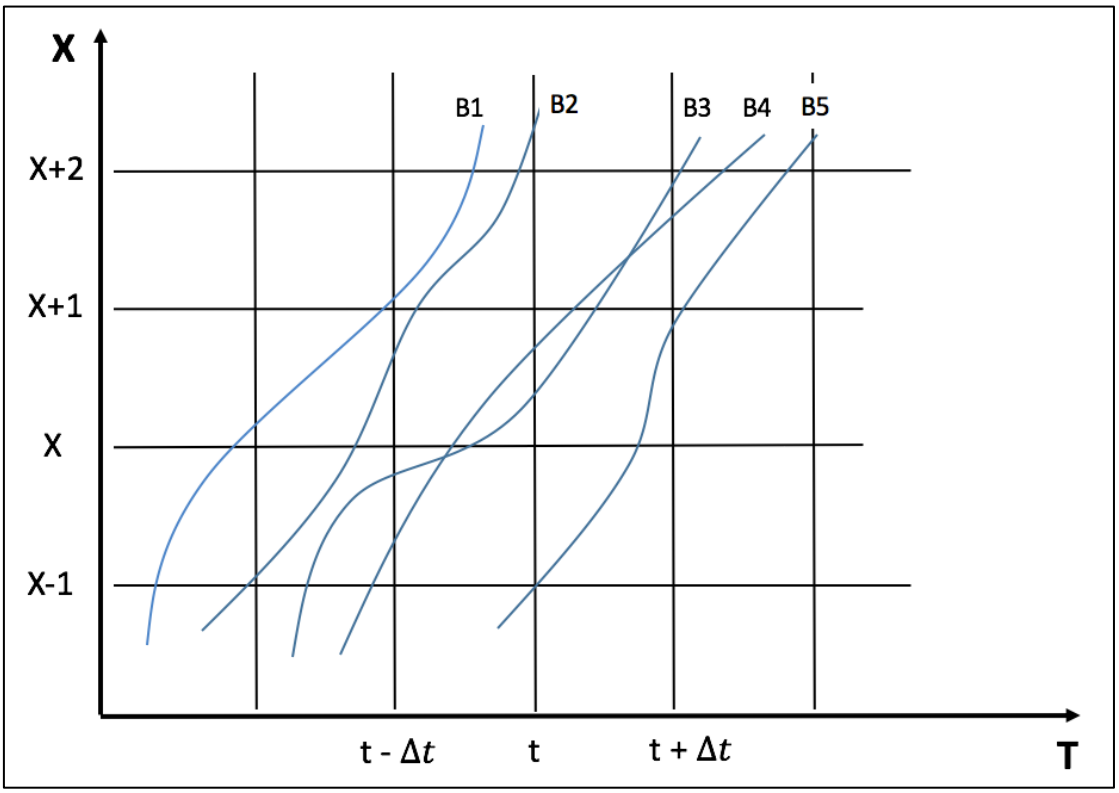


Figura 4 - 8. Acercamiento a diagrama cuadriculado de espacio-tiempo

Como se mencionó antes, las posiciones reportadas por GPS instalados en los buses del Transantiago, no son posiciones continuas, sino que pulsos cada 30 segundos. Esto genera que se tengan posiciones discretas de los buses y que estas no siempre calcen con el paradero. Como el objetivo de la tesis es predecir velocidades entre paraderos, cuando uno de estos pulsos no coincide con un paradero, se interpola la posición en función de la distancia para obtener la hora de pasada por el paradero. La metodología utilizada para esto fue desarrollada por Cortés et al (2011) y modificada posteriormente por TransitUC.

El hecho que los pulsos sean cada 30 segundos también implica que no se pueda determinar cuánto tiempo estuvo el bus detenido en el paradero. Cuando calza el pulso de la posición del bus con el paradero, no se puede determinar si este recién llegó, o si está por irse. A su vez, cuando el pulso de la posición no calza con el paradero y hay que interpolar para determinar la hora de pasada por este, se tiene el mismo problema. Lo que genera que los datos de pasada por los paraderos no sean tan limpios y puedan haber segundos de desfase con la realidad.

La creación de variables explicativas de velocidad consiste en que para cada predicción se tiene un set de variables de entrada, creadas tanto de datos en tiempo real como histórico, correspondientes a promedios espaciales de velocidades en cada uno de los rectángulos de la grilla. Sin embargo, no tiene sentido que para la predicción de velocidad de un cierto tramo se ocupen datos de todos los tramos del trayecto, por lo que siguiendo la teoría de las ondas de choque se decidió ocupar datos del tramo a predecir, más un tramo aguas abajo y uno aguas arriba. Para datos de velocidades en tiempo real, se utilizó intervalos en el eje del tiempo desde 15 a 30 minutos antes de la hora actual hasta la hora actual, donde el largo del intervalo depende de la frecuencia del servicio de bus analizado. Ahora, para datos históricos, se ocupó ese mismo intervalo anterior y además un intervalo que va desde la hora actual hasta 15 a 30 minutos más adelante. En la Figura 4-9 se puede apreciar mejor la explicación.

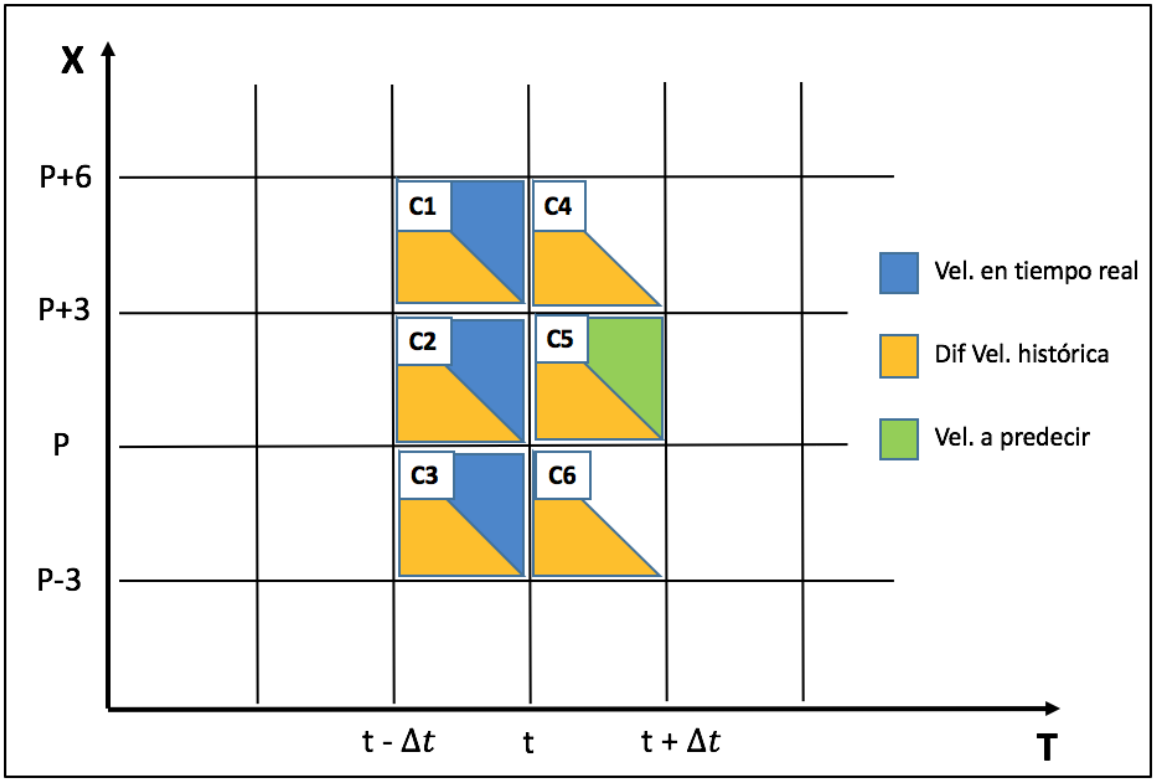


Figura 4 - 9. Grilla de variables de velocidades

El eje de la altura de los rectangulos corresponde a tres tramos, equivalente a la distancia entre 3 paraderos. Esta distancia no es fija, ya que depende del paradero en que se esté analizando. Sin embargo en promedio los paraderos tienen un distanciamiento de 300 metros, lo que hace que en promedio la altura del rectángulo sea de 900 metros. Se eligió ese distanciamiento, ya que Julio (2015) determina que separación de 3 paraderos es el punto óptimo entre predicciones no muy agregadas espacialmente, pero si lo suficiente para que esten menos sujetas a factores tales como semáforos, vehículos detenidos o circulando lento, entre muchas otras. Respecto a la cantidad de buses que pasan por la celda en el intervalo de tiempo [t - Δt , t], se demostró que no hay diferencia significativa si pasan por esa celda 1, 2 o más buses a la hora de utilizarlas para realizar las predicciones. Sin embargo, se considera que por cada celda circula al menos un bus, por lo que se omitió de las predicciones datos que no tuvieran información.

En la Figura 4-9, se observan triángulos de distintos colores y en la reseña a la derecha de la figura la explicación de cada uno. Sin embargo, es importante destacar que las velocidades en tiempo real (triángulos azules) se ocuparon con sus valores correspondientes (velocidades entre 0 y 70 km/hr), no así los rectángulos amarillos, los cuales son la diferencia entre la velocidad historica del día de la semana, menos la velocidad en tiempo real en cada rectángulo. De esa forma se evita tener alta correlación entre las variables (0.75 o mayor), mejorando así el desempeño en algunos de los modelos. Para que quede más claro, en la Ecuación 4.1 se puede observar lo explicado, donde C_i representa simplemente la celda *i*, para que al lector le quede más claro de que variable se etá hablando.

$$Dif.Vel.Histórica_{C5} = Vel.Histórica_{C5} - Vel.Tiempo\ Real_{C2} \qquad (4.1)$$

Como se mencionó, el largo del intervalo de tiempo de a grilla depende de la frecuencia o *headway* de los servicios. De la Tabla 4-2 presentada en la Sección 4.1.4, la cual muestra las características de los *headways* en cada servicio de buses analizado, se puede deducir que elegir intervalos de tiempo en la grilla igual a 15 minutos para los servicios 212R y 203R, es razonable. No obstante, para el servicio C04I, dado que la media de sus *headway* es 17,7 minutos, se requieren intervalos más grandes, por lo que se escogió intervalos de 30 minutos.

b) Variables de factores de demanda y comportamiento de pasajeros

Los pasajeros pueden jugar un rol importante en la demora de un bus de un paradero a otro, por lo cual en esta tesis se decide incluir cuatro variables influenciadas por los pasajeros, a saber: subidas pagadas, subidas no pagadas, bajadas y carga del bus. Las tres primeras de ellas están relacionadas al comportamiento de los usuarios en cada paradero y la última, a como viaja el bus entre paraderos. En Santiago, el transporte público se paga con una tarjeta llamada BIP!, por lo que para saber cuántas subidas pagadas hubo en cada paradero del recorrido de un bus, se consultó el registro de pago de tarjetas BIP!. En caso que el paradero corresponda a zona paga, es decir, que se pague por entrar al paradero y no por subirse al bus, se consideró que el número

de subidas a los buses que pasan por ese paradero son proporcionales a la frecuencia de pasada de los distintos servicios. Por otra parte, las subidas no pagadas son una evasión del sistema de pago, con lo cual no queda registro de cuántas hubo en cada paradero. Sin embargo, gracias a registros observados y a la tesis de Guarda (2015), se pudo estimar subidas no pagadas en función de las subidas pagadas, periodo del día, sentido del servicio y la parada. Las bajadas, en cambio, fueron estimadas mediante matrices Origen-Destino desarrolladas por Munizaga et al (2012), las cuales determinan la probabilidad de que una persona que se sube en el paradero A se baje en el paradero B. Por último, la carga del bus en un cierto tramo viene dada por la Ecuación 4.2. Donde *i* representa el tramo a predecir y *p* el paradero entre el tramo *i-1* y el tramo *i.*

$$Carga_i = Carga_{i-1} + Subidas\ Pagadas_p + Subidas\ No\ Pagadas_p - Bajadas_p \quad (4.2)$$

Además, a partir de estas variables se creó una nueva que resultó tener buenos resultados a la hora de incluirla en las predicciones, la cual es *no hay subidas ni bajadas* en un paradero, que como su nombre lo dice, es una variable *dummy* con valor uno si no hay subidas ni bajadas en el paradero, y cero en caso contrario.

La variable *bajadas* es la única que está disponible a futuro ya que es una estimación de bajadas en función de la gente que se subió en los tramos anteriores. Sin embargo, *subidas pagadas* y *no pagadas*, *carga del bus* y *no hay subidas ni bajadas*, son variables que solo se dispone de su valor en los tramos anteriores y no en el tramo a predecir. Ahora, como se trabaja con una base de datos, se dispone de esa información a futuro, es decir, se sabe cuál fue el valor real que tuvieron las variables en el tramo que se quiere predecir. Para medir cual es el efecto real de estas variables en la predicción de las velocidades de buses, se ocupa su valor como si se supiera el futuro. Ahora bien, ocupar el futuro es irrealista, por lo que se propone también incluir predicciones de estas variables ocupando los mismos tres modelos usados para predecir la velocidad de los buses: RLM, ANN y SVM. No se descarta que haya

métodos mejores para predecir estas variables, sin embargo, estos quedan fuera del alcance de esta tesis y le queda al lector probar con otros métodos en busca de mejores resultados.

c) Variables de infraestructura

Las variables de infraestructura, son variables estáticas, donde en un periodo largo de tiempo no suelen cambiar. Las variables de este tipo incluidas en los modelos fueron la distancia entre paraderos y la característica de la calle, en otras palabras, si era corredor segregado o no.

Como las predicciones fueron agrupados en tramos de tres paraderos, la distancia viene siendo dada como la distancia en el tramo completo a predecir. El corredor en cambio, de no estar agrupados los paraderos, sería una variable *dummy*. Sin embargo, al agruparse los paraderos, se definió como una variable continua entre 0 y 1, conteniendo la proporción en que el tramo a predecir es corredor segregado o no.

d) Variables de entorno

Santiago no es una ciudad preparada para la lluvia, y los ductos de drenaje colapsan fácilmente, por lo que como último tipo de variable se decidió introducir el clima, específicamente los milímetros de agua caídos las últimas 6 horas, cifra que se obtiene de la Dirección Meteoreológica de Chile (Precipitación Diaria, sf). Por otro lado, la gente suele manejar más lento de noche, por lo que se incluyó también una variable *dummy* de si es de día o no.

4.2.3. Selección de variables predictivas y set de datos final

No siempre todas las variables explicativas juegan un rol significativo en las predicciones. Por lo que fue necesario evaluar, del set total de posibles variables predictivas, cuales eran estadísticamente significativas y cuales no al momento de predecir las velocidades de buses. Para realizar eso, se utilizó la metodología de *Forward Stepwize Selection*, la cual consiste en a partir de una regresión con un solo intercepto ir agregando una a una las variables que más aportan a las predicciones,

hasta que eventualmente no quedan más variables por agregar, o las que quedan no mejoran el desempeño del modelo. Se elige este método por ser el método más utilizado en la literatura y además para tener el mismo set de variables independiente del modelo de aprendizaje estadístico utilizado.

Se dispone de datos de tres servicios de buses para analizar el desempeño de los modelos, y como es de esperar la selección de variables mediante la metodología de *Forward Stepwize Selection* no dio exactamente igual en los tres servicios. Por lo que, se agregó las que fueron significativas en al menos dos de los servicios y se evaluó la inclusión de las variables que fueron significativas solo en uno.

Del total de 18 variables predictivas, compuestas por 9 variables de velocidad, 5 de factores de demanda, 2 de infraestructura y 2 de factores de entorno; se llega a un set de variables final compuesto por 12 variables. De las 12 variables seleccionadas, 5 son variables de velocidad las cuales se pueden ver en la Figura 4-10; y las otras 7 son de factores de demanda, infraestructura y entorno. Todas ellas se pueden ver en la Tabla 4-3. De la Figura 4-10, se puede ver que solo se ocupan velocidades en tiempo real e históricas del tramo a predecir y de un tramo aguas arriba. Para los recorridos analizados, el tramo aguas abajo no tiene significancia en las predicciones.

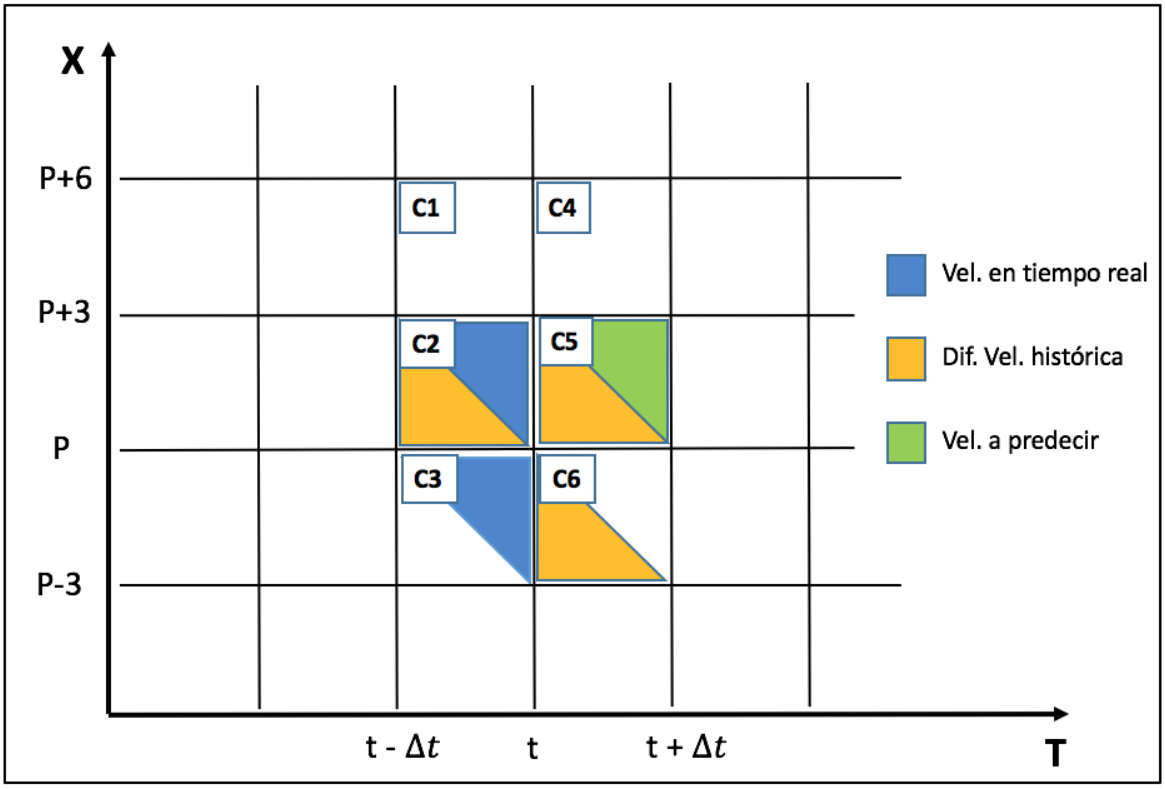


Figura 4 - 10. Grilla con set final de variables de velocidad

Tabla 4- 3. Variables predictivas finales

Variables	**Nombre**	**Tipo en estado original**	**Tipo cuando está agrupada**
Factores de demanda	Subidas Pagadas	Entera	Continua
	Bajadas	Entera	Continua
	Carga	Entera	Continua
	No hay subidas ni Bajadas	*Dummy*	Continua
Infraestructura	Distancia	Continua	Continua
	Corredor Segregado	Continua	Continua
Entorno	Lluvia (mm caídos)	Continua	Continua

Donde la matriz de correlaciones de las variables seleccionadas para el servicio 212R se puede ver en la Tabla 4-4 (el resto de los servicios tienen resultados semejantes, por lo que no se agregan sus matrices). Se observa que la velocidad a predecir (Vel. Real) tiene una alta correlación con las variables de velocidad en tiempo real en cuadrante C2 (Vel. T.R. C2), lo que tiene mucho sentido, ya que es la velocidad promedio de los buses que pasaron entre 15 minutos antes hasta la actualidad en el tramo a predecir. No así con la velocidad en tiempo real del tramo aguas arriba (Vel. T.R. C3) que tiene una correlación baja, indicando que no hay mucha relación entre la velocidad del tramo anterior y el a predecir.

Tabla 4- 4. Tabla de correlación de variables predictivas en servicio 212R

	Vel. T.R. C3	Vel. T.R. C2	Dif. V.H. C2	Dif. V.H. C6	Dif. V.H. C5	**Vel. Real (a predecir)**
Vel. T.R. C3	1	0,29	0,02	0,54	0,02	**0,21**
Vel. T.R. C2	0,29	1	0,48	0,00	0,50	**0,72**
Dif. V.H. C2	0,02	0,48	1	0,03	0,77	**0,29**
Dif. V.H. C6	0,54	0,00	0,03	1	0,04	**-0,02**
Dif. V.H. C5	0,02	0,50	0,77	0,04	1	**0,26**
Vel. Real (a predecir)	**0,21**	**0,72**	**0,29**	**-0,02**	**0,26**	**1**

	Carga Bus	Subidas Pagadas	Bajadas	No hay Sub. ni Baj.	Distancia	Corredor	Lluvia	**Vel. Real (a predecir)**
Carga Bus	1	0,51	0,58	-0,56	0,04	-0,04	-0,07	**-0,29**
Subidas Pagadas	0,51	1	0,45	-0,52	-0,04	-0,15	-0,03	**-0,32**
Bajadas	0,58	0,45	1	-0,50	0,12	-0,18	-0,03	**-0,32**
No hay Sub. ni Baj.	-0,56	-0,52	-0,50	1	0,03	0,21	0,03	**0,41**
Distancia	0,04	-0,04	0,12	0,03	1	0,54	-0,01	**0,14**
Corredor	-0,04	-0,15	-0,18	0,21	0,54	1	-0,01	**0,28**
Lluvia	-0,07	-0,03	-0,03	0,03	-0,01	-0,01	1	**-0,02**
Vel. Real (a predecir)	**-0,29**	**-0,32**	**-0,32**	**0,41**	**0,14**	**0,28**	**-0,02**	**1**

Respecto a las variables de infraestructura, factores de demanda y entorno (Tabla 4-4 inferior) se puede observar que *Carga de bus*, *Subidas pagadas y Bajadas*, como es de esperar tienen correlación negativa. Es decir, que mientras mayor sea su valor menor es la velocidad real que se tiene en el tramo a predecir. Por otro lado, la variable *no hay subidas ni bajadas*, tiene una correlación mayor al resto, indicando que esta variable creada a partir de las otras podría tener un alto impacto en las predicciones. Por último, la variable *Lluvia*, tiene una baja correlación, por lo que no se espera que esta sea una variable determinante en las regresiones.

Ahora, al mirar las correlaciones entre las variables explicativas, se observa que en general se tiene correlaciones bajas entre las variables, lo cual es bueno, ya que altas correlaciones podrían empeorar el desempeño de los modelos. Ahora las variables que tienen una mayor correlación entre ellas son Dif. V.H. C2 y Dif. V.H. C5, con un valor de 0,77. Se considera como correlación alta valores sobre 0,75, y como la correlación entre ambas variables apenas sobrepasa dicho valor no debería tener un efecto negativo.

Dado que muchos de los modelos se calibran más rápido al normalizar los datos, no se ocupó los datos originales del set, sino que cada variable se normalizó de acuerdo a la Ecuación 4.3, obteniéndose de esa forma variables que distribuyen normal con media cero y varianza uno. Luego de realizar la predicción se realizó el proceso inverso para así obtener los valores reales de las predicciones, por lo que las medidas de desempeño fueron evaluadas con el set de datos original y predicciones no normalizadas.

$$x_{i\,normalizado} = \frac{x_i - \bar{x}}{\sigma} \tag{4.3}$$

Respecto a la base de datos utilizada, se utilizó información recopilada durante los meses de Marzo y Abril de 2016, donde es importante recordar que un supuesto es que se dispone información en todas las celdas. Por ende, es necesario que al menos un bus pase por cada celda de la grilla para que esta sea introducida en el set final.

Considerando dicho supuesto y la elección del tamaño de las celdas de la grilla para cada servicio de buses analizado, se llega a la Tabla 4-5.

Tabla 4- 5: Resumen set de datos final en base a elección de parámetros

Servicio	**Fecha inicio**	**Fecha termino**	**Extensión Celda**		**Cantidad de datos set original**	**Porcentaje captados**	**Cantidad de datos set final**
			Temporal (min)	**Espacial (nº paradas)**			
212R	15-03-16	15-04-16	15	3	34837	87,1%	30343
203R	15-03-16	15-04-16	15	3	55075	92,0%	50669
C04I	01-03-16	27-03-16	30	3	7743	93,9%	7271

5. CALIBRACIÓN DE LOS MODELOS

Un aspecto fundamental a la hora de medir el desempeño de los modelos es la calibración de estos, donde los resultados pueden variar drásticamente en función de los parámetros que se eligen. En este capítulo se describen las medidas de desempeño para evaluar los modelos, para luego explicar cómo se calibra cada uno. Finalmente se presentan los modelos base o *benchmark* de los cuales se habla en las secciones anteriores.

Es importante mencionar que, tanto para calibrarlos como para su uso, se utiliza el programa R con el paquete estadístico *Caret*. Se prueba con otros paquetes, aunque por su rapidez, simpleza y cobertura de todos los métodos de aprendizaje estadístico utilizados en este trabajo, se recomienda al lector dicho paquete.

5.1. Selección de medidas de desempeño

Para medir el desempeño de los modelos y, poder realizar comparaciones válidas entre ellos, se utilizan tres medidas: raíz del error cuadrático medio (RMSE por sus siglas en inglés), error absoluto medio (MAE) y error porcentual absoluto medio (MAPE). En las Ecuaciones 5.1, 5.2 y 5.3 se puede ver la formulación de cada error. Donde n es la cantidad de datos, p_i es la predicción en la instancia i de la muestra y o_i es la observación en dicha instancia. Es importante que el lector sepa que mientras menor sea el valor en cada uno de los errores, mejor es el desempeño de los modelos.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - o_i)^2}{n}} \tag{5.1}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|p_i - o_i|}{n} \tag{5.2}$$

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n}\frac{|p_i - o_i|}{o_i}}{n} \tag{5.3}$$

De la formulación del error RMSE, se puede deducir que a diferencia de los otros errores, este penaliza fuertemente predicciones alejadas de las observaciones. Por la misma razón, este es considerado el error principal al momento de calibrar y comparar el desempeño de los modelos. Esto porque al predecir velocidades de buses, errores pequeños no causan mayores problemas al momento de tomar decisiones, no así cuando la predicción es muy alejada de la realidad. El error MAE, en cambio, no penaliza cuadráticamente predicciones muy alejadas, sin embargo, tiene la utilidad de ver cuánto es el error absoluto medio que se esta cometiendo, y por ende permite ver en promedio qué tan alejadas estan las predicciones de las observaciones. Por último, la formulación del error MAPE es parecida a la del MAE. Sin embargo, es un error porcentual, haciendolo muy útil para comparar desempeños entre distintos rangos de velocidades, o incluso entre distintos recorridos.

Ahora, las medidas de desempeño dependen de la muestra de entrenamiento y validación seleccionadas. Para evitar eso, se utiliza la técnica de validación cruzada con *k-fold* igual a 5. Esta consiste en separar la muestra total en 5 porciones iguales, luego calibrar el modelo con 4 de estas y validarlo obteniendo las medidas de desempeño con la porción que se deja afuera. Se realiza ese paso cinco veces, dejando siempre una porcion distinta afuera. Finalmente se promedian las medidas de desempeño de las cinco iteraciones, obteniendose asi una medida de error representativa del set de datos. En la Figura 5-1 se muestra este hecho.

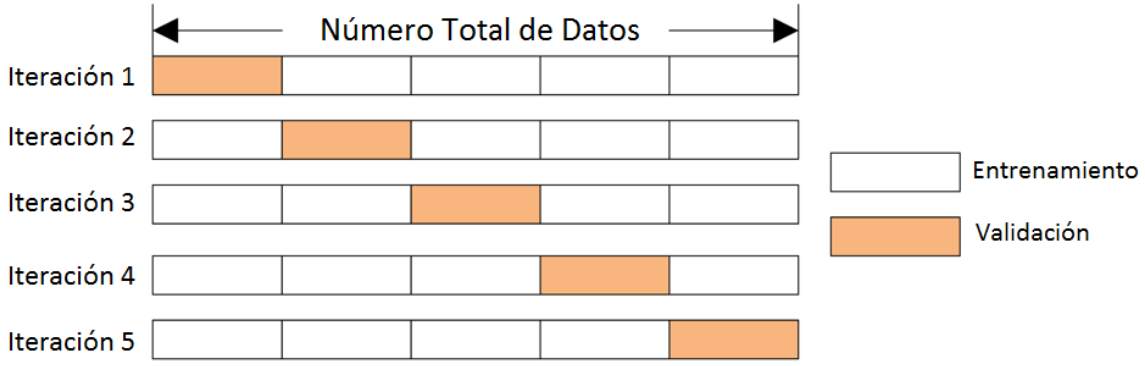


Figura 5 - 1. Validación cruzada con *k-fold* igual a 5. Fuente: StackOverFlow, s.f.

5.2. Definición de modelos base o *benchmark*

Los modelos base o *benchmark*, son modelos simples, que no requieren ningún tipo de entrenamiento ni calibración. Son modelos rápidos de desarrollar, pero tienen supuestos que en la mayoría de las situaciones no se cumplen (van Lint, 2004).

En este trabajo se eligen dos modelos *benchmark* para comparar el desempeño de los modelos de aprendizaje estadístico. Uno de ellos denominado *benchmark histórico*, y el otro *benchmark en tiempo real*.

5.2.1. *Benchmark* histórico (BH)

Este modelo se puede resumir en la Ecuación 5.4 y Figura 5-2, donde vemos que la predicción es meramente la velocidad histórica promedio que ha habido en el tramo a predecir. Vemos que es un modelo muy simple que no se necesita realizar cálculos en los instantes que se quiere predecir. A su vez, no requiere de ningún tipo de información en línea, lo cual lo hace útil en caso de no tener conexión constante de GPS.

$$Predicción_{\,[p\,,\,p+3][t\,,\,t+\Delta t]} = Vel.\ Histórica_{\,[p,p+3][t\,,\,t+\Delta t]} \tag{5.4}$$

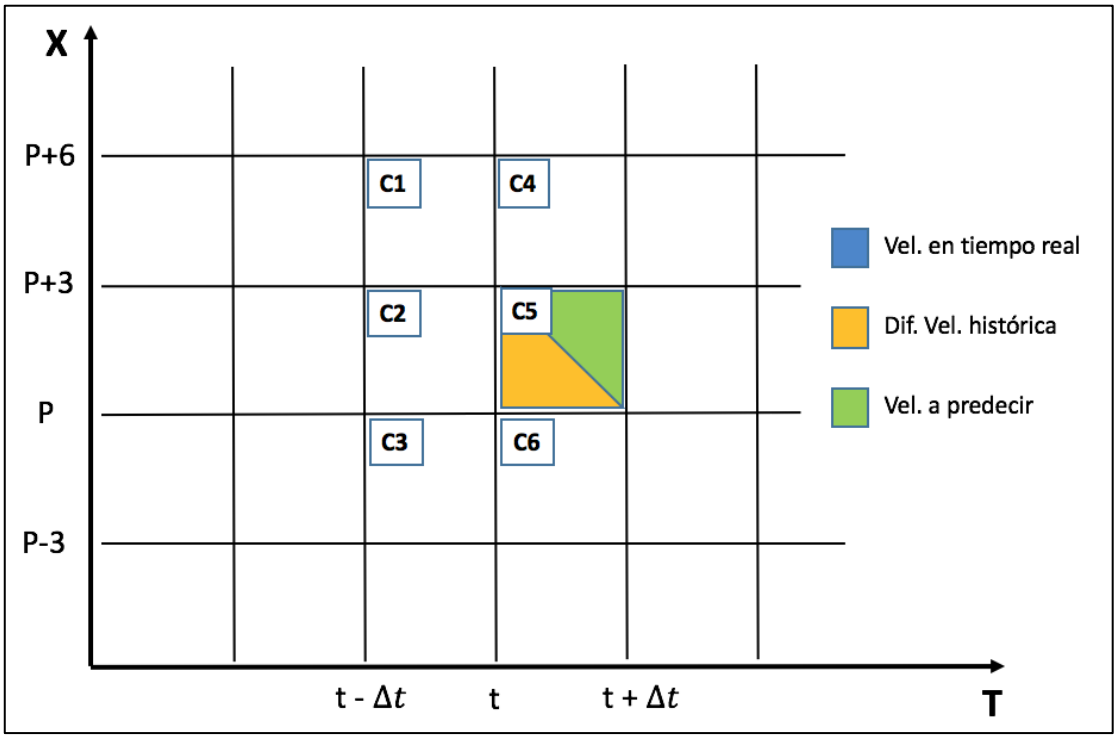


Figura 5 - 2. Modelo *Benchmark* Histórico

5.2.2. *Benchmark* en tiempo real (BTR)

Este modelo es semejante al anterior, aunque su predicción tiene relación con la velocidad que tuvieron los buses de adelante cuando pasaron por el tramo a predecir. En la Ecuación 5.5 y Figura 5-3 se puede ver su formulación, donde la predicción de la velocidad para un tramo es el promedio de las velocidades en tiempo real de los buses que pasaron en dicho tramo en un intervalo de tiempo desde la hora actual a 15 o 30 minutos antes (largo del intervalo depende de la frecuencia del servicio analizado). El modelo asume que los GPS están siempre funcionando y que por ende se tiene al menos cada 30 segundos la posición de cada bus.

$$Predicción_{\,[p\,,\,p+3][t\,,\,t+\Delta t]} = Vel.\ Tiempo\ Real_{\,[p,p+3][t-\Delta t\,,\,t]} \qquad (5.5)$$

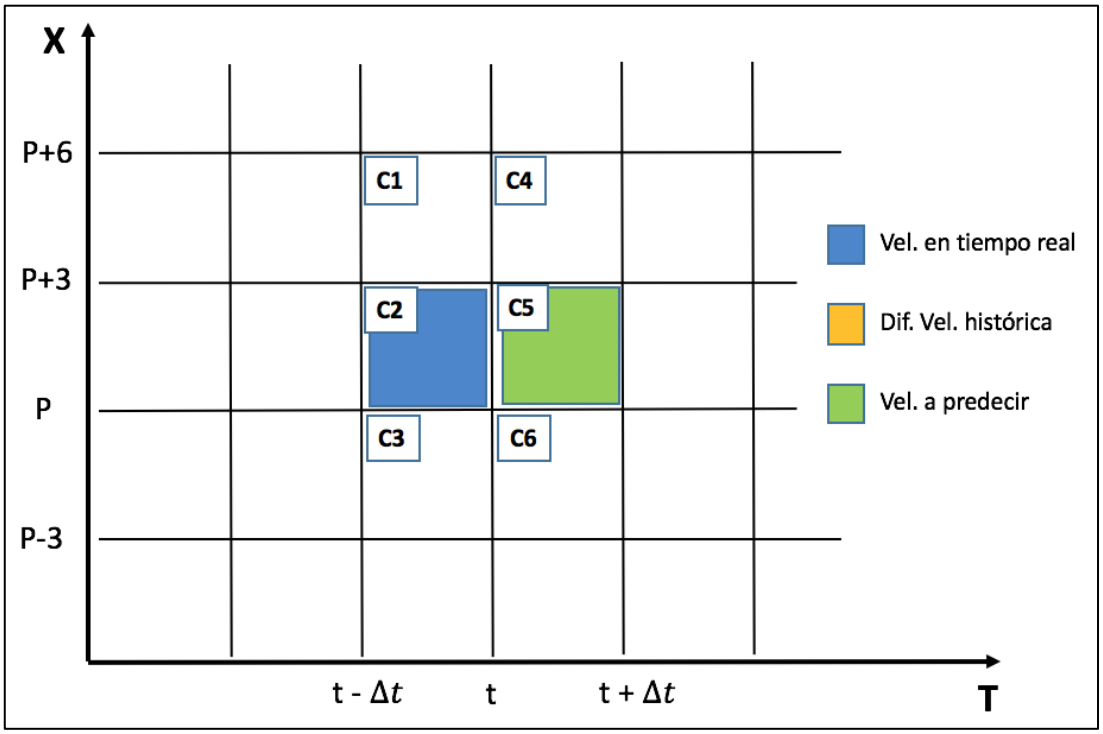


Figura 5 - 3. Modelo *Benchmark* en Tiempo Real

5.3. Calibración de los modelos de aprendizaje estadístico propuestos

5.3.1. MLR

En la Sección 2.1.1 se presenta los modelos de regresión lineal múltiple y de cómo se puede transformar las variables explicativas para ajustarse a relaciones no lineales. Para dejar más claro este hecho, se puede ver la Figura 5-4, donde en la imagen de la izquierda se ajusta una recta mediante una regresión lineal. En dicha recta se utiliza solo el término X, y claramente se puede ver que la recta no ajusta muy bien con los datos. Sin embargo, al realizar la misma regresión lineal y agregarle el término X^2, se observa que la nueva curva (imagen derecha) se ajusta muy bien a los datos. El modelo en si sigue siendo lineal, aunque las variables explicativas no necesariamente lo son.

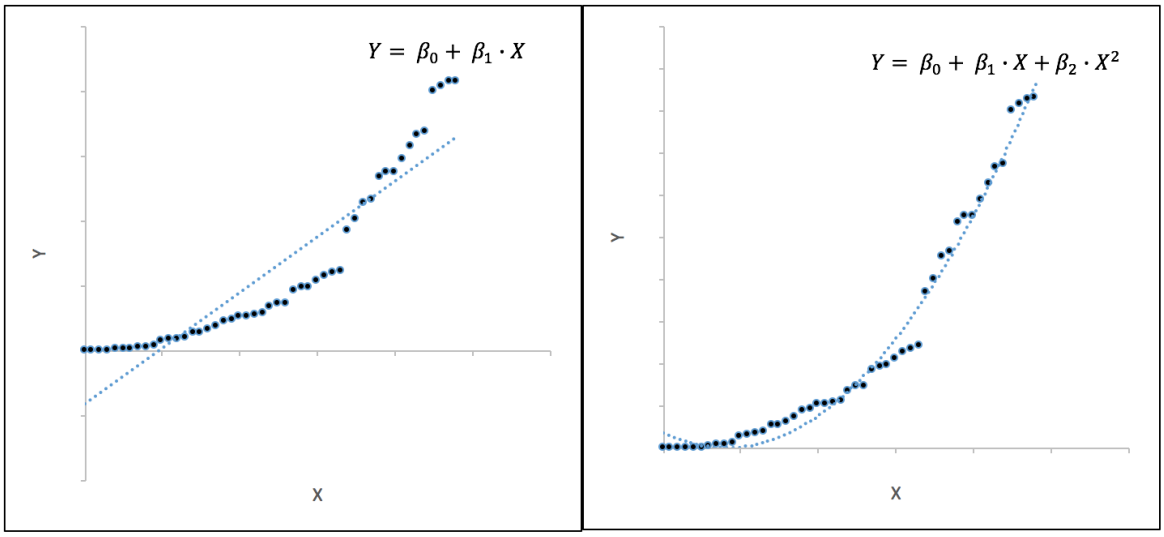


Figura 5 - 4. Ajuste lineal versus uno cuadrático en una regresión lineal. Fuente: elaboración propia.

El mismo concepto se utiliza al calibrar los modelos de regresión lineal utilizados para los tres servicios analizados. Aquí se prueban distintas combinaciones con polinomios de distintos grados para cada variable explicativa. Luego de probar distintas combinaciones, se registra cual tuvo menor RMSE y se trabaja con dichos valores. Esto se realizó empíricamente probando todas las combinaciones polinómicas entre grados del 1 al 6. A su vez, se probó incluir interacciones entre variables, pero no reporto mejores resultados.

En la Tabla 5-1 se muestra el grado de polinomio de cada variable utilizada en la regresión para cada servicio analizado. Se puede observar que, a excepción de algunas variables, los grados de polinomio óptimos son iguales entre los distintos servicios de buses. Lo anterior indica que la influencia de las distintas variables predictivas en la dependiente tiene un comportamiento similar independiente del servicio de buses analizado.

Es importante destacar que, para los servicios analizados, calibrar un modelo de Regresión Lineal Múltiple sin utilizar transformaciones polinómicas; es decir, ocupar la variable sin elevar a ninguna potencia, no empeora mucho las predicciones (La

raíz del error cuadrático medio aumenta alrededor de un 1.5%). En esta tesis se busca encontrar la mejor calibración para cada modelo propuesto, por lo que a pesar de agregar complejidad y no mejorar mucho los resultados, se decide igualmente trabajar con combinaciones polinómicas.

Tabla 5 - 1. Grado de polinomio de las variables utilizadas en el modelo MLR

	Grado de Polinomio		
	212R	203R	C04I
Vel. Tiempo Real $_{[p\ ,\ p+3]\ [t-t15\ ,\ t]}$	4	4	4
Vel. Tiempo Real $_{[p-3\ ,\ p]\ [t-t15\ ,\ t]}$	4	4	4
Dif. Vel. Histórica $_{[p\ ,\ p+3]\ [t-15\ ,\ t]}$	2	3	1
Dif. Vel. Histórica $_{[p\ ,\ p+3]\ [t\ ,\ t+15]}$	3	3	2
Dif. Vel. Histórica $_{[p\ ,\ p+3]\ [t\ ,\ t+15]}$	3	3	2
Lluvia (mm caídos)	4	4	-
Subidas Pagadas	3	3	3
Bajadas	3	3	3
Carga del Bus	3	1	1
No hay Subidas ni Bajadas	1	1	1
Distancia	5	5	3
Corredor Segregado	2	2	-

Por último, los modelos de Regresión Lineal Múltiple, se basan en dos supuestos fundamentales:

- Errores aleatorios tienen una distribución normal homocedastica.
- Ausencia de multicolinealidad.

El primer supuesto garantiza que la distribución condicional de la variable dependiente sea Normal y que tenga varianza constante, dado un conjunto de valores tomado por los regresores (Ortúzar y Willumsen, 2011). El segundo, confirma que no haya colinealidad entre las variables, es decir que no haya una o más variables que son una combinación lineal de otra.

Para medir el cumplimiento de los supuestos, se puede realizar distintos test. En el Anexo B se puede ver los resultados de los test aplicados para evaluar los supuestos en cada servicio. De dichos resultados se concluye que hay una ausencia de multicolinearidad en los tres servicios de buses. Sin embargo, los errores alateorios no tienen una distribución normal ni homocedastica. La violación de este supuesto implica que los resultados obtenidos por MLR pueden no ser confiables, lo que podría poner en duda el funcionamiento real de este modelo para la predicción de velocidades de buses.

5.3.2. SVM

En la Sección 2.1.2 se presenta las Máquinas de Soporte Vectorial y de las funciones de *kernels* que se utilizan para realizar la regresión. Existen una serie de funciones como lineal, polinómica, radial, entre otras más. En este trabajo se ocupa solo funciones de base radial, por dos razones: (i) tienen mejores resultados que las otras funciones y (ii) en la literatura se utiliza en su gran mayoría funciones de base radial al realizarse predicciones de velocidades y/o tiempos de viaje de buses (Yu et al., 2011).

Por otro lado, hay dos parámetros que se tienen que calibrar en un modelo SVR: ε y C. Donde ε es el ancho del margen suave y C el costo o penalización de dejar un punto fuera de estos márgenes. Para calibrar los modelos se probó con distintas combinaciones de ε y C, donde ambos se variaron entre cero e infinito, obteniéndose de esa forma, el conjunto con los mejores resultados. Igual para ε valores superiores a 0,5 empeoraban drásticamente el resultado y para C valores superiores a 5 mostraban un desempeño claramente más pobre. En la Tabla 5-2 se puede ver los valores de los parámetros mencionados en función de los tres recorridos analizados.

Tabla 5 - 2. Valor de parámetros calibrados en modelo SVM

	212R	203R	C04I
C	1	1	1,5
ε	0,3	0,3	0,3

5.3.3. ANN

Como mencionamos en la Sección 2.1.3, existen variaciones de Redes Neuronales. Sin embargo, en este trabajo se utilizó solo *Radial Basis Function Neural Network* (RBFNN). A diferencia de otras variaciones de algoritmos de Redes Neuronales, RBFNN solo tiene una capa oculta, y se tiene que calibrar el número de neuronas al interior de esta. Una capa oculta basta para aproximar casi cualquier relación entre variables de entrada y salida (Wang, Zuo, & Fu, 2014). A su vez, reduce fuertemente los tiempos de calibración, lo cual hace ideal a RBFNN como algoritmo de Redes Neuronales cuando se tiene una cantidad de datos que genera que calibrar el modelo tome horas, como es el caso de este trabajo.

No existe un consenso de cuál es el número óptimo de neuronas por capa oculta, sin embargo, Heaton (2008) presenta tres reglas empíricas que sirven para guiarse al momento de elegir el número de neuronas ocultas:

- Debe variar entre el número de variables de entrada y salida.
- Debe ser 2/3 del número de variables de entrada, más el número de variables de salida.
- Debe ser menor al doble del número de variables de entrada.

El set de datos final tiene 12 variables de entrada y una de salida. Con lo cual, bajo dichas reglas, la capa oculta debiera tener entre 1 y 24 neuronas. Ahora, la experiencia mostró que al normalizar los datos del set no es necesario un gran número de neuronas, donde los valores son cercanos a aplicar la segunda regla; es decir a 2/3 del número de variables de entrada más el número de variables de salida. En la Tabla 5-3, se muestra cual fue el número de neuronas en la capa oculta en función de los servicios y de los distintos sets de variables utilizados.

Ahora, en RBFNN se tiene que calibrar también el decaimiento de los pesos, en inglés *weight decay*. Este decaimiento de pesos, genera una reducción en algunos

coeficientes a cero, lo cual asegura encontrar un óptimo local con parámetros de pequeña magnitud. Esto suele ser crucial para evitar problemas de sobreajuste, donde valores mayores a 0,1 son considerados como un decaimiento fuerte (Hastie et al, 2001). En la Tabla 5-3 se puede ver los valores del número de neuronas y el decaimiento de los pesos para los modelos de Redes Neuronales calibrados en los distintos servicios.

Tabla 5 - 3. Valores de calibración de los parámetros de un modelo ANN

	212R	203R	C04I
# neuronas capa oculta	10	11	6
Weight Decay	0,05	0,1	0,01

6. RESULTADOS

En este capítulo se detalla los resultados obtenidos por los distintos modelos para cada servicio de buses. Los resultados de cada servicio son presentados por separado para mantener el orden. Sin embargo, en la antepenúltima sección se hace un análisis comparativo para analizar en términos generales el desempeño de los modelos. Luego, se muestra cómo influyen las distintas variables predictivas en el desempeño del mejor modelo de cada servicio, para finalmente, en la última sección, presentar que significa en términos de tiempos los errores en las predicciones de velocidades.

6.1. Servicio 212 retorno (212R)

La Tabla 6-1 presenta los resultados obtenidos por los distintos modelos, al predecir velocidades en el servicio 212R. Se puede observar que el mejor modelo, en términos de la raíz del error cuadrático medio, es la red neuronal (ANN). Sin embargo, la regresión lineal múltiple (MLR) tiene un resultado muy parecido. La máquina de soporte vectorial (SVM) no se desempeña tan bien como los otros dos; aunque lo hace mejor que los modelos *benchmark*. De estos dos últimos, el *benchmark* en tiempo real (BTR) tiene un desempeño peor que los algoritmos de aprendizaje estadístico, específicamente un 10.7% peor que ANN; si bien bastante superior al *benchmark* histórico, el cual se desempeña con un error un 21.4% sobre la red neuronal.

Tabla 6 - 1. Resultados de las predicciones de los modelos en el servicio 212R

	BH	BTR	MLR	ANN	SVM
RMSE (km/hr)	6,95	6,12	5,47	5,46	5,57
MAE (km/hr)	4,75	3,80	3,81	3,81	3,71
MAPE (%)	20,42	15,39	17,14	17,19	16,35

Tal como se mencionó en la Sección 5.1, RMSE es la medida de error usada para rankear el desempeño de los modelos, sin embargo, se analizaron también otros

errores. El MAE es el error absoluto medio en km/hr, es decir, el error promedio entre la velocidad predicha y la observada. De la Tabla 6-1, se desprende que los algoritmos cometen en promedio un error absoluto de más o menos 3,8 km/hr en las predicciones, a excepción del BH, cuyo MAE es significativamente superior al resto. En el Anexo C-a se muestran gráficos de velocidad predicha versus real para cada modelo, donde se puede observar que en general se tiende a sobreestimar velocidades bajas y a subestimar velocidades altas.

Es importante destacar que ANN tiene un mejor desempeño en RMSE, pero no en el resto de los errores. Para entender este fenómeno es necesario separar el desempeño de los modelos en función de rangos de velocidades. Para eso se eligieron tres rangos correspondientes a velocidades bajas, medias y altas. Para realizar la separación se utilizó un modelo *k-means*, el cual separa las velocidades reales en tres grupos: velocidades bajas que van entre 0 y 20 km/hr; medias, entre 20 y 32 km/hr; y altas, velocidades superiores a 32 km/hr. Para una explicación del modelo *k-means* se puede ver Hastie et al. (2001).

Tabla 6 – 2. Resultados de los modelos según rango de velocidad en servicio 212R

		BH	BTR	MLR	ANN	SVM
Baja	RMSE (km/hr)	3,60	2,64	3,62	3,61	3,25
	MAE (km/hr)	2,84	2,05	2,86	2,85	2,57
	MAPE (%)	20,69	14,05	19,77	19,80	18,14
Media	RMSE (km/hr)	4,89	4,07	3,92	3,96	3,96
	MAE (km/hr)	3,87	2,98	3,11	3,13	3,05
	MAPE (%)	15,51	12,02	12,73	12,84	12,12
Alta	RMSE (km/hr)	16,98	15,70	12,63	12,59	13,65
	MAE (km/hr)	15,35	13,69	10,44	10,38	11,51
	MAPE (%)	37,79	33,40	25,23	25,08	27,89

La Tabla 6-2, muestra los resultados en función de los distintos rangos. Es importante mencionar que estos son los errores de las predicciones separados en tres rangos y no nuevas predicciones de tres bases de datos separadas en función de los

rangos. En la tabla se puede observar que para velocidades bajas, el modelo de *benchmark* en tiempo real obtiene los mejores resultados en todas las medidas de desempeño, con resultados bastante mejores que los otros modelos, los cuales tuvieron desempeños parecidos entre ellos. Respecto a velocidades medias, los modelos de aprendizaje estadístico tuvieron resultados muy similares entre ellos y ligeramente superiores al BTR. Finalmente, en el rango de velocidades altas es donde se obtiene la mayor diferencia entre modelos y a su vez los mayores errores, donde ANN y MLR obtienen el mejor desempeño, seguido de las máquinas de soporte vectorial y, finalmente, los modelos de *benchmark* con un desempeño mucho más pobre.

De la tabla es interesante también comparar el MAPE en los distintos rangos de velocidades. Observamos que este error es menor para velocidades medias, luego lo sigue velocidades bajas y finalmente es mucho mayor para velocidades altas. No así el MAE, donde este es menor para velocidades bajas y mayor para velocidades altas. Esto quiere decir que en promedio, a medida que aumenta la velocidad, aumenta el error. Sin embargo, en términos porcentuales se cometen errores menores en velocidades medias, luego en bajas. Por otra parte, lo más difícil es predecir velocidades altas. Lo que concuerda con la intuición, ya que velocidades bajas son generalmente causadas por congestión, donde no hay mucha variabilidad en la velocidad de un tramo entre un bus y otro, explicando el bajo MAE. Por otra parte, por ser velocidades menores un pequeño error puede causar un mayor error porcentual, lo que explica un mayor MAPE. En cuanto a velocidades altas, estas son causadas por hecho fortuitos como varios semáforos en verde, pocas o ninguna subida ni bajada, conductores que manejan más rápido que otros, así como otros posibles hechos difíciles de predecir. Además, buses consecutivos no necesariamente tienen comportamientos similares.

En la Sección 6.7 se hace un análisis más detallado de los rangos de velocidades, con el fin de mostrar qué implica en términos de tiempos de viaje los errores de las predicciones de velocidades.

6.2. Servicio 203 retorno (203R)

Los resultados de los modelos calibrados para el servicio 203R se pueden ver en la Tabla 6-3, donde se observa que nuevamente el mejor modelo en términos de la raíz del error cuadrático medio es la red neuronal, con un desempeño ligeramente superior a la regresión lineal múltiple y a la máquina de soporte vectorial. Se desprende también que ocupar ANN como modelo tiene un error 9,5% menor que al ocupar un modelo simple de *benchmark* con información en tiempo real y, un 18.2% menor que al ocupar un *benchmark* de velocidades históricas.

Tabla 6 – 3.Resultados de las predicciones de los modelos en el servicio 203R

	BH	BTR	MLR	ANN	SVM
RMSE (km/hr)	7,05	6,37	5,78	5,76	5,85
MAE (km/hr)	4,78	4,01	3,99	3,98	3,88
MAPE (%)	21,28	16,91	18,79	18,73	17,75

Respecto al error absoluto medio (MAE), a excepción del BH, los otros modelos tienen resultados muy parecidos, con valores que rondan los 4 km/hr. Lo que quiere decir que en promedio la diferencia entre la velocidad predicha y la observada es de 4 km/hr, donde la velocidad promedio de este recorrido completo es de 21,9 km/hr.

Gráficos de velocidad predicha versus real para cada modelo calibrado en el servicio 203R se pueden encontrar en Anexo C-b. Donde, al igual que en el servicio 212R se tiende a sobreestimar velocidades bajas y a subestimar velocidades altas.

Al igual que en el servicio 212R, ANN es el mejor modelo en términos del RMSE, pero no en las otras dos medidas de desempeño. Por lo que al separar el análisis en tres rangos de velocidad: baja, media y alta, específicas a este servicio, realizado con el método de *k-means*, donde las primeras van desde los 0 km/hr a los 19,8 km/hr, las segundas desde los 19,8 km/hr a los 32 km/hr y las últimas a velocidades superiores a 32 km/hr, se obtiene la Tabla 6-4. En ella se observa comportamiento

similar al recorrido 212R; es decir, para velocidades bajas el modelo que mejor predice es el *benchmark* en tiempo real, con resultados bastante mejores que los otros modelos. Para velocidades medias los desempeños son más parejos con resultados bastante parecidos en los tres modelos de aprendizaje estadístico y mejores que los modelos *benchmark*. Finalmente, para velocidades altas, hay una mayor diferencia entre modelos, donde ANN es el que mejor predice, seguido por MLR y luego SVM. Para este rango, los modelos *benchmark* desempeñan bastante peor en todas las medidas de desempeño y por sobre todo cuando se ocupa solo información histórica.

Tabla 6 – 4. Resultados de los modelos según rango de velocidad en servicio 203R

		BH	BTR	MLR	ANN	SVM
Baja	RMSE (km/hr)	3,77	2,94	3,98	3,95	3,55
	MAE (km/hr)	2,97	2,29	3,17	3,14	2,79
	MAPE (%)	22,84	16,47	23,12	22,96	20,73
Media	RMSE (km/hr)	4,82	4,29	3,96	4,01	4,13
	MAE (km/hr)	3,83	3,21	3,16	3,18	3,20
	MAPE (%)	15,44	13,03	13,08	13,16	12,75
Alta	RMSE (km/hr)	18,31	17,13	14,29	14,17	15,19
	MAE (km/hr)	16,55	15,00	11,72	11,57	12,77
	MAPE (%)	39,95	35,88	27,59	27,24	30,24

Ahora, si observamos el MAE y el MAPE, pareciera que el primero aumenta a medida que aumenta la velocidad. Sin embargo, entre velocidad baja y media, este aumento no es tan claro como en el recorrido 212R, lo que explica que se tengan MAPEs de casi el doble entre velocidades bajas y medias. Con las velocidades altas, como era de esperar, las tres medidas de desempeño se disparan en comparación a los otros rangos, lo que muestra que es más difícil predecir velocidades altas, con un error porcentual medio cercano al 30%.

6.3. Servicio C04 ida (C04I)

En la Tabla 6-5, se puede observar las medidas de desempeño de los modelos calibrados para predecir velocidades en el servicio C04I. Nuevamente vemos que el modelo con menor RMSE es la red neuronal, seguida muy de cerca por la regresión lineal y luego por la máquina vectorial. Respecto a los modelos *benchmark*, se observa que el de información en tiempo real tiene un mejor desempeño que el de información histórica, con un error 11,4% menor. Al ocupar redes neuronales mejora aún más, con una disminución de un 11,3% la raíz del error cuadrático medio, respecto al BTR y de un 21,8% respecto al BH. En cuanto al MAE, observamos que, a excepción del modelo BH, los valores rondan por los 3 km/hr, donde la velocidad promedio de este servicio es de 20,8 km/hr.

Tabla 6 - 5. Resultados de las predicciones de los modelos en el servicio C04I

	BH	BTR	MLR	ANN	SVM
RMSE (km/hr)	5,62	4,98	4,42	4,40	4,51
MAE (km/hr)	3,87	3,08	3,05	3,01	3,01
MAPE (%)	18,81	13,63	14,65	14,35	14,40

En el Anexo C-c se puede ver los gráficos de velocidad predicha versus real para cada modelo calibrado en el servicio C04I. Donde al igual que en los otros dos servicios se tiende a sobreestimar velocidades bajas y a subestimar velocidades altas.

Si este análisis lo hacemos separado en los tres rangos de velocidad, donde el intervalo de velocidades bajas es entre 0 km/hr y 19 km/hr, el de velocidades medias entre 19 km/hr y 30 km/hr, y finalmente el de alta velocidades mayores a 30 km/hr; obtenemos la Tabla 6-6. En esta, observamos que para velocidades bajas BTR es el mejor modelo y BH el peor. Por otro lado, para velocidades medias los modelos de aprendizaje estadístico y el *benchmark* en tiempo real obtienen resultados similares, y nuevamente *benchmark* histórico se desempeña peor. Finalmente, las predicciones para velocidades altas son más dispersas, donde ANN es el que obtiene mejores

resultados, seguido de cerca por la regresión lineal múltiple, luego por las máquinas de soporte vectorial, y ya con un desempeño bastante peor están los modelos *benchmark.*

Tabla 6 - 6. Resultados de los modelos según rango de velocidad en servicio C04I

		BH	BTR	MLR	ANN	SVM
Baja	RMSE (km/hr)	3,57	2,34	2,78	2,69	2,66
	MAE (km/hr)	2,86	1,81	2,18	2,08	2,11
	MAPE (%)	22,17	13,16	16,26	15,45	16,05
Media	RMSE (km/hr)	3,97	3,58	3,48	3,59	3,32
	MAE (km/hr)	3,07	2,65	2,79	2,84	2,61
	MAPE (%)	12,82	11,11	11,86	12,04	11,06
Alta	RMSE (km/hr)	13,50	12,67	10,17	10,03	10,87
	MAE (km/hr)	11,78	10,42	7,88	7,78	8,64
	MAPE (%)	30,65	26,75	19,99	19,77	21,99

Al igual que los resultados de los otros servicios, el MAE aumenta a medida que aumenta la velocidad, donde la diferencia de este error entre velocidades bajas y medias no es mucha, aunque en velocidades altas el valor se dispara a casi el triple. Ahora, si observamos el error porcentual, es decir el MAPE, vemos nuevamente que en las predicciones de velocidades medias se tienen los menores errores, seguido de las bajas, y finalmente las altas. A diferencia de los otros servicios, en las velocidades altas, si bien se incurre en un mayor error, este no es tan alto como con los otros recorridos, lo que puede ser porque el servicio C04I solo transita por calles comunes, teniendo una desviación estándar y dispersión de las velocidades altas menor.

6.4. Análisis comparativo

De los resultados obtenidos en los tres servicios podemos concluir que ANN es el mejor modelo en términos de RMSE a la hora de predecir velocidades de buses del Transantiago. Sin embargo, MLR obtiene resultados ligeramente peores, en

promedio un 0,2% peor, y es un modelo fácil y rápido de calibrar, por lo que no debiera descartarse como elección de modelo si se considera que en Santiago hay más de 350 líneas, y por ende el doble de servicios (DPTM, 2013).

Por otro lado, los modelos *benchmark* se desempeñan en todos los servicios peor que cualquier modelo de aprendizaje estadístico, en donde en promedio el modelo *benchmark* en tiempo real tiene un error cuadrático medio 10,5% más alto que ANN y el modelo *benchmark* histórico 20,5%. Lo que muestra el valor de la información en tiempo real y de utilizar modelos de aprendizaje estadísticos para realizar las predicciones.

Por otra parte, si analizamos las medidas de desempeño en cada servicio, vemos que en promedio el servicio 203R tiene las peores medidas de desempeño, luego el servicio 212R. El mejor es el servicio C04I. Pareciera ser que hay una relación entre la desviación estándar de los datos, que se puede ver en la Tabla 4-1, y las medidas de desempeño de los modelos. Sin embargo, llegar a ese conclusión es un poco apresurado, pero si podría tomarse como medida para predecir cómo va a ser la calidad de las predicciones de un cierto servicio.

En los resultados se obtiene un hecho no muy intuitivo; que los modelos de redes neuronales son los que predicen mejor en términos de la raíz del error cuadrático medio, aunque no lo son en términos del MAE ni del MAPE. En los tres servicios en términos de MAE, las máquinas de soporte vectorial fueron el mejor modelo y en términos de MAPE, el mejor fue *benchmark* en tiempo real. En realidad, este último es el que llama la atención, ya que a pesar de tener un enfoque muy ingenuo, en donde su predicción es meramente la velocidad promedio en ese tramo los últimos 15 o 30 minutos, tiene un menor error porcentual en las predicciones. Para entender este hecho es necesario mirar los análisis de intervalos de velocidades: baja, media y alta; donde en velocidades altas y medias los algoritmos de aprendizaje estadístico sobrepasan o en el peor de los casos, igualan el desempeño del BTR. Sin embargo, en velocidades bajas, este último tiene un desempeño mucho mejor que el resto. Esa

diferencia hace que en términos generales este obtenga un menor MAPE aun cuando para velocidades medias y altas este no lo tiene.

Respecto al análisis de velocidades bajas, medias y altas, se puede concluir del MAPE, que en términos porcentuales las velocidades en que se incurre menos error al predecir son las medias, seguidas de las bajas y después de las altas. Como ya se mencionó antes, el mayor MAPE en velocidades bajas se puede deber a que una pequeña diferencia entre la velocidad predicha y la real, genera un error significativo porcentualmente. Por otro lado, las velocidades altas son muy difíciles de predecir, ya que se pueden generar por hechos fortuitos, como varias luces en verde en un cierto tramo, que el bus no tenga que parar en paraderos por no haber subidas ni bajadas o incluso por conductores que manejan más rápido que otros. De hecho, si se mira los gráficos de la velocidad predicha versus la real de los Anexos C, vemos que rara vez se predice velocidades sobre los 40 km/hr, siendo que en ciertos casos los buses tienen velocidades de hasta 70 km/hr.

Si bien parece atractivo separar el análisis en rangos de velocidades, entrenando modelos específicos para cada uno, o sea, generar modelos mixtos o *hybrid models* en inglés, donde se tiene un modelo que prediga velocidades bajas, otro velocidades medias y un último las altas, en que previo a cada predicción se tiene que categorizar a que rango de velocidad pertenece la velocidad del tramo que se quiere predecir. Se ajustó modelos específicos para cada intervalo de velocidades y en caso de saberse que rango de velocidad se iba a tener en el tramo a predecir se obtenían mejores resultados. Sin embargo, debido a la dificultad de clasificar previamente en qué rango de velocidades se estaba prediciendo, es decir, si utilizar el modelo ajustado para velocidades bajas, o el ajustado a velocidades medias o altas, los resultados finales obtenidos no fueron buenos, por lo que se decidió no reportarlos en la tesis. En la Tabla 6-7 se puede ver las instancias correctamente clasificadas (en inglés *Correctly Classified Instances*, CCI) mediante el uso de Máquinas de Soporte Vectorial para clasificación, en relación a los rangos de velocidades separados por el método *k-means*. Se logra clasificar correctamente alrededor del 75% de la muestra,

en donde el 25% mal clasificado empeora mucho los resultados del modelo mixto, lo que explica su pobre desempeño.

Tabla 6 - 7. Resultados clasificación con modelo ANN

	CCI
212R	73,6%
203R	73,1%
C04I	77,7%

Para entender porque se tiene un alto error en la clasificación se puede ver la Figura 6-1. El gráfico izquierdo muestra una clasificación con el algoritmo de *k-means* de la velocidad real que se tuvo en el tramo a predecir, donde esta se separa en tres rangos y a cada rango se le asigna un color. El eje x está compuesto por esta velocidad real y el eje y, por la predicha por el modelo SVM. Del gráfico izquierdo se observa claramente los rangos en los cuales las velocidades son consideradas bajas, medias y altas; separadas por líneas verticales imaginarias. Ahora en el gráfico derecho, se tienen los mismos valores, no obstante, los colores de las clases son los predichos por el modelo de Máquinas de Soporte Vectorial para clasificación. Donde la separación de clases ya no es vertical, sino que horizontal (acorde al eje de velocidades predichas). Este hecho muestra porque se tienen altos errores de clasificación, haciendo que sea difícil implementar un modelo mixto.

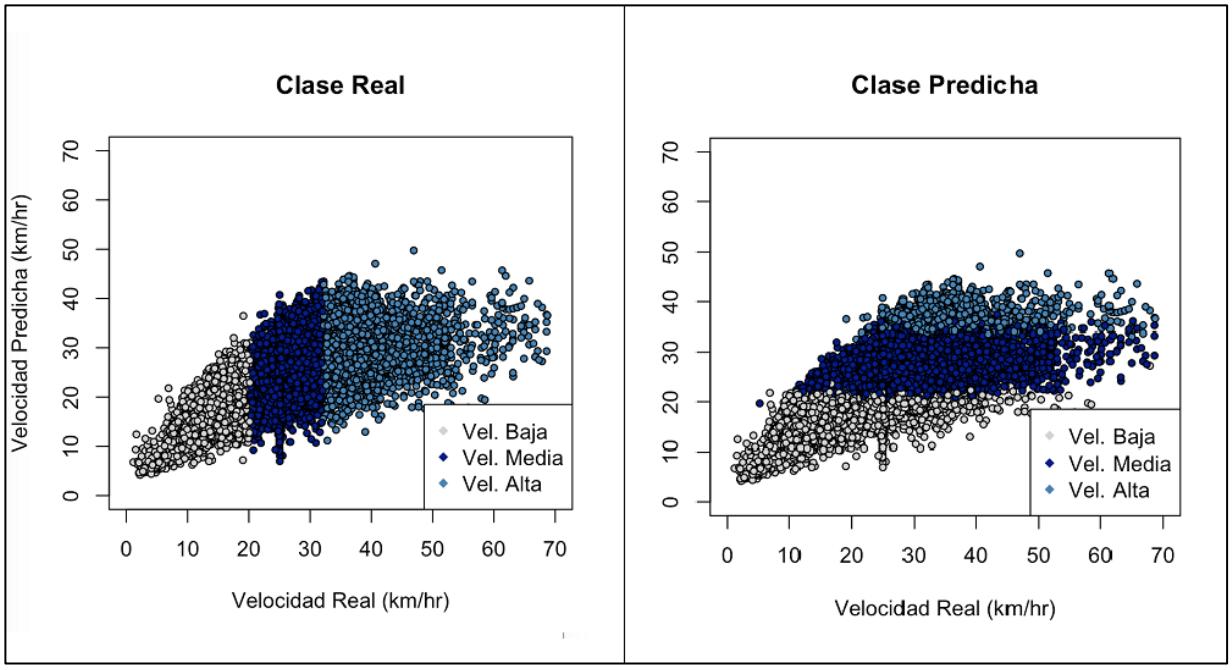


Figura 6 - 1. (Izquierda) Clasificación de las velocidades reales mediante modelo k-means. (Derecha) Clasificación SVM del rango de velocidad a predecir.

6.5. Análisis de las distintas variables predictivas

El segundo objetivo de esta tesis es ver cómo impactan las distintas variables explicativas en la predicción de velocidades de buses, especialmente ver cuánto mejora, si lo hace; agregar variables de comportamiento de demanda, infraestructura y de clima, a un set base de variables de velocidad.

Para realizar el análisis se utiliza el mejor modelo en cada servicio, es decir el modelo de redes neuronales, el cual se va calibrando nuevamente en función de las variables adicionadas o sustraídas. En la Tabla 6-8 se pueden ver los resultados obtenidos para cada servicio en términos de mejora porcentual en la raíz del error cuadrático medio (RMSE) para cada variable, o set de variables agregados a un set base de solo variables de velocidad. Se presenta únicamente el cambio porcentual del RMSE, ya que con los otros errores los cambios porcentuales fueron similares. La primera columna muestra en qué servicio fue el cambio; la segunda el RMSE base, es decir el error incurrido al solo agregar variables de velocidad; y el resto de las

columnas muestran en cuánto se disminuye (signo menos) o se aumenta (signo más) el RMSE base al agregarle cada variable o set de estas. El *set infra*. es el set que agrupa las variables *corredor* y *distancia*. Por su parte, el *set Fact. Demanda* es el set que agrupa todas las variables de factores de demanda. La última columna, es el resultado de tener todas las variables en el modelo; es decir, el mismo resultado que se presenta para el modelo ANN en las secciones anteriores.

Tabla 6 - 8. Mejora porcentual al agregar distintas variables a un set base

	SoloVe + (RMSE base)	Corredor	Distan-cia	Set Infra.	Lluvia	Bajadas	Carga Bus	No Sub ni Baja,	Subida Pagadas	Set Fact. Demanda	**TODAS**
212R	**5,557**	-0,04%	-0,21%	-0,18%	0,03%	-0,70%	-0,74%	-1,51%	-1,10%	-1,70%	**-1,78%**
203R	**5,849**	-0,03%	-0,14%	-0,13%	0,03%	-0,24%	-0,46%	-0,89%	-1,20%	-1,45%	**-1,53%**
C04I	**4,530**	-	-0,27%	-0,27%	-	-0,27%	-0,45%	-1,93%	-2,93%	-2,80%	**-3,04%**

Si observamos la última columna de la tabla vemos que la adición de todas las variables predictivas mejora en poco la raíz del error cuadrático medio que si se agregan solo variables de velocidad (máxima mejora de 3% en el servicio C04I). Por lo que cabe cuestionarse si vale el esfuerzo la recopilación de dichas variables. Para efectos del análisis que viene a continuación se considerará cualquier mejora como significativa.

Por otro lado, vemos que agregar cada variable por sí sola, o en conjunto, mejora el RMSE, excepto la variable *Lluvia*, la cual empeora el error. Esta se introduce porque en el método de *Forward Stepwize Selection*, que se utiliza para ver que variables son significativas a la hora de predecir en un modelo MLR, mejora un poco las predicciones, pero al utilizar ANN como modelo, no lo hace. Con las variables *Corredor* y *Distancia* ocurre hecho similar, donde la variable distancia por si sola genera un disminución igual o mayor del error RMSE a que si se introducen ambas juntas. Ahora la disminución es prácticamente insignificante, sin embargo, dada la

construcción de las variables de velocidad, probablemente si se disponga del dato distancia, por lo que se recomienda agregarlo igual.

Respecto a las variables de factores de demanda, vemos que estas tienen un mayor impacto en la mejora del RMSE. Un hecho curioso es que la variable *Subidas Pagadas* por sí sola genera casi el mismo impacto que agregar todo el set de este tipo. A su vez, teniendo el dato de subidas pagadas se puede obtener la variable *Bajadas* y por ende también la variable *No hay Subidas ni Bajadas*, ya que como se mencionó en la Sección 4.2.2.b), las bajadas son estimaciones en función de las subidas en un cierto paradero a una cierta hora, y la variable *No hay Subidas ni Bajadas*, es meramente la proporción de paraderos en que no hubo ninguna subida ni bajada en el tramo a predecir.

Es importante recordar, que las variables *Subidas Pagadas* y *No hay Subidas ni Bajadas*, son variables que se ocupan a futuro. Es decir, para efectos de esta tesis se ocupa su valor como si se supiera cuantas subidas van a haber en el siguiente tramo, o que proporción de paraderos no va a tener ninguna subida ni bajada. Por lo que, si bien, queda fuera de los alcances de esta tesis determinar metodologías para predecir estas variables, se utiliza los mismos métodos de aprendizaje estadísticos para realizar estimaciones en el tramo a predecir. A su vez, en caso de no tener información en tiempo real de las subidas en los paraderos, se prueba con variables históricas, donde las *Subidas Históricas* vendría siendo la tasa histórica de subidas en un paradero a una cierta hora por el tiempo que transcurrió desde que pasó el bus anterior por ese mismo paradero. En la Ecuación 6.1 se puede ver la formulación matemática. *No hay Subidas ni Bajadas Históricas*, en cambio, es simplemente la proporción histórica de dicha variable en el tramo a predecir a una cierta hora de un cierto día.

$$Subidas\ Históricas_{[p,p+3][t\,,\,t+\Delta t]} = Tasa\ histórica\ de\ subidas_{[p,p+3][t\,,\,t+\Delta t]} * \left(Hora\ pasada\ bus\ adelante_p - Hora\ pasada\ bus\ actual_p\right) \quad (6.1)$$

En la Tabla 6-9 se puede ver cómo afecta la inclusión de las variables *Subidas Pagadas* (S.P) y *no hay Subidas ni Bajadas* (noSniB) tanto históricas como predichas con los modelos de aprendizaje estadístico. Vemos que al ocupar la predicción de estas variables como variable de entrada para la predicción de velocidades de buses, se pierde al menos la mitad de lo mejorado que al ocupar sus valores reales. En otras palabras, nuestra mejor predicción de estas dos variables no llega a captar ni la mitad del efecto de ocupar su valor real. Ahora si vemos ambas variables con valores históricos, estas no tienen una mejora significativa en las predicciones, sin embargo, no descartaría su uso, ya que por pequeño que sea no se requiere mayor esfuerzo obtener dichos valores.

Tabla 6 - 9. Mejora porcentual al agregar variables históricas y predichas en set de variables base

	Solo Vel + (RMSE base)	noSniB Real	noSniB predicha	noSniB histórica	S.P. Real	S.P. predicha	S.P. histórica	Todas Predichas	**Todas Real**
212R	5,557	-1,51%	-0,71%	-0,08%	-1,10%	-0,49%	-0,17%	-0,98%	**-1,78%**
203R	5,849	-0,89%	-0,22%	0,01%	-1,20%	-0,22%	-0,07%	-0,78%	**-1,53%**
C04I	4,530	-1,93%	-0,41%	-0,15%	-2,93%	-0,60%	-0,50%	-0,53%	**-3,04%**

Por último, en la columna Todas Predichas, se ve el efecto de hacer las predicciones con las 12 variables originales del set, pero con predicciones de las *Subidas Pagadas, No hay subidas ni bajadas, Carga del Bus y Bajadas*. Donde, como se mencionó en la Sección 4.2.2.b), las dos últimas ya se ocupaban con sus valores estimados. De esta columna se observa que se logra captar un poco más de la mitad de la mejora del RMSE que cuando se ocupan los valores reales y no estimados de las variables. Sin embargo, se vuelve a cuestionar su inclusión, debido al alto esfuerzo de recopilarlas y la baja mejora que generan.

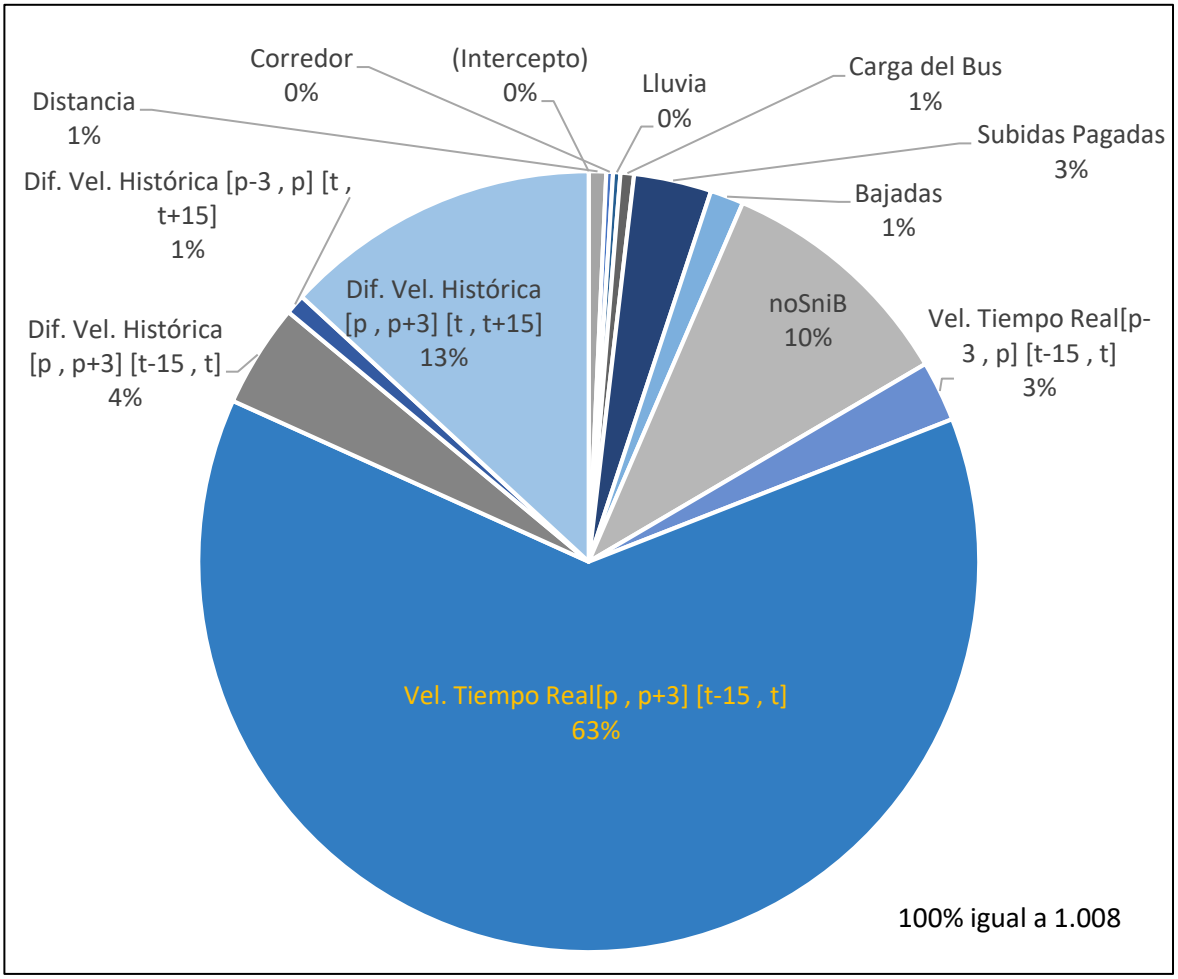
6.6. Peso de las variables predictivas

Una de las razones por la cual se incluye el modelo de regresión lineal múltiple como modelo de aprendizaje estadístico para predecir velocidades de buses, es porque a diferencia de las redes neuronales y las máquinas de soporte vectorial, este no es un modelo de caja negra. Es decir, es fácil de interpretar y se puede ver con claridad el peso que tiene cada variable explicativa en las predicciones.

Para que el análisis fuera más fácil de entender, se trabaja con las doce variables originales, sin considerar transformaciones polinómicas de estas. Si bien, con transformaciones polinómicas el resultado no es exactamente igual, el cambio es pequeño, por lo que el análisis es semejante.

En la Figura 6-2 se puede observar el peso, expresado en porcentaje, que tiene cada variable explicativa sobre la dependiente (velocidad a predecir). El peso de cada variable viene explicado por el cociente entre el coeficiente de la variable y la suma de todos los coeficientes (se toma el valor absoluto de cada valor). Es importante recordar que las variables están normalizadas; es decir, todas tienen media cero y varianza igual a uno, por lo que a mayor valor absoluto del coeficiente de la variable mayor es el peso que tiene sobre la predicción.

Figura 6 - 2. Peso de las distintas variables explicativas en el modelo MLR para el servicio 212R



En la Figura 6-2 se muestra solo el gráfico del servicio 212R, ya que el resto arroja resultados similares. Igualmente, en el Anexo B se puede ver los valores de los coeficientes de las variables explicativas para los tres servicios de buses analizados.

De la Figura 6-2 se observa, que la variable que tiene el mayor peso en la regresión es Vel. Tiempo Real$_{[p\ ,\ p+3]\ [t-15\ ,\ t]}$. Es decir, la variable que guarda la velocidad promedio que tuvieron los buses que pasaron por el tramo a predecir entre 15 minutos antes hasta la hora de la predicción. El valor del coeficiente que lo sigue es de 50 puntos porcentuales menos, lo que indica, que dicha variable es por lejos la que tiene mayor peso en las predicciones.

El resto de las variables tiene por lo general un peso bastante bajo, donde destacan los coeficientes de las variables *Corredor* y *Lluvia*, por su valor prácticamente nulo. Lo que explica que en la Sección 6.5, la inclusión de estas variables en las predicciones no haya generado mejoras significativas.

En la sección 6.5 se observó que disponer de variables de factores de demanda, infraestructura y clima mejoren en tan poco las predicciones, siendo que la lluvia, cantidad de subidas, cantidad de bajadas, por nombrar algunas de las variables predictivas; pueden afectar fuertemente las velocidades de viaje. Sin embargo, la respuesta pareciera estar en el peso de las variables, donde la variable que tiene por lejos mayor peso en las predicciones es la Velocidad en Tiempo $Real_{[p,\ p+3]\ [t-15\ ,\ t]}$. Por su construcción, esta variable, incluye estados de tráfico, subidas, bajadas, factores de infraestructura, accidentes, y todo lo que haya determinado la velocidad de los buses en un cierto tramo. Por lo que en su valor están implícitos dichos factores. Dicho de otra manera, la inclusión de todos ellos no genera una gran mejora en las predicciones, por ya estar implícitamente considerados en las variables de velocidad utilizadas. Las variables *Subidas Pagadas* y *No hay subidas ni bajadas*, son variables que pueden variar fuertemente entre dos buses consecutivos, ya que dependen de factores como el distanciamiento entre ellos, o la cantidad de pasajeros que se encuentran viajando. Por lo que su impacto en las velocidades de los buses no se ve completamente captado por la velocidad promedio de los buses en el tramo a predecir. Lo que explica, que la inclusión de estas dos variables tenga un mayor impacto en la disminución del error de las predicciones, pero aun así, su impacto sea pequeño.

6.7. Análisis de error de las predicciones en términos de tiempo

Como se mostró en la Sección 2.2, muchos de los trabajos reportados en la literatura son predicciones de tiempos de viaje de los buses, más que velocidades. Por las razones expuestas en la Sección 3.2, se decidió trabajar con velocidades. Sin

embargo, no deja de ser interesante un análisis de que implica en términos de tiempos de viaje, los errores cometidos en las predicciones de velocidades.

En la Tabla 6-10, muestra un análisis de errores cometidos en términos de tiempos de viaje, para los tres servicios analizados. La fila Tiempo real de viaje, muestra cual fue el tiempo promedio de viaje en minutos que realmente tuvo el bus en los tramos (acordarse de que los tramos son agrupaciones de tres paraderos), donde este tiempo viene dado simplemente por el cociente entre la distancia y la velocidad que se tuvo en dicho tramo. La fila Error medio, representa el error medio que se tuvo en minutos entre el tiempo real de viaje menos el tiempo predicho (Nuevamente se utilizó el cociente entre distancia y velocidad predicha en el tramo para obtener el tiempo predicho. Para la velocidad estimada se utilizó las reportadas por el modelo ANN, ya que fue el que tuvo mejores resultados). La fila Intervalo de confianza, como lo dice su nombre, es el intervalo de confianza del error medio al 95%, expresado en minutos. Como los errores no distribuyen normal, el intervalo no es necesariamente simétrico para ambos lados de la media. Por último, el análisis está dividido en los rangos de velocidad propuestos en las Secciones 6.1, 6.2 y 6.3, donde la columna Todos es el rango completo de velocidades.

De la tabla se observa que, en general, los tiempos de viaje de los buses son bastante similares entre los recorridos, donde las pequeñas diferencias vienen dadas por las distancias promedio de los tramos de los servicios. Ahora, respecto al error medio, es importante mencionar que este no está considerado en valor absoluto, por lo que un error medio positivo indica que el modelo tiende a sobrestimar las velocidades, y por ende a subestimar los tiempos de viaje.

Tabla 6 - 10. Análisis de tiempos de viaje en función de las velocidades predichas en los tres servicios de buses.

Servicio 212R	**Rango Tiempos (Rango Velocidades)**			
	Todos	Alto (Bajas)	Medio (Medias)	Bajo (Altas)
Tiempo real de viaje (min)	3,3	4,4	2,7	1,8
Error medio (min)	0,17	0,56	-0,02	-0,59
Intervalo confianza del error medio al 95% (min)	[-1,22 ; 1,71]	[-0,84 ; 2,36]	[-1,17 ; 0,83]	[-1,77 ; 0,28]

Distancia promedio de los tramos: 1,082 m

Servicio 203R	**Rango Tiempos (Rango Velocidades)**			
	Todos	Alto (Bajas)	Medio (Medias)	Bajo (Altas)
Tiempo real de viaje (min)	3,1	4,1	2,4	1,6
Error medio (min)	0,18	0,57	-0,03	-0,6
Intervalo confianza del error medio al 95% (min)	[-1,17 ; 1,73]	[-0,97 ; 2,35]	[-1,07 ; 0,72]	[-1,60 ; 0,10]

Distancia promedio de los tramos: 960 m

Servicio C04I	**Rango Tiempos (Rango Velocidades)**			
	Todos	Alto (Bajas)	Medio (Medias)	Bajo (Altas)
Tiempo real de viaje (min)	2,8	3,8	2,1	1,2
Error medio (min)	0,08	0,28	-0,02	-0,29
Intervalo confianza del error medio al 95% (min)	[-0,92 ; 1,20]	[-0,99 ; 1,58]	[-0,84 ; 0,56]	[-1,08 ; 0,18]

Distancia promedio de los tramos: 820 m

Respecto a los intervalos de confianza, se observa que para velocidades bajas y medias, y por ende tiempos de viaje altos y medios, en un 95% de los casos el error no sobrepasa el 50% del tiempo real de viaje. En donde las predicciones se pueden ir ajustando cada 30 segundos a medida que el bus va reportando su posición mediante pulsos de GPS. Ahora, para velocidades altas y por ende tiempos de viaje bajos, el tamaño de este aumenta considerablemente. Donde el intervalo contiene valores que podrían significar un error de casi un 100% del tiempo promedio real de viaje. Esto representa un problema, ya que no solo el intervalo de confianza del error es mayor,

sino que se dispone de menos instancias donde el bus puede reportar su posición y de esa forma poder actualizar las predicciones.

Cabe destacar que se probó predecir tiempos de viaje en vez de velocidades. Donde los resultados obtenidos fueron similares. Sin embargo, la gran diferencia es que prediciendo tiempos en vez de velocidades de viaje se cometen errores aún mayores en las velocidades altas (tiempos de viaje pequeños), y menores en velocidades bajas (tiempos de viaje grande). Por lo que, para este set de datos es más conveniente predecir velocidades a que tiempos.

7. CONCLUSIONES

En este Capítulo se obtienen las conclusiones del trabajo. Estas fueron separadas en tres secciones: (i) valor de la información en tiempo real, (ii) valor de los modelos de aprendizaje estadístico y (iii) valor de las distintas variables predictivas.

7.1. Valor de la información en tiempo real

Un aspecto importante de esta tesis es la disponibilidad de información en tiempo real. Con información histórica solo es posible desarrollar modelos como el *benchmark* histórico. En el Capítulo 6 vimos que este modelo no obtiene muy buenos resultados, y su performance se encuentra muy por debajo a los modelos de aprendizaje estadístico y al *benchmark* en tiempo real.

En los tres servicios analizados, el modelo *benchmark* histórico tiene un mayor error cuadrático medio que el *benchmark* en tiempo real. En promedio el modelo BH tuvo un RMSE de un 12,4% mayor que el modelo BTR, siendo similar en los tres servicios de buses analizados. Este valor muestra la utilidad de la información en tiempo real para mejorar las predicciones, donde solo información histórica es un enfoque muy ingenuo para predecir velocidades de buses.

La diferencia en los resultados de ambos modelos *benchmark*, también muestra la relación que hay entre los buses de un recorrido. Al tenerse solo información histórica se dispone información del estado de tráfico histórico en el tramo a predecir. Sin embargo, en caso de haber más (o menos) tráfico de lo habitual, o de tenerse incidentes como accidentes, bloqueos, entre otros; no se puede prever estos hechos, por lo que el modelo no es capaz de ajustar sus predicciones a la realidad. Ahora, cuando se dispone de información en tiempo real, el estado de tráfico del tramo a predecir queda implícito en la velocidad de los buses de adelante, por lo que el modelo es capaz de ajustar sus predicciones con información en tiempo real del estado de tráfico. Lo que explica la diferencia de los resultados de ambos modelos *benchmark*.

Por otro lado, al observar los resultados obtenidos por rangos de velocidades, se observa que al disponerse de información en tiempo real, las predicciones mejoran mucho, y por sobre todo en el rango de velocidades bajas. Esto se debe a que cuando el bus tiene velocidades bajas, pasa la mayor parte del tiempo viajando entre un paradero y otro, por lo que la información del estado de tráfico en ese tramo es crucial para realizar las predicciones. No tanto así, para velocidades medias y altas, donde otros variables, como si hubieron subidas y/o bajadas en un paradero, empiezan a tomar una mayor relevancia en la velocidad que va a tener el bus.

Ahora, respecto al valor de la información en tiempo real, no en las variables de velocidad, sino que en los factores de demanda. Si bien, la inclusión de estas variables tiene un menor impacto en las predicciones que introducir variables de velocidades en tiempo real, sus resultados no dejan de mostrar la importancia de esta. Donde, el beneficio de introducir variables de factores de demanda se pierde casi completamente si se incorpora estas variables con valores históricos.

7.2. Valor de los modelos de aprendizaje estadístico

Dentro de los objetivos principales de esta tesis, se encuentra la evaluación de los modelos de aprendizaje estadístico para predecir velocidades de buses. En los tres servicios analizados, la utilización de estos modelos, mejoró ampliamente las predicciones en comparación a los modelos base o *benchmark*. Donde el error cuadrático medio, al ocupar redes neuronales, fue en promedio un 11,9% menor que al ocupar el modelo de *benchmark* en tiempo real, y un 25,8% menor que al ocupar el modelo de *benchmark* con información histórica. A su vez, comparando los modelos benchmark con las máquinas de soporte vectorial (modelo de aprendizaje estadístico que tuvo peor desempeño en términos de RMSE), igual se observa una mejora significativa, con un error cuadrático medio de un 9,7% menor que al modelo BTR y 22,3% menor al modelo BH. Esto muestra la mejora que hay en las predicciones al tener modelos que sean capaces de incorporar más de una variable,

donde al agregarle complejidad y poder computacional se obtiene una mejora significativa en las predicciones.

Entre los modelos de aprendizaje estadístico, el que tuvo mejor resultado en los tres servicios de buses analizados, es la red neuronal artificial (ANN). Con un RMSE en promedio de un 0,3% menor a la regresión lineal múltiple (MLR), y de un 2% menor a las máquinas de soporte vectorial (SVM). Por lo que ANN se considera como el mejor modelo testeado para predecir velocidades de buses.

Dicho lo anterior, debido a la pequeña diferencia entre ANN y MLR, no descartaría este último como el modelo de elección para predecir velocidades de buses en toda la red del Transantiago. El hecho que no sea un modelo de caja negra; es decir, que se pueda ver fácilmente como influyen las variables predictivas en la predicción; y a su vez, la facilidad y rapidez para calibrarlo incluso en presencia de grandes bases de datos, lo hacen un modelo atractivo cuando se quiere predecir a gran escala. Por lo que, considerando que en Santiago hay alrededor de 350 líneas y por ende el doble de servicios, recomendaría su elección por sobre las Redes Neuronales Artificiales.

7.3. Valor de las distintas variables predictivas

En la Sección 6.5, se pudo ver la mejora al incluir distintas variables predictivas a un set base de solo variables de velocidad. Al sumarle a este set base, variables de factores de demanda, infraestructura y clima, solo reporto una mejora máxima de 3,04% en el servicio C04I, y 2,1% en promedio entre los servicios. No hay que olvidarse que esa mejora se da ocupando los valores reales de las variables de factores de demanda, sin embargo, al ocupar las predicciones realizadas para las variables *Subidas Pagas* y *No hay subidas ni bajadas*, la mejora fue de solo un 0,8%. La baja mejoría y el alto esfuerzo de recopilar dichas variables hace cuestionarse si vale la pena incluirlas en las predicciones. Se deja a criterio del lector la decisión de costo/beneficio entre esfuerzo de recopilación y mejora de predicciones.

Respecto a las variables de infraestructura, la inclusión de la variable *Corredor*, prácticamente no reportó mejora en las predicciones, por lo que no se recomienda su

inclusión. No así la variable *Distancia*, la cual género en promedio una mejora de un 0,21%. Si bien su impacto es pequeño, al construir las variables de velocidad necesariamente se utiliza la distancia entre los paraderos, por lo que se recomienda su uso ya que se dispone de su valor.

La inclusión de la variable *Lluvia* no genera una mejora en el error cuadrático medio, por lo que tampoco se recomienda incluirla dentro de las variables predictivas.

Es extraño pensar que disponer de variables de factores de demanda, infraestructura y clima mejoren en tan poco las predicciones, siendo que la lluvia, cantidad de subidas, cantidad de bajadas, por nombrar algunas de las variables predictivas; pueden afectar fuertemente las velocidades de viaje. Ahora, por la construcción de las variables de velocidad, estas incluyen información del estado de tráfico, subidas, bajadas, factores de infraestructura, accidentes y todo lo que haya determinado la velocidad de los buses en un cierto tramo. Por lo que la incorporación de dichas variables no mejora en mucho las predicciones, por ya estar considerados sus valores implícitamente en las variables de velocidad. De la misma forma se extiende el análisis a cualquier variable no considerada en este trabajo, pero que este implícitamente considerada en las variables de velocidad. Ahora, las variables *Subidas Pagadas* y *No hay subidas ni bajadas*, pueden variar considerablemente entre un bus y otro, ya que dependen de factores como el distanciamiento entre ellos, o la cantidad de pasajeros que se encuentran viajando. Por lo que su valor no queda completamente implícito en las variables de velocidad, lo que explica su mayor impacto en los resultados de las predicciones al incluir estas variables.

7.4. Extensiones

Como todo trabajo queda pendiente una serie de incógnitas que no se pudieron abordar debido a la base de datos que se tiene y a la construcción de esta. A continuación, se lista una serie de extensiones que se asocian a lo tratado durante esta tesis y en especial a las conclusiones realizadas.

- Como se mencionó anteriormente, la variable velocidad incluye tiempos de frenado, subidas, bajadas, entre otros factores; que ensucian el verdadero valor de la velocidad en el tramo a predecir. Por lo que se propone realizar este mismo trabajo, pero teniendo como input, tiempo de detención de los buses en paraderos y tiempos de viaje entre estos, por separado.
- Los pulsos de los GPS son cada 30 segundos. Tiempo en el cual se puede haber avanzado más de un paradero o al menos un buen tramo. Por lo que muchas veces el tiempo de pasada por un paradero que se dispone en la base de datos, es meramente una interpolación entre la distancia recorrida en esos 30 segundos y la distancia desde el paradero que se pasó. Se propone realizar este mismo trabajo, pero con pulsos cada 10 o 15 segundos. Esperándose tiempos de pasada y por ende velocidades más fidedignas.
- Si bien se probaron tres algoritmos de aprendizaje estadístico, quedaron muchos otros por probarse. Por lo que se propone probar con los restantes, en busca de uno que reporte aún mejores resultados.
- Se probó realizar predicciones con un modelo mixto o híbrido. Sin embargo, no se obtuvieron mejores resultados debido a la dificultad de clasificar en que rango de velocidades se estaba prediciendo. Por lo que se podría mejorar este modelo mixto al cambiar la probabilidad con que se clasifica una clase. Es decir, cambiar bajo que probabilidad se clasifica cada rango de velocidades, buscando una que se obtenga un mayor porcentaje de acierto en la clasificación.
- Respecto a las estimaciones de las variables *Subidas Pagadas* y *No hay subidas ni bajadas*, la metodología utilizada no pareciera ser la más adecuada. Por lo que se propone buscar otros métodos de predecir los valores de dichas variables.
- Se comete un mayor error prediciendo velocidades altas. Un factor que podría explicar esto es el perfil de conducción de los choferes. Podría ser

interesante introducir dicha variable en las variables explicativas de los modelos.

- Por último, considerando que en Santiago hay alrededor de 700 servicios de buses, calibrar modelos para cada servicio no necesariamente es lo óptimo. Por lo que, se podría probar resultados calibrando modelos por zonas de la ciudad o ejes que circulan los buses.

BIBLIOGRAFÍA

(DPTM), D. d. (2013). *Informe de Gestión 2013.*

Coffey, C., Pozdnoukhov, A., & Calabrese, F. (Noviembre de 2011). Time of arrival predictability horizons for public bus routes. *In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, 1-5.

Cortés, C. E., Gibson, J., Gschwender, A., Munizaga, M., & Zúñiga, M. (2011). Commercial bus speed diagnosis based on GPS-monitored data. *Transportation Research Part C: Emerging Technologies, 19(4)*, 695-707.

Cruz Pizarro, A. (24 de Enero de 2015). *Administración y Transportes*. Recuperado el 19 de Julio de 2016, de http://administracionytransportes.cl/

Du, L., Peeta, S., & Kim, Y. H. (2012). An Adaptive Information Fusion Model to Predict the Short-Term Link Travel Time Distribution in Dynamic Traffic Networks. *Transportation Research part B 46*, 235-253.

Dziekan, K., & Kottenhoff, K. (2007). Dinamyc at-stop Real-time information displays for public transport: Effects on customers. *Transport Research. Part A: Policy. Pract., 41(6)*, 489-501.

(2015). *Encuesta Origen-Destino de Viajes de Santiago.* Ministerio de Transporte y Télecomunicaciones, Santiago.

Guarda, P. (2015). *¿Que más hay detras de la evasión en el transporte público? Un enfoque econométrico.* Pontificia Universidad Católica de Chile.

Gurmu, Z. K., & Fan, W. D. (2014). Artificial neural network travel time prediction model for buses using only GPS data. *Journal of Public Transportation, 17(2)*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning.* Palo Alto, California: Springer.

Heaton, J. (2008). Introduction to neural networks with Java.

Imagen sin titulo [imagen]. (2016). Recuperado el 8 de Agosto de 2016, de http://www.statsoft.com/Textbook/Support-Vector-Machines

James, G., Witten, D., & Hastie, T. (2014). *An Introduction to Statistical Learning: with Applications in R.* Springer.

Jeong, R., & Rilett, L. (2004). Bus arrival time prediction using artificial neural network model. *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference* (págs. 988-993). IEEE.

Julio, N. (2015). *Aplicación de algoritmos de aprendizaje estadístico para predecir velocidades de buses con información en tiempo real.* Pontificia Universidad Católica de Chile, Santiago.

Julio, N., Giesen, R., & Lizana, P. (2015). Real-Time Prediction of Bus Travel Speeds Using Traffic Shockwaves and Machine Learning Algorithms.

Julio, N., Giesen, R., & Lizana, P. (2016). Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. *Research in transportation economics*.

Kotagiri, Y., & Pulugurtha, S. S. (2016). Modeling Bus Travel Delay and Travel Time for Improved Arrival Prediction. *International Conference on Transportation and Development 2016*, (págs. 562-573).

Kumar, V., Kumar, B. A., Vanajakshi, L., & Subramanian, S. C. (2014). Comparison of model based and machine learning approaches for bus arrival time prediction. *Proceedings of the 93rd Annual Meeting. Transportation Research Board*, 14-2518.

Lu, Y., Sundararajan, N., & Saratchandran, P. (1998). Performance evaluation of sequencial minimal radial basis function (RBF) neural network learning algorithm. *Neural Networks, IEEE Transactions on, 9*, 308-318.

Mazloumi, E., Currie, G., & Rose, G. (2010). Using traffic flow data to predict bus travel time variability through an enhanced artificial neural network. *In World Congress on Transport Research, 12th*(03377).

Mazloumi, E., Rose, G., Currie, G., & Moridpour, S. (2011). Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence, 24(3)*, 534-542.

McCulloch, W., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of MAthematical Biophisycs, Vol. 5*, 115-133.

Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems, Transportmetrica A. *Transport Science, 11:2*, 119-157.

Munizaga, M., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C 24*, 9-18.

Non-linear SVM [imagen]. (2012). Recuperado el 8 de Agosto de 2016, de Stackoverflow: http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

O'brien, R. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673-690.

Peng, Z. R., Yu, D., & Beimborn, E. (2002). Transit user perception of the benefits of automatic vehicle location. *Transportation Research Records*, 127-133.

Precipitación Diaria. (s.f.). *Dirección Metereológica de Chile*. Recuperado el 5 de julio de 2016, de http://www.meteochile.gob.cl/inicio.php

Rahman, M., Wirasinghe, S., & Kattan, L. (2016). The effect of time interval of bus location data on real-time bus arrival estimations. *Transportmetrica A: Transport Science, 12:8*, 700-720.

Ripley, B. (1996). Pattern recognition and neural networks. *Cambridge*.
Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222.

Soft Margin Linear SVM [imagen]. (2012). Recuperado el 8 de Agosto de 2016, de Stackoverflow: http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

StackOverFlow. (s.f.). Recuperado el 24 de Octubre de 2016, de How to implement walk forward testing sklearn: http://stackoverflow.com/questions/31947183/how-to-implement-walk-forward-testing-in-sklearn

Suport Vector Machine Regression. (2016). Recuperado el 4 de Agosto de 2016, de Non-linear Regression Function: http://kernelsvm.tripod.com/

van Lint, J. (2004). *Reliable Travel Time Prediction for Freeways.* PhD diss., Delft University of Technology.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.
Vapnik, V., & Chervonenkis, A. (1964). A note on one class of perceptros. *Automation and Remote Control, 25*, 112-120.

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. Automation and Remove Control. *Automatio and Remote Control,, 24*, 774-780.

Wang, L., Zuo, Z., & Fu, J. (2014). Bus Arrival Time Prediction Using RBF Neural Networks Adjusted by Online Data. *Procedia-Social and Behavioral Sciences*(138), 67-75.

Xinghao, S., Jing, T., Guojun, C., & Qichong, S. (2013). Predicting bus real-time travel time basing on both GPS and RFID data. *Procedia-Social and Behavioral Sciences, 93*, 2287-2299.

Xiong, G., & et al. (2015). Continuous Travel Time Prediction for Transit Signal Priority Based on a Deep Network. *IEEE 18th International Conference on Intelligent Transportation Systems* (págs. 523-528). IEEE Computer Society.

Yu, B., Lam, W. H., & Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, *19(6)*, págs. 1157-1170.

Zheng, C. J., Zhang, Y. H., & Feng, X. J. (2012). Improved iterative prediction for multiple stop arrival time using a support vector machine. *Transport, 27(2)*, 158-164.

Zhong, S., Hu, J., Ke, S., Wuang, X., Zhao, J., & Yao, B. (2015). A Hybrid Model Based on Support Vector Machine for Bus Travel-Time Prediction. *Traffic & Transportation, 27*, 291-300.

ANEXOS

Anexo A: Cumplimiento de supuestos de Regresión Lineal Múltiple

Los modelos de regresión lineal tienen una serie de supuestos que no siempre se cumplen. Estos supuestos son:

- Errores aleatorios tienen una distribución normal homocedastica.
- Ausencia de multicolinealidad.

El supuesto de la normalidad de los errores aleatorios puede ser medido con los test Shapiro-Wilk y Lilliefors, los cuales tienen como hipótesis nula que la muestra de residuos aleatorios tiene una distribución normal. A su vez, la homocedasticidad es probada con el test Breusch-Pagan, el cual tienen como hipótesis nula la existencia de homocedasticidad de los errores aleatorios, donde se verifica si su varianza es explicada por un subconjunto de las variables explicativas incluidas en el modelo de regresión lineal.

Los resultados de estos test se pueden ver en la Tabla A-1. Del test Shapiro-Wilk y Lilliefors vemos que al tener valor *p* menor al valor α escogido (se elige α de forma de tener un 95% de confianza), entonces se rechaza la hipótesis nula de distribución normal de los residuos. A su vez, los resultados del test de Breush-Pagan rechazan la hipótesis de homocedasticidad. Por ende, el supuesto de que los errores aleatorios tienen una distribución normal homocedastica se rechaza para el modelo MLR de los tres servicios de buses analizados.

Tabla A - 1. Resultados test para validar supuestos de modelo MLR

Test	Estadístico			Valor α
	212R	203R	C04I	
Shapiro-Wilk	0,00 (W=0,865)	0,00 (W=0,852)	0,00 (W=0,874)	0,05
Lilliefors	0,00 (D=0,122)	0,00 (D=0,115)	0,00 (D=0,108)	0,05
Breush-Pagan	0,00 (BP=1619)	0,00 (BP=1451)	0,00 (BP=465)	0,05

El Supuesto de ausencia de multicolinealidad se puede medir con el test de VIF (del inglés, *Variance Inflation Factor*). Donde variables con valores superiores a un cierto

valor critico indicarían presencia de multicolinealidad. La Tabla A-2 muestra los coeficientes del test VIF obtenidos para cada variable en cada servicio de bus. No hay consenso de que valor crítico ocupar, por lo que se utilizó el valor cuatro, el cuál es el menor valor crítico utilizado en la literatura (O'brien, 2007). Vemos que ningún valor de la tabla sobrepasa el valor cuatro, por lo que se puede afirmar que no hay una presencia de multicolinearidad significativa.

Tabla A - 2. Resultados Test VIF

Variables	Valor Coeficiente VIF		
	212R	203R	C04I
Distancia	1,16	1,06	1,18
Lluvia	1,01	1,00	-
Carga Bus	1,20	1,37	2,06
Subidas Pagadas	1,18	1,10	1,24
Bajadas	1,32	1,19	1,32
noSniB	1,61	1,67	1,99
Corredor	1,34	1,16	-
Vel. Tiempo Real$_{[p-3\ ,\ p]\ [t-\ 15,\ t]}$	1,17	1,17	1,08
Vel. Tiempo Real$_{[p\ ,\ p+3]\ [t-15\ ,\ t]}$	1,19	1,21	1,12
Dif. Vel. Histórica $_{[p\ ,\ p+3]\ [t-15\ ,\ t]}$	1,44	1,38	1,22
Dif. Vel. Histórica $_{[p-3\ ,\ p]\ [t\ ,\ t+15]}$	1,18	1,17	1,03
Dif. Vel. Histórica $_{[p\ ,\ p+3]\ [t\ ,\ t+15]}$	1,30	1,38	1,04

Anexo B: Peso de las variables explicativas en un modelo MLR ajustado en los tres servicios de buses.

De la Tabla B-1 se pueden ver los valores de los coeficientes para cada variable en un modelo de Regresión Lineal Múltiple calibrado para cada servicio de buses. Como las variables están normalizadas; es decir, poseen media cero y varianza igual a uno, el valor del coeficiente muestra el peso que tiene la variable explicativa sobre la dependiente (velocidad del bus en el tramo a predecir). En general, las variables tienen un peso similar entre los distintos servicios, por lo que su efecto no es específico a cada servicio de buses y pareciera ser extrapolable a toda la red.

Tabla B - 1. Valores de los coeficientes y del valor-t de las variables en modelos MLR ajustados para cada servicio

	212R		**203R**		**C04I**	
	Coef.	Valor-t	Coef.	Valor-t	Coef.	Valor-t
(Intercepto)	1,7E-16	0,0	-7,1E-16	0,0	1,1E-16	0,0
Distancia	-8,4E-03	-1,7	-1,2E-03	-0,3	3,1E-02	3,9
Corredor	-3,2E-03	-0,6	4,8E-02	12,5	-	
Lluvia	-3,5E-03	-0,9	-5,2E-04	-0,2	-	
Carga del Bus	-6,5E-03	-1,2	-8,7E-03	-2,1	4,8E-02	3,7
Subidas Pagadas	-3,7E-02	-7,7	-5,8E-02	-15,6	-6,3E-02	-6,5
Bajadas	1,6E-02	3,0	1,9E-02	4,8	-3,7E-03	-0,3
No hay Sub. ni Baj.	1,2E-01	22,3	8,9E-02	20,7	1,2E-01	11,5
$Vel.Tiempo\ Real_{\ [p-3\ ,p][t-\Delta t\ ,t]}$	-2,9E-02	-5,6	-4,1E-02	-9,4	1,7E-03	0,2
$Vel.Tiempo\ Real_{\ [p\ ,p+3][t-\Delta t\ ,t]}$	7,3E-01	122,5	7,0E-01	138,4	8,2E-01	98,0
$Dif.Vel.Histórica_{\ [p\ ,p+3][t-\Delta t\ ,t]}$	5,0E-02	8,0	6,7E-02	13,0	-6,8E-02	-8,9
$Dif.Vel.Histórica_{\ [p-3\ ,p][t\ ,t+\Delta t]}$	-9,6E-03	-2,0	1,9E-03	0,5	-1,1E-02	-1,6
$Dif.Vel.Histórica_{\ [p\ ,p+3][t\ ,t+\Delta t]}$	-1,5E-01	-24,1	-1,6E-01	-30,6	-1,3E-01	-18,2

Anexo C: Gráficos de dispersión de predicciones

a. Servicio 212R

En la Figura C-1 se puede ver los distintos gráficos de dispersión de las predicciones realizadas por los distintos modelos calibrados al servicio 212R. En el eje x se tiene la velocidad real que tuvo el bus, y en el eje y, la velocidad predicha por el modelo correspondiente. La línea roja que cruza diagonalmente cada gráfico, es una línea de pendiente uno e intercepto igual a cero; es decir, es la línea que mientras más cercana está la predicción a ella (punto negro), mejor fue esta.

De los gráficos se puede ver que en general se tiende a sobreestimar las velocidades bajas y a subestimar las velocidades. Respecto a estas últimas, se ve que rara vez el modelo predice velocidades sobre los 40 km/hr, aun cuando en la realidad estas alcancen valores cercanos a los 70 km/hr. Esto muestra que es difícil predecir velocidades altas y que probablemente sean hechos aislados causado por situaciones fortuitas (varias luces en verde seguidas o algún otro hecho).

A simple vista, no se puede determinar qué modelo es mejor. Sin embargo, si se puede ver que el modelo *benchmark* histórico, se desempeña peor que el resto. Donde la nube de puntos pareciera ser más plana y alejada de la línea roja. Esto lo confirma el valor del R^2, valor el cual muestra la relación entre los predictores y la variable a predecir para cada modelo utilizado. Donde mientras más alto es este, mayor relación hay entre dichas variables. Acorde con los resultados del RMSE presentados en la Sección 6.1, el modelo que presenta el mayor R^2 es ANN, seguido muy de cerca por MLR y luego por SVM. Un poco más alejado, pero con una buena relación todavía, se encuentra el modelo BTR, y ya con una relación bastante más pobre el modelo BH.

Al reducir el tamaño de los gráficos, se pierde un poco la visual de cómo están realmente distribuidos los puntos de velocidad predicha versus real. En la Figura C-2 se puede ver en grande el gráfico de dispersión del modelo ANN, y de esa forma tener una mejor percepción de las predicciones y la distribución de estas.

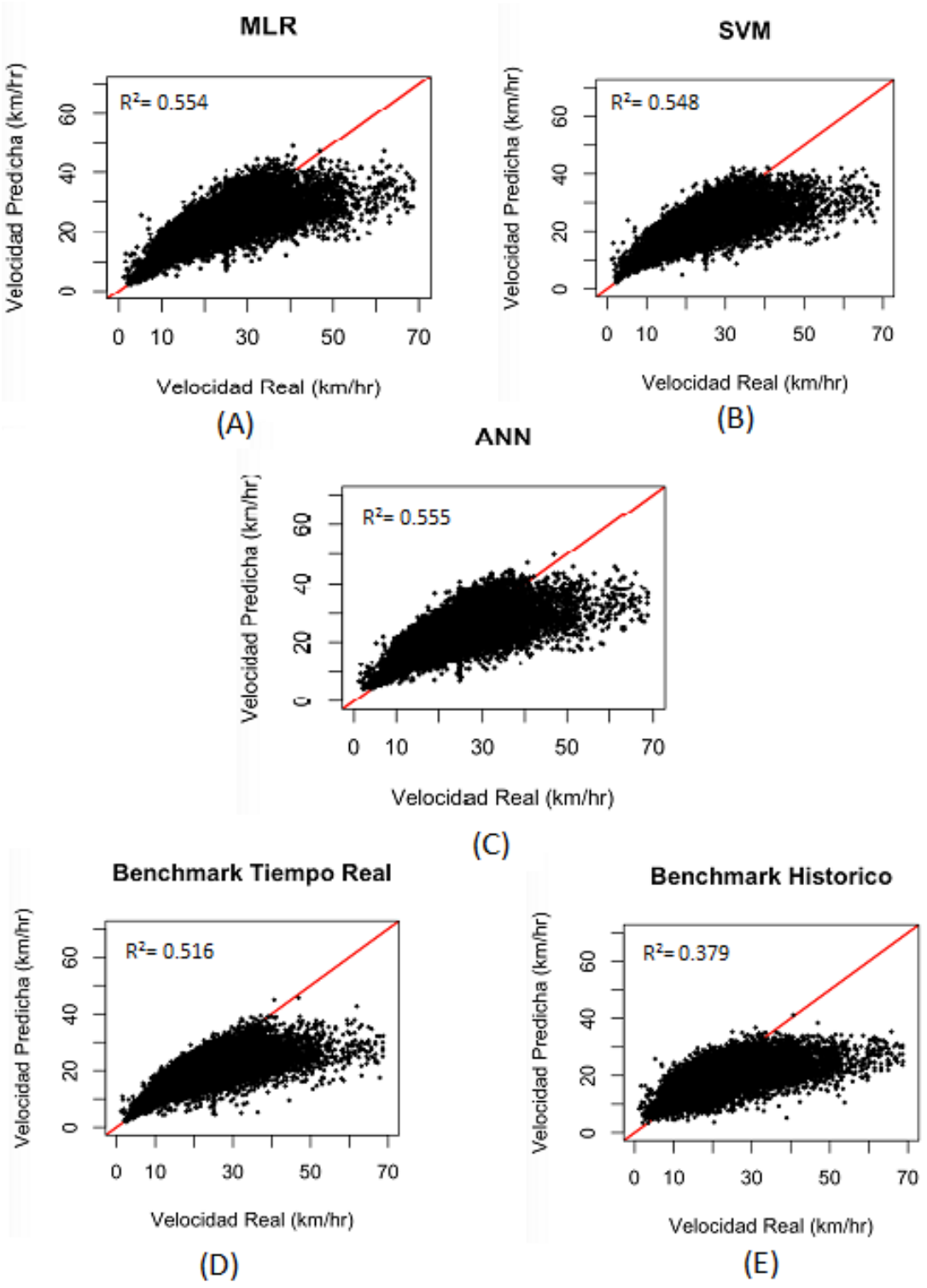


Figura C - 1. Gráficos de dispersión de las predicciones para los distintos modelos ajustado al servicio 212R

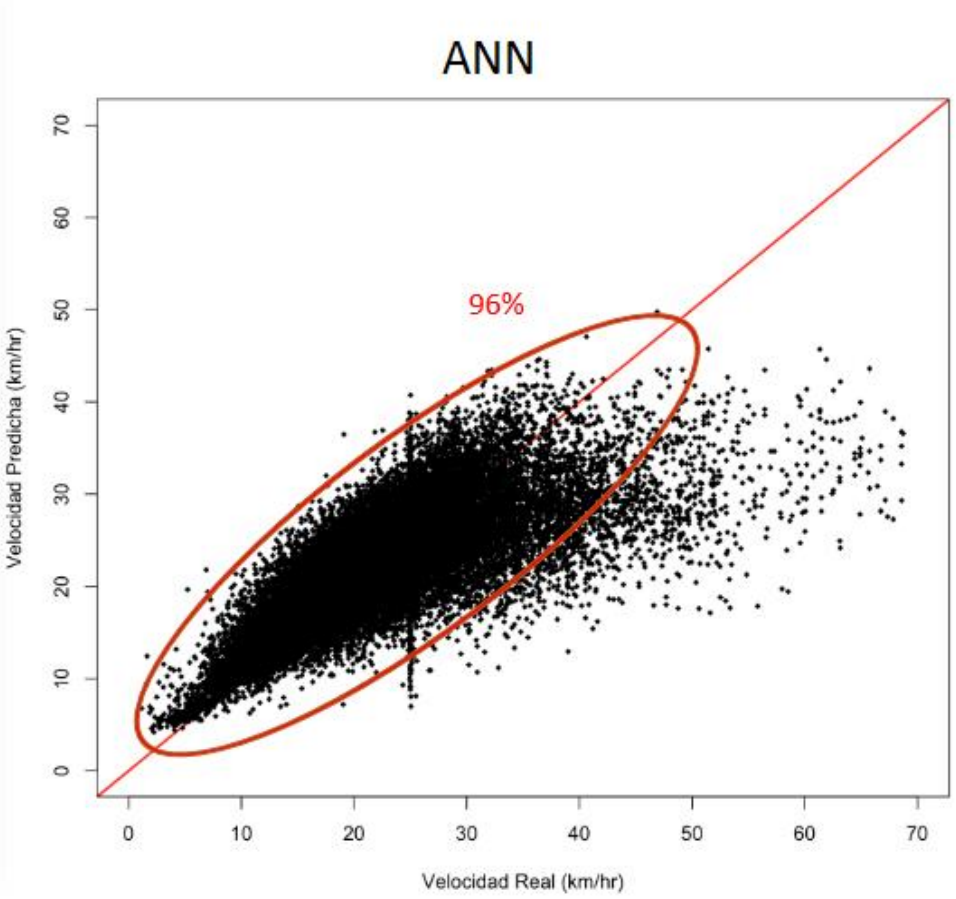


Figura C - 2. Gráfico de dispersión del modelo ANN en el servicio 212R

En esta figura se observa que la cola de velocidades altas que están siendo subestimadas por el modelo, representa menos de un 4% de los datos. Por lo que en general el modelo predice bastante bien al menos el 96% de las velocidades.

b. Servicio 203R

Los gráficos se dispersión de las predicciones realizadas por los distintos modelos en el servicio 203R se pueden ver en la Figura C-3. La notación de estos es la misma que la utilizada en el servicio 212R, por lo que se recomienda ver explicación en Anexo C-a para entender los gráficos.

Al igual que en el servicio 212R, se tiende a sobreestimar las velocidades bajas y subestimar las velocidades altas. Donde nuevamente en la predicción de velocidades altas es donde se tiene más problemas, con estimaciones bastante inferiores a sus valores reales.

De los valores R^2 de cada modelo, se puede ver que los modelos de aprendizaje estadístico explican mejor las relaciones entre las variables explicativas y la dependiente (velocidad del bus). No así, los modelos *benchmark*, donde el modelo BTR tiene un valor del R^2 superior al modelo BH, pero este sigue siendo más bajo que el resto.

Finalmente, a simple vista se puede ver que el modelo BH predice peor. Sin embargo, los gráficos de dispersión del resto de los modelos son bastante similares, por lo que se tiene que recurrir a las medidas de desempeño para poder rankear el funcionamiento de estos.

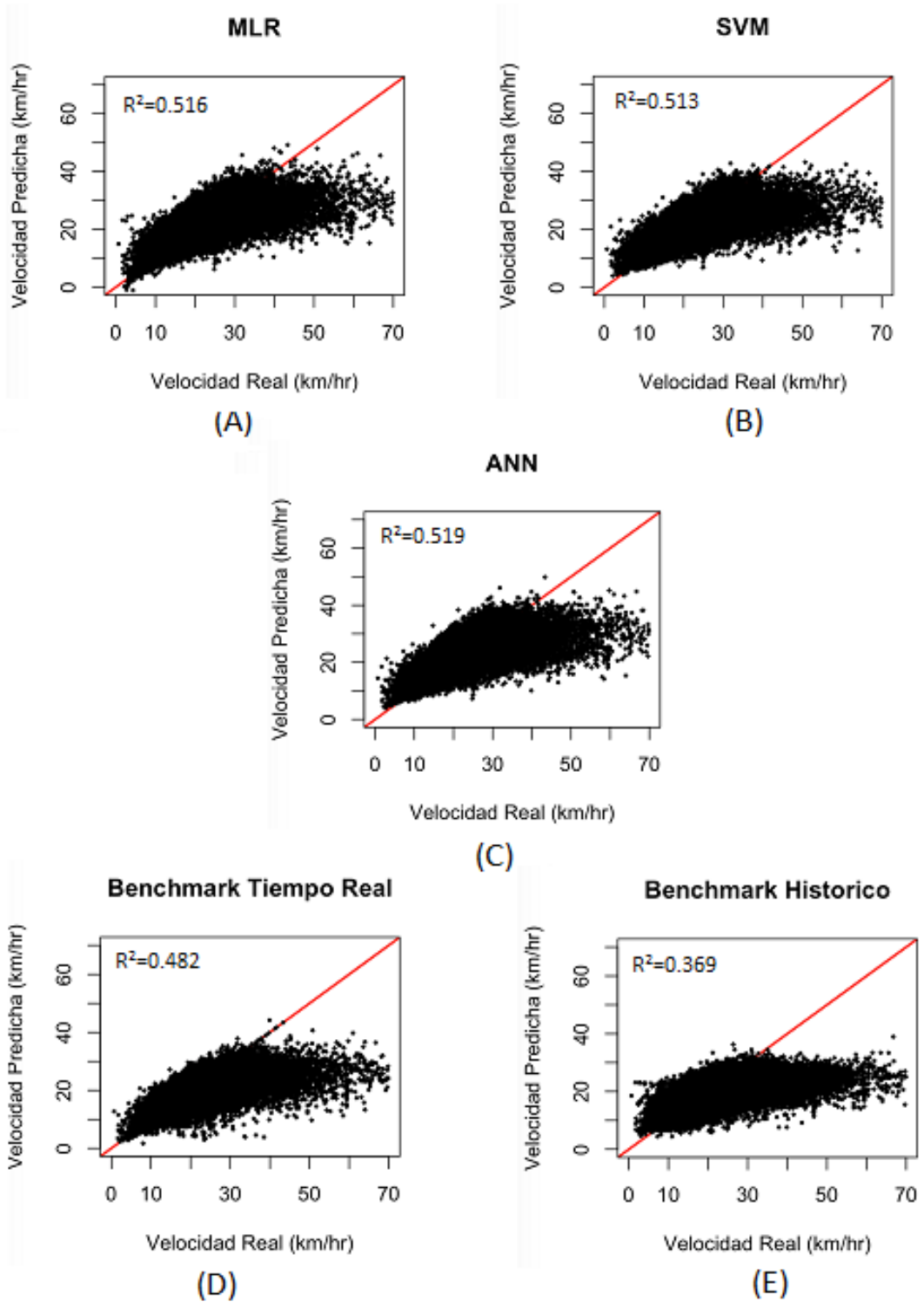


Figura C - 3. Gráficos de dispersión de las predicciones para los distintos modelos ajustado al servicio 203R

c. Servicio C04I

Los gráficos de dispersión del servicio C04I se pueden ver en la Figura C-4. Se recomienda utilizar el Anexo C-a como referencia para poder interpretar de mejor manera los gráficos presentados.

De los gráficos se observa, que a diferencia de los otros servicios, en este rara vez se pasa los 50 km/hr lo que facilita las predicciones. A su vez, los gráficos se ven más compactos y cercanos a la línea roja diagonal, lo que explica que en este servicio se haya tenido errores menores. Incluso, el valor de R^2 alcanzado por los modelos es bastante superior a los otros servicios, donde nuevamente los modelos de aprendizaje estadístico destacan por sobre los modelos *benchmark*.

A simple vista, pareciera que las velocidades bajas se estiman bien, pero en las altas se vuelve a subestimar. No obstante, se subestima de menor manera que los otros servicios, lo que podría ser explicado por ser un servicio sin corredores segregados y por ende tener velocidades menores entre los tramos.

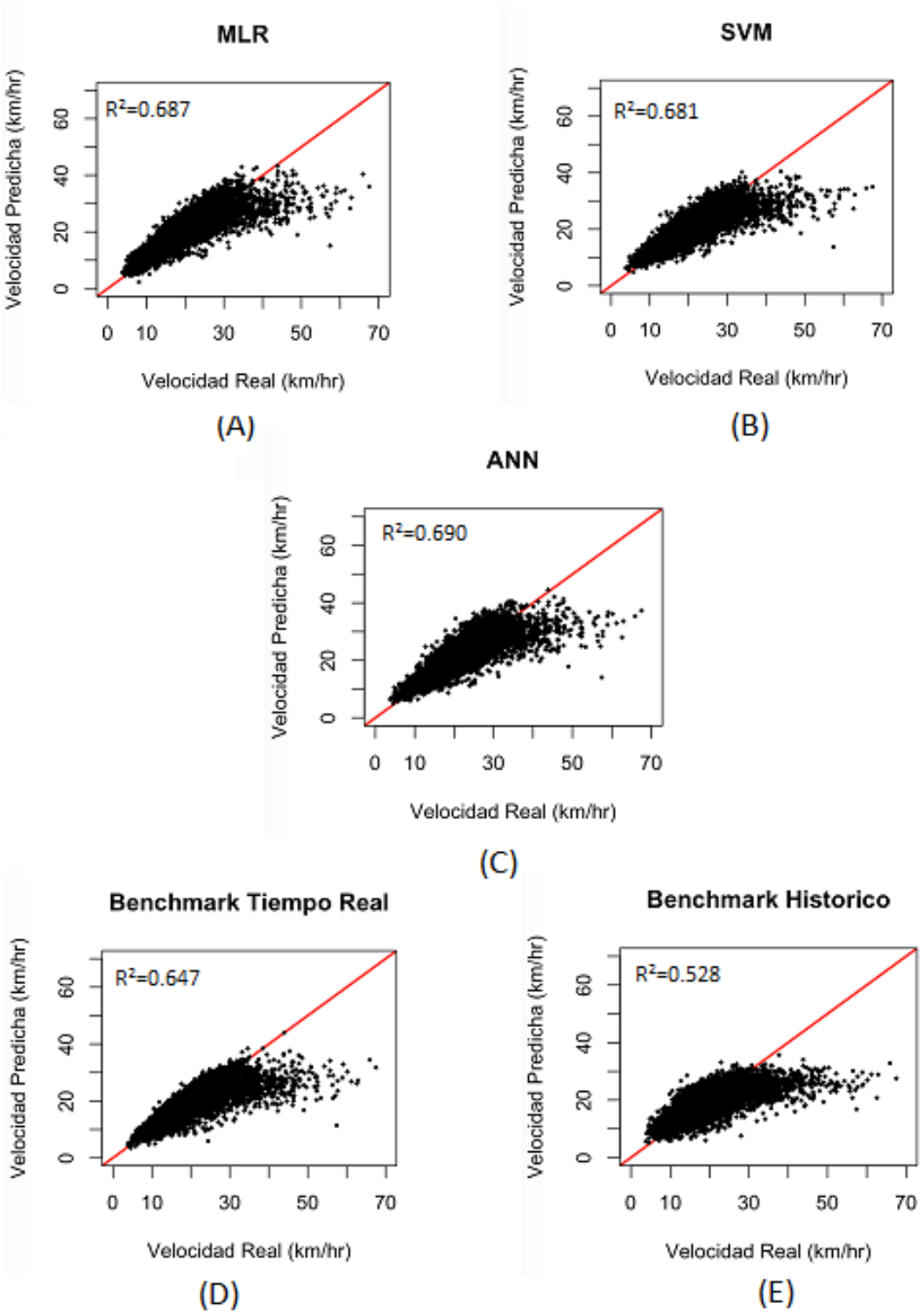


Figura C - 4.Gráficos de dispersión de las predicciones para los distintos modelos ajustado al servicio C04I