



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

# **PREDICCIÓN DE TIEMPO DE VIAJE CON INFORMACIÓN DE EJES PARALELOS Y BAJA FRECUENCIA**

**RAMÓN ANDRÉS BAHAMONDE BIRKE**

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

Profesores Supervisores:

RICARDO GIESEN

HANS LÖBEL

Santiago de Chile, Junio 2018

© 2018, RAMÓN ANDRÉS BAHAMONDE BIRKE



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

# **PREDICCIÓN DE TIEMPO DE VIAJE CON INFORMACIÓN DE EJES PARALELOS Y BAJA FRECUENCIA**

**RAMÓN ANDRÉS BAHAMONDE BIRKE**

Miembros del Comité:

RICARDO GIESEN

HANS LÖBEL

FELIPE DELGADO

MARCELA MUNIZAGA

ESTEBAN SAEZ

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

Santiago de Chile, Junio 2018

© 2018, RAMÓN ANDRÉS BAHAMONDE BIRKE

*A mi familia*

## AGRADECIMIENTOS

En mi primer lugar quiero agradecer a mis profesores guías, Ricardo Giesen y Hans Löbel, cuyo aporte fue crucial en el desarrollo de este trabajo. A Ricardo, por su gran experiencia en el área y su motivación sobre el tema de la presente investigación, lo que facilitó que iniciara esta etapa. A Hans, por su disponibilidad permanente para resolver cualquier tipo de duda y sus invaluable consejos en momentos finales y críticos de este largo proceso, sin los cuales la dificultad habría sido mayor. Además, a Christopher Bucknell, por su ayuda en la obtención de información con la que se pudo dar inicio al desarrollo esta investigación. Agradezco al proyecto Fondecyt N°1171049, al Centro de Desarrollo Urbano Sustentable (CEDEUS, CONICYT/FONDAP 15110020) y BRT+ Centre of Excellence.

También deseo agradecer a los miembros del Departamento de Ingeniería de Transporte y Logística, con quienes se genera un clima agradable y distendido, donde se puede compartir entre docentes y alumnos, lo que contribuye a la generación de un ambiente de trabajo motivador. A mis compañeros de postgrado, donde cada uno aporta algo para generar este ambiente grato, de colaboración y ayuda, donde se pasa bien, se comparte, pero se trabaja en forma intensa. En general, fue muy grato pertenecer a este departamento.

Finalmente, termino agradeciendo a mi familia por su apoyo durante estos años. A pesar de la distancia, estuvieron siempre presentes de alguna forma, con consejos o visitas, con palabras oportunas, lo que permitió que la distancia no constituyera una dificultad adicional durante mi vida universitaria. Y, además, agradecer a mis amigos, con quienes compartí durante estos años y me brindaron momentos de alegría y apoyo cuando fue necesario.

## ÍNDICE GENERAL

AGRADECIMIENTOS	IV
Índice de figuras	VIII
Índice de tablas	X
ABSTRACT	XI
RESUMEN	XII
Capítulo 1. INTRODUCCIÓN	1
Capítulo 2. REVISIÓN BIBLIOGRÁFICA	4
2.1. Modelos de aprendizaje de máquina . . . . .	4
2.1.1. Regresión Lineal (Lineal Regression, LR) . . . . .	4
2.1.2. Regresión de Vectores de Soporte (Support Vector Regression, SVR) . . . . .	6
2.1.3. Redes Neuronales (Neuronal Network, NN) . . . . .	8
2.2. Trabajos Anteriores . . . . .	10
Capítulo 3. DATOS	15
3.1. Origen Datos . . . . .	15
3.2. Procesamiento . . . . .	16
3.2.1. Construcción de Trayectos . . . . .	17
3.2.2. Revisión de datos . . . . .	18
3.2.3. Generación de horarios . . . . .	20
3.3. Consolidación de los datos . . . . .	21
Capítulo 4. METODOLOGÍA	22
4.1. Variables modelo . . . . .	22
4.1.1. Diagrama Variables Vector Input . . . . .	23

4.2.	Muestreo y Estructura . . . . .	24
4.3.	Medidas de Desempeño . . . . .	26
4.4.	Calibración Modelos y Benchmark de comparación . . . . .	27
4.4.1.	Promedio Histórico(Historical Average, HA) . . . . .	27
4.4.2.	Promedio Último Periodo(Last Period Average, LPA) . . . . .	28
4.4.3.	Regresión Lineal(LR) . . . . .	28
4.4.4.	Regresión de Vectores de Soporte(SVR) . . . . .	29
4.4.5.	Redes Neuronales(NN) . . . . .	29
4.5.	Validación Cruzada . . . . .	29
Capítulo 5. CASO DE ESTUDIO		31
5.1.	Zona . . . . .	31
5.1.1.	Pedro de Valdivia(PdV) . . . . .	32
5.1.2.	Los Leones(LL) . . . . .	33
5.1.3.	Manuel Montt(MM) . . . . .	34
5.2.	Comparación Calles . . . . .	36
Capítulo 6. RESULTADOS		38
6.1.	Puntos Comparación . . . . .	39
6.1.1.	Información Histórica . . . . .	39
6.1.2.	Información Actual . . . . .	40
6.1.3.	Información Actual sin Calle Interés . . . . .	41
6.1.4.	Histórico e Información Actual . . . . .	43
6.2.	SVR por Hora . . . . .	44
6.2.1.	Información Actual 2.5 km . . . . .	44
6.2.2.	Información Histórica 2.5 km . . . . .	48
Capítulo 7. CONCLUSIONES		52
7.1.	Principales Contribuciones y Resultados . . . . .	52
7.2.	Trabajo Futuro y Extensiones . . . . .	54

BIBLIOGRAFÍA 56

Anexo 59

    A. Resultados SVR por Hora 0.5 km con Información Actual . . . . . 60

    B. Resultados SVR por Hora 0.5 km con Información Histórica . . . . . 63

## ÍNDICE DE FIGURAS

2.1.	Regresión Lineal . . . . .	5
2.2.	Vectores de Soporte . . . . .	7
2.3.	Capas Red Neuronal . . . . .	9
2.4.	Nodo i Red Neuronal . . . . .	9
3.1.	Velocidad Semana 3 de Abril . . . . .	16
3.2.	Trayecto D09 . . . . .	18
3.3.	Generación de trayectos recorrido D09 por cambio de sentido . . . . .	19
3.4.	Selección de Horario . . . . .	20
4.1.	Diagrama Variables . . . . .	24
4.2.	Diagrama HA . . . . .	27
4.3.	Diagrama LPA . . . . .	28
4.4.	Diagrama Validación Cruzada . . . . .	30
5.1.	Zona Estudio . . . . .	31
6.1.	SVR MAPE 2.5 km Información Actual 00:00-08:00 . . . . .	45
6.2.	SVR MAPE 2.5 km Información Actual 08:00-16:00 . . . . .	47
6.3.	SVR MAPE 2.5 km Información Actual 16:00-00:00 . . . . .	48
6.4.	SVR MAPE 2.5 km Histórico 00:00-08:00 . . . . .	49
6.5.	SVR MAPE 2.5 km Histórico 08:00-16:00 . . . . .	50

6.6. SVR MAPE 2.5 km Histórico 16:00-00:00 . . . . . 50

A.1. SVR MAPE 0.5 km Información Actual 00:00-08:00 . . . . . 60

A.2. SVR MAPE 0.5 km Información Actual 08:00-16:00 . . . . . 61

A.3. SVR MAPE 0.5 km Información Actual 16:00-00:00 . . . . . 62

B.1. SVR MAPE 0.5 km Histórico 00:00-08:00 . . . . . 63

B.2. SVR MAPE 0.5 km Histórico 08:00-16:00 . . . . . 64

B.3. SVR MAPE 0.5 km Histórico 16:00-00:00 . . . . . 65

## ÍNDICE DE TABLAS

2.1.	Resumen Trabajos Anteriores . . . . .	11
5.1.	Resumen Pedro de Valdivia . . . . .	33
5.2.	Resumen Los Leones . . . . .	34
5.3.	Resumen Manuel Montt . . . . .	35
5.4.	Comparación Tiempos de Viaje 0.5 km . . . . .	36
5.5.	Comparación Tiempos de Viaje 2.5 km . . . . .	37
6.1.	Comparación Porcentaje de Datos . . . . .	38
6.2.	Comparación Base . . . . .	39
6.3.	Histórico 2.5 km . . . . .	40
6.4.	Histórico 0.5 km . . . . .	40
6.5.	Información Actual 2.5 km . . . . .	41
6.6.	Información Actual 0.5 km . . . . .	41
6.7.	Información Actual Sin Calle Interés 2.5 km . . . . .	42
6.8.	Información Actual Sin Calle Interés 0.5 km . . . . .	42
6.9.	Histórico e Información Actual 2.5 km . . . . .	43
6.10.	Histórico e Información Actual 0.5 km . . . . .	43

## ABSTRACT

The decrease in technology costs has allowed to massively provide tracking devices to vehicles, increasing the availability and collection of global positioning information (GPS). As a consequence, new methodologies have been developed on the basis of this data, allowing a better estimation of travel time for vehicle fleets. However, the implementation of these methods is associated with large data requirements.

In order to face the problem of data volume, this work proposes a methodology that allows estimating travel times with limited GPS positioning information. For this purpose, travel time information of vehicles on parallel axes is incorporated as input. The method is evaluated on the basis of GPS information generated by the public transport system Transantiago. The tests are performed using linear regression models, support vectors regressions and neural networks.

The results indicate that using travel time information from parallel axes allows improving the estimates if the axes are chosen appropriately. Moreover, the exclusive use of information from parallel axes allows obtaining good estimates when there is no current information of the axis in question.

Keywords: travel time prediction, linear regression, support vectors regression, neural networks, parallel axes, low frequency, low amount of data, real time, GPS.

## RESUMEN

La disminución de costos de la tecnología ha permitido dotar masivamente de dispositivos de localización a los vehículos, incrementado la disponibilidad y recolección de información de posicionamiento global (GPS). Debido a lo anterior, se han desarrollado nuevas metodologías que permiten, a partir de este tipo de datos, enfrentar de mejor manera el problema de la estimación de tiempo de viaje para flotas de vehículos. Sin embargo, la implementación de estos métodos requiere de altos volúmenes de información.

Con el fin de enfrentar el problema de volumen de datos, en este trabajo se plantea una metodología que permite realizar estimaciones de tiempo de viaje cuando se posee información de posicionamiento GPS limitada. Para ello, se incorpora información asociada al tiempo de viaje de vehículos en ejes paralelos. El método es evaluado usando información GPS generada por el sistema de transporte público Transantiago. Las pruebas se realizan utilizando modelos de regresión lineal, regresión de vectores de soporte y redes neuronales.

Los resultados indican que la información de tiempo de viaje en ejes paralelos permite mejorar las estimaciones si se escoge apropiadamente los ejes a utilizar. Además, el uso exclusivo de información de ejes paralelos permite obtener buenas estimaciones cuando no se posee información actual del eje en cuestión.

Palabras Claves: predicción tiempo de viaje, regresión lineal, regresión de vectores de soporte, redes neuronales, ejes paralelos, baja frecuencia, baja cantidad de datos, tiempo real, GPS.

## **CAPÍTULO 1. INTRODUCCIÓN**

La predicción de velocidades es un problema que deben enfrentar distintos tipos de flotas de vehículos. Tanto aquellas que realizan ruteos distintos cada día, como aquellas que circulan por el mismo trayecto diariamente, como es el caso de vehículos de transporte público. Una buena predicción de velocidad permite que los operadores puedan planificar de mejor manera los tiempos que sus vehículos necesitarán para realizar un cierto recorrido. Una mejor planificación de los vehículos impactará tanto en sus costos, como en la calidad de servicio entregado. Esto puede ser reflejado en tiempos más acotados en la entrega de productos con envío a domicilio o mayor exactitud en los horarios de arribo a destino lo que disminuye la ansiedad de un usuario que espera de un producto o que espera en un paradero un servicio de transporte. También permite que los conductores tengan mayor certeza en sus tiempos de viaje y hora de arribo.

El uso de datos de posicionamiento GPS e información en tiempo real es un enfoque que se ha utilizado de manera exitosa en el pasado para la estimación de velocidad o tiempo de viaje. Mediante este tipo de información, estudios anteriores se han centrado en los tiempos de viaje por arcos utilizando una gran flota de vehículos (Jenelius y Koutsopoulos, 2013; Zheng y Van Zuylen, 2013; Zhan, Hasan, Ukkusuri, y Kamga, 2013). En general, estos estudios reportan una estimación de arcos entre intersecciones, pero con una flota de taxis que permite tener vehículos circulando constantemente por los arcos. Otros estudios consideran rutas que comprenden varios arcos (Wang, Zheng, y Xue, 2014; Rahmani, Jenelius, y Koutsopoulos, 2015), pero al igual que los estudios anteriores, con grandes flotas de vehículos. Estos estudios hacen uso de diversas herramientas para la estimación, como modelos de aprendizaje, modelos de optimización o máxima verosimilitud, logrando buenos resultados.

Un aspecto que no cubren los trabajos anteriormente mencionados es el problema que enfrentan flotas pequeñas al tener una baja cantidad de datos o ausencia de éstos en una

ruta o arco que se requiera estimar. En muchos casos, las velocidades de operaciones de distintos tipos de flotas no se asemejan entre ellas. Por esta razón, no se puede utilizar la velocidad de otra flota directamente como predictor. Debido a esto, es complejo tener buenas estimaciones de velocidad cuando se tiene una flota pequeña, que reporte pocos datos, o cuando se tiene una muestra parcial de datos para realizar una predicción de un arco o ruta de interés.

Con el fin de enfrentar de mejor manera el problema anteriormente mencionado, en este trabajo proponemos estudiar el efecto sobre las estimaciones de tiempos de viaje al incorporar información asociada a calles paralelas al arco que se requiere estimar. De esta forma, se pretende tener una visión global de la zona, que permita mejorar la estimación cuando se tiene una baja cantidad de datos o ausencia de éstos en el eje de interés. Para ello, se utilizarán 3 modelos de aprendizaje, regresión lineal múltiple, regresión de vectores de soporte y redes neuronales. Los resultados serán comparados con promedios históricos y promedios de último periodo mediante datos en tiempo real. Para estudiar el efecto de las calles paralelas se utilizarán las posiciones GPS reportadas por el sistema de transporte público de Santiago, Transantiago (DTPM, 2017). Las posiciones GPS de cada bus son reportadas cada 30 segundos y con frecuencias de hasta 30 buses por hora en algunos de los ejes escogidos para el estudio. Por ello y para simular una flota pequeña, se trabaja con submuestras de los buses que circulan por cada avenida.

La tesis se encuentra dividida en 7 capítulos. En el capítulo 2 se realiza una revisión bibliográfica de los modelos que son utilizados para las estimaciones y, posteriormente, se presenta un resumen de los estudios que han abordado la estimación de tiempos de viaje o velocidad con datos GPS en tiempo real. Luego, en el capítulo 3 se detalla el origen de los datos utilizados y el preprocesamiento necesario para su uso. El capítulo 4 presenta la metodología aplicada para calibrar cada uno de los modelos y las medidas de

desempeño con las que éstos se comparan. En el capítulo 5 se introduce el caso de estudio con características de la zona y tiempos de viaje para los ejes seleccionados. El capítulo 6 presenta los resultados para cada uno de los modelos y un detalle del mejor modelo bajo las condiciones propuestas en el estudio. Finalmente, en el capítulo 7, se presentan las conclusiones y posibles extensiones del trabajo desarrollado.

## **CAPÍTULO 2. REVISIÓN BIBLIOGRÁFICA**

En esta sección se realiza una revisión de los modelos que serán utilizados para la predicción de tiempos de viaje. Estos corresponden a regresión lineal, regresión de vectores de soporte y redes neuronales. También se presenta una revisión de distintos trabajos que han tratado la estimación de tiempo de viaje o velocidad utilizando datos GPS con distintos enfoques.

### **2.1. Modelos de aprendizaje de máquina**

La detección de patrones en bases de datos, con el fin de realizar una predicción de variables, es una tarea que se realiza mediante métodos de aprendizaje automático, que deben ser entrenados y calibrados con información del pasado. Los métodos de aprendizaje se pueden utilizar para predecir categorías o valores de variables dependiendo de los requerimientos.

Para el análisis propuesto se utilizarán tres modelos de aprendizaje, regresión lineal múltiple, regresión de vectores de soporte y redes neuronales. Se utilizará entrenamiento supervisado y se realizará predicción de un valor de una variable. A continuación, se explica en que consiste cada uno de estos métodos.

#### **2.1.1. Regresión Lineal (Lineal Regression, LR)**

La regresión lineal es un modelo de aprendizaje estadístico que busca predecir el valor de una variable dependiente mediante una combinación lineal de variables independientes (descriptivas), ponderadas por un parámetro  $\beta$  que debe ser estimado. La ecuación (2.1) representa una regresión lineal con  $Y$  como variable dependiente,  $X_i$  como variables independientes.  $\epsilon$  es un término de error que captura la información que el modelo no logra capturar con las variables utilizadas (ya sea por falta de información o por la existencia de procesos estocásticos).

$$Y = \beta_0 + \sum_{i=1}^i \beta_i X_i + \epsilon \quad (2.1)$$

Respecto a los parámetros  $\beta$ , éstos se pueden obtener mediante una serie de métodos de estimación. En el contexto de este trabajo se utiliza una estimación por mínimos cuadrados ordinarios. Esto implica calcular los valores para  $\beta$  que minimizan el error cuadrático de la regresión lineal, tal como se indica en la ecuación (2.2) donde  $Y$  representa el valor real e  $\hat{Y}$  representa el valor estimado.

$$RSS(\beta) = \sum_{j=1}^j (Y_j - \hat{Y}_j)^2 \quad (2.2)$$

En la figura 2.1 se presenta una regresión lineal de una variable dependiente que minimiza el error cuadrático entre la estimación y el valor real.

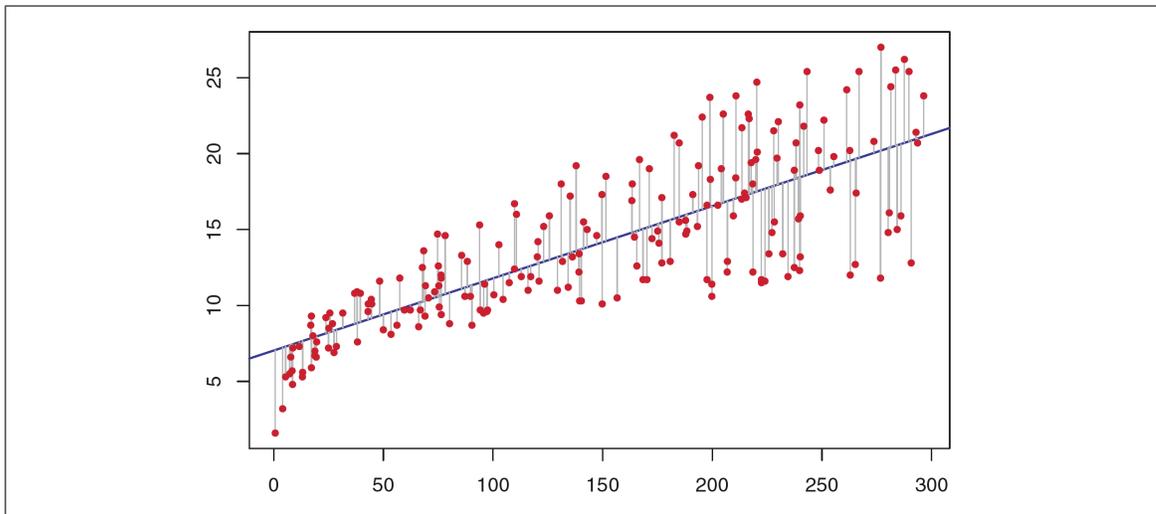


Figura 2.1. Regresión Lineal  
Fuente: (Tibshirani et al., 2013)

Para más detalles sobre modelos de regresión lineal múltiple se recomienda revisar (Friedman, Hastie, y Tibshirani, 2001) y (Tibshirani et al., 2013).

### 2.1.2. Regresión de Vectores de Soporte (Support Vector Regression, SVR)

La regresión de vectores de soporte (SVR) es un tipo de modelo desarrollado a partir de las máquinas de vectores de soporte por (Vapnik y Lerner, 1963; Vapnik y Chervonenkis, 1964) en los años 60. Son un tipo de algoritmo de aprendizaje, utilizado para clasificación de datos y que busca determinar una superficie que separe el conjunto de datos, en un determinado número de subconjuntos (categorías) resolviendo un problema de maximización de la distancia entre éstos. Las superficies de decisión pueden tener distintas formas, las que dependen del tipo de función que se utilice. Estas funciones son denominadas kernel y permiten obtener superficies de decisión no lineales, pero que mantiene el problema matemático lineal. Los kernel pueden ser lineales, polinomiales, sigmoidales, radiales, etc. A continuación, se presenta la formulación lineal de este algoritmo, donde  $w^t$  representa el conjunto de parámetros del modelo y  $x^t$  el vector de escalares:

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 \\ \text{sa:} \quad & r^t(w^t x^t + w_0) \geq +1 \quad \forall t \end{aligned} \tag{2.3}$$

A la ecuación (2.3) se aplican los multiplicadores de Lagrange ( $\alpha$ ) y se obtiene la formulación dual del SVM que permite apreciar de manera clara las características centrales de este:

$$\begin{aligned} \max \quad & \sum_t \alpha^t - \frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s (x^t)^T x^s \\ \text{sa:} \quad & \sum_t \alpha^t r^t = 0 \\ & \alpha^t \geq 0 \quad \forall t \end{aligned} \tag{2.4}$$

Al solucionar el problema de maximización (2.4) para  $\alpha^t$ , la mayoría de los valores corresponde a  $\alpha^t = 0$ . Solo en un pequeño porcentaje se obtiene  $\alpha^t > 0$  y estos son los  $x^t$

llamados vectores de soporte que construyen la superficie de decisión. Para cada uno de estos  $x^t$  se tiene que  $w = \sum_t \alpha^t r^t x^t$ . En la figura 2.2 se observa un problema de dos clases representadas como cruces y puntos. Los datos marcados con un círculo corresponden a los vectores de soporte y son los que configuran los márgenes del hiperplano. Los demás datos no afectan el hiperplano construido y si estos fueran eliminados de la muestra, no afectarían la solución (Alpaydin, 2014).

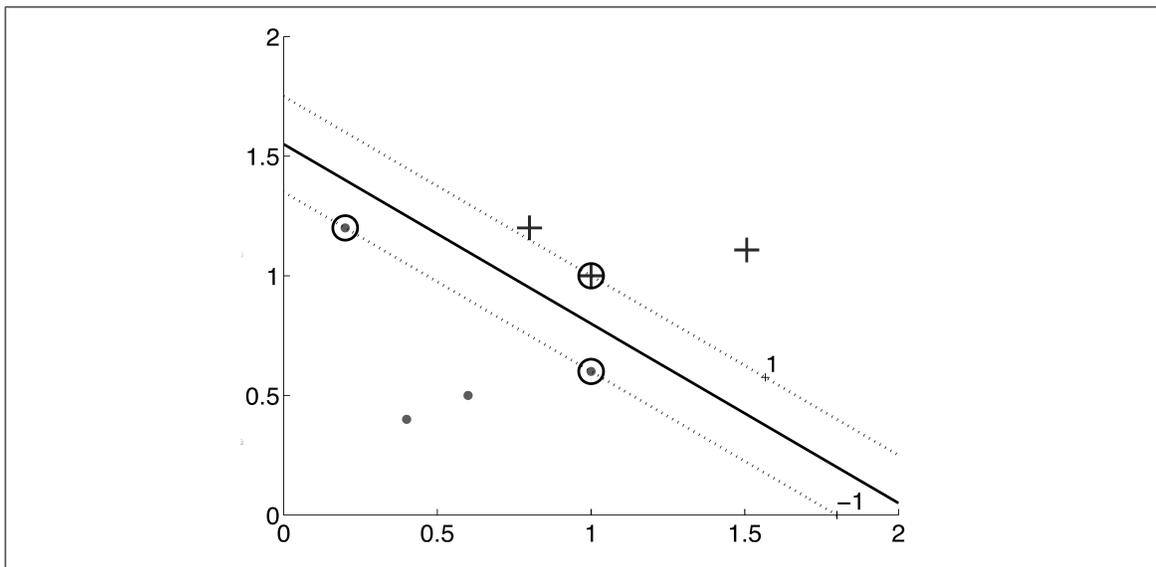


Figura 2.2. Vectores de Soporte  
Fuente: (Alpaydin, 2014)

Como se mencionó anteriormente, a partir de las máquinas de vectores de soporte se desarrollaron, en la década de los 90, los algoritmos de SVR (Vapnik, 1995). A diferencia de su modelo predecesor, los algoritmos de aprendizaje de SVR sirven para predecir valores para variables continuas como puede ser el tiempo de viaje. Al igual que en el método de clasificación se busca generar márgenes, pero incorporando una medida de error  $\epsilon$  permitido y una penalización  $C$  a aquellos datos que sobrepasen ese valor. El modelo queda planteado en la formulación (2.5) y su desarrollo mediante una formulación de Lagrange, permite la estimación de modelos SVR.

$$\begin{aligned}
\min \quad & \frac{1}{2} \|w\|^2 + C \sum_t (\xi_+^t + \xi_-^t) \\
\text{sa:} \quad & r^t - (w^t x^t + w_0) \leq +\epsilon + \xi_+^t, \forall t \\
& (w^t x^t + w_0) - r^t \leq +\epsilon + \xi_-^t, \forall t \\
& \xi_+^t, \xi_-^t \geq 0
\end{aligned} \tag{2.5}$$

Para más detalles sobre modelos de máquinas de vectores de soporte se recomienda revisar (Alpaydin, 2014), (Nilsson, 1996) y (Shalev-Shwartz y Ben-David, 2014)

### 2.1.3. Redes Neuronales (Neuronal Network, NN)

Las redes neuronales son un tipo de algoritmo de aprendizaje que puede ser utilizado para categorizar o estimar valores de variables continuas. Nacen a partir de un intento de modelar matemáticamente las neuronas biológicas en (McCulloch y Pitts, 1943) y su continuo desarrollo ha permitido que sean consideradas como aproximadores universales. Una red neuronal con la suficiente cantidad de datos de entrenamiento es capaz de aproximar cualquier tipo de función, con cualquier grado de exactitud (Hornik, Stinchcombe, y White, 1989).

Las redes neuronales están estructuras en capas interconectadas en un solo sentido que pueden ser capas de input, output o capas ocultas como se puede observar en la figura 2.3 representadas como  $V_0$  y  $V_2$  para las capas de input y output y como  $V_1$  para la capa oculta. Las capas ocultas son aquellas que procesan los datos y los relacionan. Las capas están compuestas de nodos, unidad básica de una red neuronal. Cada nodo aplica una función  $f$  a la información recibida, la que es ponderada mediante pesos  $w_{ij}$ , donde  $i$  representa el nodo que aplica la función y  $j$  la información de un nodo predecesor. Además, se considera un parámetro de penalización  $\alpha$  que realiza una compensación entre el aprendizaje y los pesos  $w_{ij}$  (Kriesel, 2007). El output de cada nodo puede ser utilizado como input por otros nodos o como output final de una red.

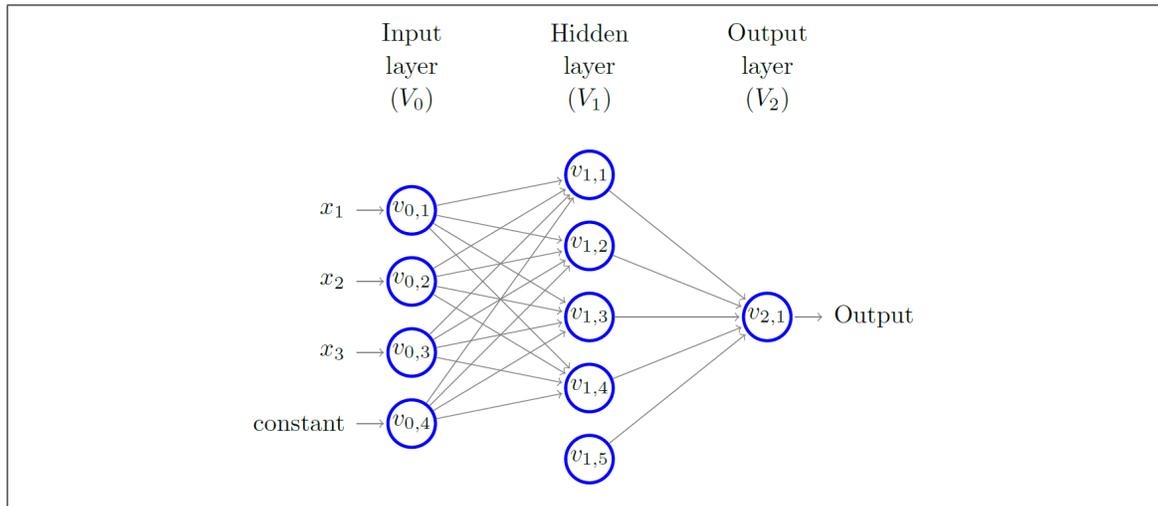


Figura 2.3. Capas Red Neuronal  
Fuente: (Shalev-Shwartz y Ben-David, 2014)

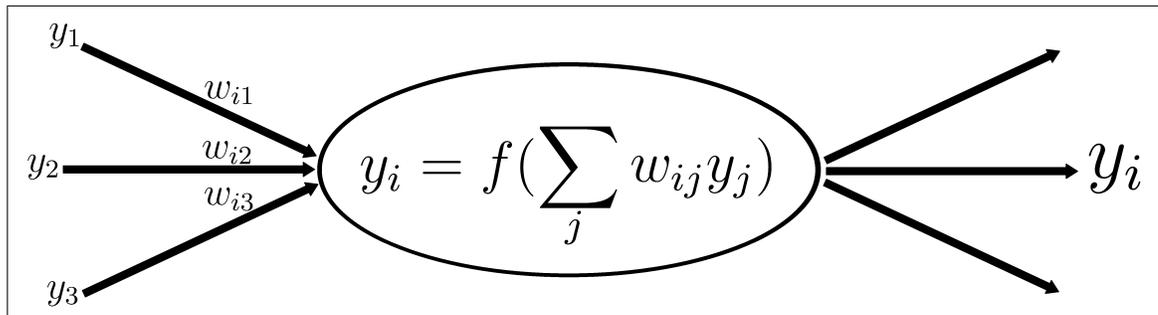


Figura 2.4. Nodo i Red Neuronal

La figura 2.4 es una representación de un nodo de una red neuronal, donde  $y_{123}$  representan datos de input, los que pueden ser output de nodos predecesores o variables  $x_i$  del vector de input de un modelo dependiendo en que capa de la red se encuentre el nodo. Cada uno de estos datos de input está ponderado por los pesos  $w_{ij}$ , que varían por el aprendizaje de la red.  $y_i$  es el output del nodo que será utilizado por otros nodos o como resultado final y es el resultado de aplicar la función  $f$  a la ponderación de los datos de input.

Para más detalles sobre modelos de redes neuronales se recomienda revisar (Alpaydin, 2014), (Nilsson, 1996) y (Shalev-Shwartz y Ben-David, 2014)

## **2.2. Trabajos Anteriores**

El número de investigaciones relacionadas con la predicción de velocidades o tiempos de viaje, así como el interés en este tema, han aumentado en los últimos años, gracias al mayor acceso a información. La posibilidad de contar con grandes bases de datos de posicionamiento (GPS) y, además, que estos se generen en tiempo real, permite que los modelos de aprendizaje de máquinas puedan generar estimaciones con mayor precisión y exactitud. Trabajos previos han desarrollado el tema desde distintos enfoques, utilizando variados modelos de aprendizaje o predicción y bases de datos completas o con información parcial. En la tabla 2.1 se presenta un resumen de trabajos consultados sobre predicción de velocidades o tiempos de viaje para flotas de vehículos.

Ramezani y Geroliminis (2012) utilizan los reportes GPS para calcular el tiempo de viaje por arcos, los cuales son utilizados para estimar velocidades en ruta. Cadenas de Markov son utilizadas para correlacionar arcos consecutivos, y así obtener la probabilidad de una distribución de tiempos de viaje en la ruta. Los autores concluyen que la metodología es más exacta que realizar una convolución de distribución de cada arco. Además, el método tiene buen comportamiento con baja cantidad de vehículos reportando su posición.

Tabla 2.1. Resumen Trabajos Anteriores

Trabajo	Fuente Datos	Input	Modelos	Output	Estimación
Ramezani y Geroliminis (2012)	GPS	Tiempo de viaje arcos	Cadena Markov	Distribución tiempo de viaje	Ruta
Jenelius y Koutsopoulos(2013)	GPS *	Tiempo de viaje	Maxima Verosimilitud	Distribución tiempo de viaje	Arcos
Zhan et al.(2013)	GPS	Origen, destino y tiempo de viaje	MNL y RL	Tiempo de viaje promedio hora	Arcos
Zheng y Van Zuylen (2013)	VISSIM *	Tiempo de viaje	ANN	Tiempo de viaje	Arcos
Gurmu y Fan (2014)	GPS	Tiempo de viaje	NN	Tiempo de llegada	
Wang et al. (2014)	GPS	Tiempo de viaje arcos	Optimización	Tiempo de viaje	Ruta
Fusco et al. (2015)	GPS	Tiempo de viaje	NN y BN	Tiempo de viaje promedio 5 min	Arcos
Julio (2015)	GPS	Velocidad	SVM,NN,BN	Velocidad	Arco
Rahmani et al. (2015)	GPS*	Tiempo de viaje y ruta	No-Paramétrico	Distribución tempo de viaje	Ruta

\*Baja frecuencia de reporte

Jenelius y Koutsopoulos (2013) realiza una estimación para tiempo de viaje en arcos mediante el tiempo entre pulsos GPS. También se incluyen variables de características de la vía, como velocidad máxima o largo del arco y condiciones del viaje, como día de la semana o condiciones climáticas. Los datos utilizados son de baja frecuencia de reporte por lo que se usa un promedio medio espacial para localizar trayectos GPS compartidos entre dos arcos. El modelo planteado permite determinar efectos importantes de ciertas características de los arcos y condiciones climáticas y así, usar los datos GPS como herramienta para monitorear el sistema de transporte urbano.

Zhan et al. (2013) considera la estimación del tiempo de viaje promedio durante una hora por arco mediante regresión lineal. Para ello, utiliza información parcial de GPS, proveniente de una flota de taxis, donde sólo se considera el origen, destino y el tiempo entre los dos puntos. Para estimar y seleccionar arcos se utiliza un modelo Logit multinomial (MNL)(McFadden, 1973) y el tiempo de viaje se estima mediante modelos de regresión lineal. La metodología propuesta permite realizar una estimación por espacios de una hora eficientemente.

Zheng y Van Zuylen (2013) realiza una estimación de tiempo de viaje en arcos delimitados por intersecciones. Se utiliza información GPS generada por simulación mediante VISSIM (Fellendorf, 1994). Un modelo de redes neuronales de tres capas es entrenado utilizando una baja frecuencia de reportes de posicionamiento. Entre los principales resultados se destaca que los modelos de NN superan a modelos analíticos y que la variable de velocidad instantánea no es un buen indicador en una red urbana.

Gurmu y Fan (2014) analiza el tiempo de viaje hasta el próximo paradero de una línea de buses de transporte público. Como input se utiliza el tiempo de viaje entre pulsos GPS y se estima un modelo mediante una red neuronal, la que resulta ser más exacta en las predicciones y más robusta que modelos realizados con promedios históricos.

Wang et al. (2014) utiliza el tiempo de viaje por arco, los cuales se calculan a partir de rutas obtenidas mediante reportes GPS. El tiempo de viaje en cada arco se calcula utilizando las rutas que en algún momento circulan por él. El cálculo de la velocidad en una ruta se realiza mediante una optimización que concatena cada uno de los tiempos de viaje en cada arco, considerando tiempos de viaje de datos históricos y en tiempo real. Se obtiene un modelo que tiene buena exactitud cuando hay datos faltantes, y arcos sin información.

Fusco et al. (2015) utiliza reportes GPS cada 2 minutos por vehículo, los cuales son agrupados en franjas de 5 minutos. Modelos de redes neuronales y de redes bayesianas (Pearl, 2014) son estimados. Velocidades promedio, el número de observaciones y la desviación estándar son utilizados como input. No se puede concluir cuál de los dos modelos (redes neuronales o redes bayesiana) es superior al estimar tiempos de viaje en el corto plazo.

Julio (2015) realiza una estimación de velocidad entre paraderos de líneas de transporte público. Las velocidades son calculadas en base a los tiempos de viaje reportados en el GPS de cada bus. Se estiman modelos de máquinas de vectores de soporte, redes neuronales y redes bayesianas, obteniéndose como resultado, que el modelo que presenta el mejor desempeño en estas condiciones es la red neuronal.

Rahmani et al. (2015) calcula distribuciones de tiempos de viaje para una ruta, utilizando datos GPS provenientes de una flota de taxis con una baja frecuencia de reporte de datos (inferior a dos pulsos GPS por minuto). Para la estimación de tiempos de viaje se consideran rutas que utilizan calles adyacentes a las contenidas en la ruta en cuestión. El modelo utilizado aplica pesos dependiendo de la importancia de cada dato, respecto de la ruta, en base a los cuales se calculan los valores estadísticos del modelo. Se demuestra que el modelo propuesto utilizando información proveniente de calles adyacentes permite una estimación con mayor exactitud.

En la mayoría de los trabajos consultados se observa que el uso de información GPS en tiempo real está asociada a la estimación del tiempo de viaje por arcos pequeños. En algunos de estos casos, se consideran ciertas condiciones como no conocer la ruta utilizada (Zhan et al., 2013) o bajas frecuencias de reporte (Jenelius y Koutsopoulos, 2013; Zheng y Van Zuylen, 2013; Rahmani et al., 2015), pero en todos ellos se utiliza una amplia flota de vehículos, que permite tener reportes contantes en cada arco. Por otro lado, cuando se utiliza la distribución de tiempos de viaje en cada arco para generar el tiempo de viaje en una ruta mayor, no es sencillo derivar de las distribuciones del tiempo de viaje de los enlaces que constituyen la ruta mayor (Rahmani et al., 2015). Además, en ciertas aplicaciones es preferible contar con el tiempo de viaje media de una ruta completa. En el caso de este trabajo, al utilizar una baja cantidad de vehículos reportando su posición, la estimación por arcos se hace inviable y, por tanto, la estimación se va a enfocar en rutas. Con el objetivo de compensar la baja cantidad de vehículos, se utiliza como información adicional los reportes GPS en vías paralelas. Finalmente, se utilizan tres modelos de estimación (regresión lineal múltiple, regresión de vectores de soporte y redes neuronales) para determinar cuál de ellos tiene un mejor comportamiento con la información adicional seleccionada.

## **CAPÍTULO 3. DATOS**

En esta sección se presentan los datos utilizados para el estudio y el procesamiento que se realizó, con tal de que éstos puedan ser usados en la modelación. A continuación, se describe la generación de los datos de input de los modelos, a partir de los pulsos GPS obtenidos por los vehículos de la base de datos.

### **3.1. Origen Datos**

Los datos utilizados provienen del transporte público de Santiago, Transantiago. El sistema cuenta con 6 646 buses (DTPM, 2017) que reportan su posición GPS cada 30 segundos y en conjunto generan 12.5 millones de pulsos diariamente. Estos datos se encuentran agrupados en archivos diarios, que están ordenados por la patente del vehículo y el horario de cada pulso. Cada uno de los pulsos contiene siete variables:

- Patente: Identificador del bus que envía el reporte
- Latitud
- Longitud
- Tiempogps: Fecha y hora del reporte
- Ruta: Trayecto que debe cubrir el bus
- Velocidad: Velocidad instantánea al momento de enviar reporte
- Ignición: Corresponde al estado del motor al momento del reporte.

Además, a cada uno de los pulsos se les asignó un ID correlativo y único para facilitar el seguimiento de cada bus. Respecto a la información contenida en cada pulso, la variable velocidad fue descartada en el análisis, debido a que es una velocidad instantánea, por lo que tiene una alta variabilidad y es poco representativa de la velocidad promedio en un arco. Respecto a las variables ignición, ésta puede tener dos estados, Si o No, donde No,

representa que el motor está apagado, por lo que el vehículo no se encuentra en movimiento. Los datos que se utilizarán corresponden al periodo comprendido entre el 3 de abril y el 25 de agosto de 2017, lo que corresponde a 21 semanas. De cada semana se utilizarán los días hábiles, descartando los días sábado y domingo por presentar un comportamiento distinto al habitual. En la figura 3.1 se pueden observar patrones de velocidad a distintas horas de una semana representados por escalas de colores, donde los verdes representan velocidades más altas y los colores rojos velocidades más bajas. Es posible apreciar un comportamiento distinto asociado a los días descartados para el análisis.

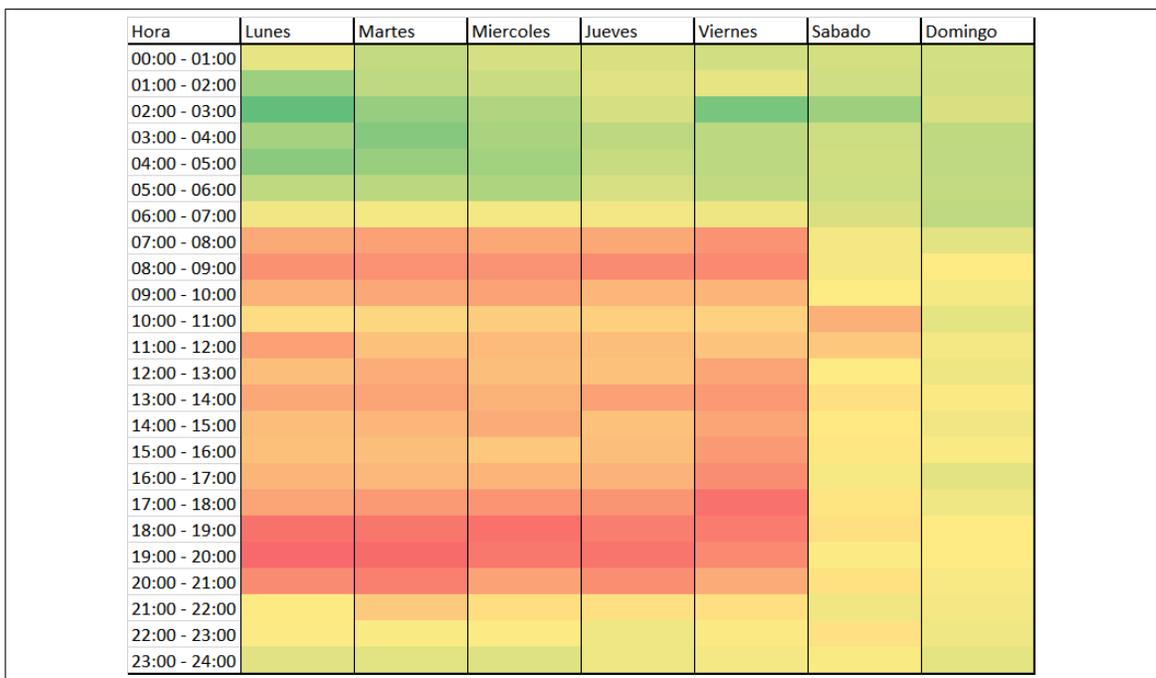


Figura 3.1. Velocidad Semana 3 de Abril

### 3.2. Procesamiento

Los pulsos GPS permiten localizar un bus durante su trayecto, pero por si solos no permiten calcular la velocidad promedio de desplazamiento de éste a lo largo de un arco y, por tanto, no pueden ser utilizados directamente. Por esto, es necesario generar el recorrido

de cada vehículo uniendo sus reportes GPS. Los pulsos GPS utilizados para este proceso corresponden sólo a los pulsos que se encuentran en el área que se estudiará.

### **3.2.1. Construcción de Trayectos**

Los trayectos corresponden al desplazamiento de un vehículo al interior de la zona de estudio en un sentido de movimiento. Este sentido de movimiento puede ser Norte, Sur, Este u Oeste, o puede variar según la orientación que tengan las calles en la zona de estudio. Los trayectos cuentan con información que caracterizan al vehículo en su recorrido por el área. Cada uno se compone de la distancia recorrida, hora de inicio o ingreso al área y el tiempo que permaneció al interior de ella. Otra característica de los trayectos es que un mismo vehículo puede generar más de un trayecto al circular por un área en un mismo horario. Esto se debe a que cada trayecto como unidad solo puede tener un sentido de movimiento y, por lo tanto, si un vehículo cambia su sentido de movimiento al interior de una zona, generará un trayecto adicional. En el caso del recorrido D09 en la figura 3.2, se construye un trayecto con sentido Sur y otro con sentido Este, ya que existe un cambio de dirección en sentido de movimiento al interior de la zona de estudio.

Respecto a la construcción de los trayectos, éstos se obtienen del seguimiento correlativo de pulsos GPS de un mismo vehículo en un sentido de movimiento. Esto se realizó identificando cada pulso GPS mediante su ID y relacionándolos con pulsos correlativos que pertenecen a la misma patente y ruta. Cuando se genera un nuevo trayecto para un vehículo, los pulsos son incorporados uno a uno revisando que el sentido de movimiento se mantenga. Por cada pulso nuevo, se suma el tiempo hasta el nuevo reporte y la distancia entre ellos. Esta distancia se calcula de forma lineal entre cada punto. En el caso de que el sentido de movimiento de un pulso, que cumpla con ruta, patente e ID correlativo, sea distinto, se analiza el sentido del siguiente pulso. Si el sentido es distinto

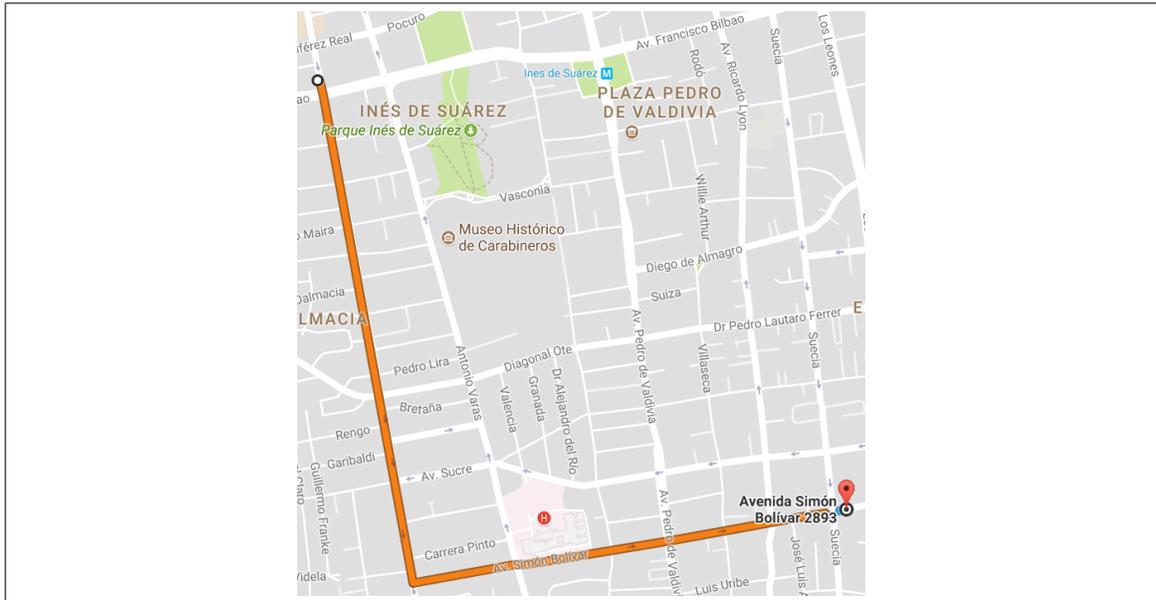


Figura 3.2. Trayecto D09

en dos pulsos consecutivos, se crea un nuevo trayecto, con el respectivo cambio de sentido de movimiento. En la figura 3.2 se observa un bus del recorrido D09 que genera dos trayectos. En rojo en sentido sur y en blanco con sentido este. En el nodo número tres se observa el cambio de sentido de movimiento y se genera un nuevo trayecto. Cuando se revisa la existencia de cambios de sentido se analiza que la distancia entre pulsos sea mayor a 3 metros para descartar que el vehículo se encuentre cambiando de pista mientras esta en un cruce o por variaciones marginales de la posición reportada por el GPS al estar detenido.

### 3.2.2. Revisión de datos

Mientras se realiza el seguimiento de los pulsos GPS en cada trayecto, se revisa que no existan incongruencias en su geolocalización. Esto se analiza mediante la distancia total de un recorrido y el tiempo de éste. La revisión se realiza para dos extensiones, 500 metros y 2.5 km. Esto se debe a que se analizará el efecto de calles paralelas con distintos largos de ruta. En el caso de 500 metros, se consideró que un trayecto debía

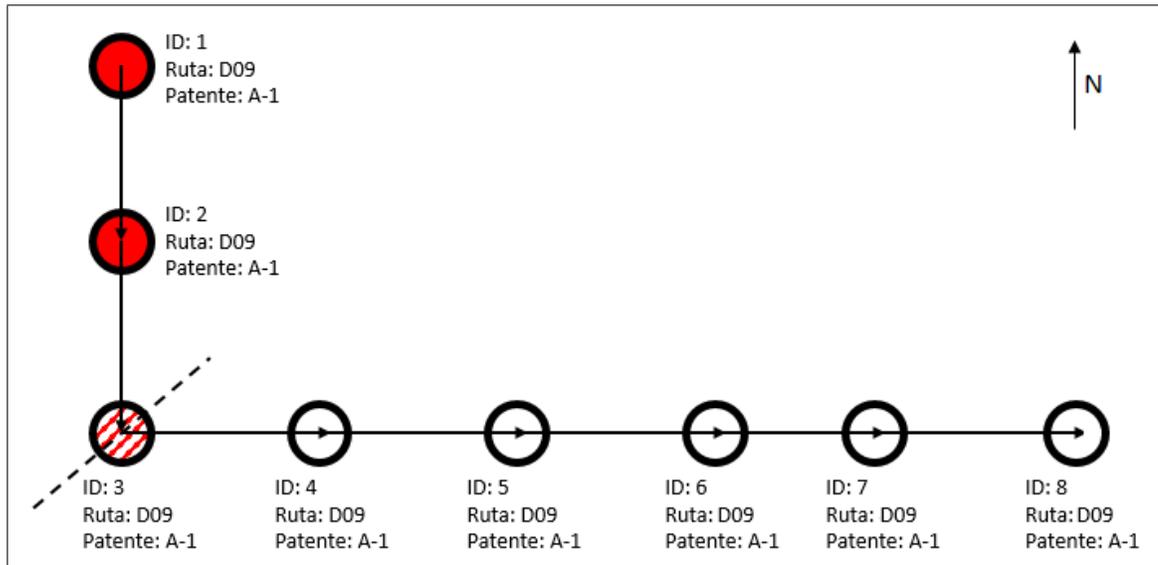


Figura 3.3. Generación de trayectos recorrido D09 por cambio de sentido

tener un mínimo de 250 metros y un máximo de 650 metros. Estos valores vienen dados por la posición de los pulsos GPS inicial y final. Si ambos se encuentran reportados en los bordes del área de estudio, se alcanzará una distancia cercana al valor máximo. Por el contrario, si los reportes se encuentran en la zona central, sólo se aceptarán aquellos recorridos con una distancia correspondiente a la mitad del arco estudiado. Respecto al tiempo empleado en recorrer el arco, los valores mínimos observados corresponden a 30 segundos, considerado un valor factible para la zona de estudio. En el caso del tiempo máximo, se consideró un valor de 1 200 segundos. Esto corresponde a 20 minutos en un arco de 500 metros. Este valor se seleccionó debido a la existencia de horarios donde la totalidad de los buses presentaban tiempos alrededor de los 1 100 segundos. Pese a que estos datos son extraños (considerando el comportamiento normal del arco de estudio) no fueron eliminados, ya que corresponde a información importante en los casos que se realiza estimación de velocidad con los datos de un periodo anterior. Para la revisión de 2.5 km, se consideró que largo de estos trayectos sea superior a 2 km e inferiores 3 km. Estos valores se seleccionaron debido a la posición del primer y último reporte GPS como se explicó para el caso de 500 metros. Respecto al tiempo de viaje, se consideró un tiempo

mínimo de 180 segundos para recorrer la distancia de 2.5 km y un tiempo máximo de 2 200 segundos. Este valor corresponde a 35 minutos en un arco de 2.5 km. Como se indicó en el caso de 500 metros, existen horarios donde la totalidad de buses presentan altos tiempos de viaje. Pese a ser datos extraños, no fueron eliminados, ya que corresponde a información importante cuando se utilizan datos del periodo anterior para realizar la estimación. Sólo se consideraron pulsos GPS que pertenezcan a alguna ruta que opere en el área. De esta forma, se descarta incluir buses que no se encuentren en operación y se estén movilizand para realizar alguna inyección de frecuencia en un punto conflictivo de la red o retornando a la matriz sin pasajeros.

### 3.2.3. Generación de horarios

Una vez generados todos los trayectos, deben ser asignados a un horario determinado. En el caso de que un vehículo circule en dos horarios durante el mismo trayecto, este trayecto es asignado al horario en el que el vehículo estuvo más tiempo. En el caso de la figura 3.4, el vehículo 1 pertenece al periodo  $P1$  y el vehículo 2, pese a que su hora de ingreso al área de estudio es en el periodo  $P1$ . Para efectos de este estudio, se considera que el vehículo pertenece al periodo  $P2$  por circular una mayor cantidad de tiempo durante éste.

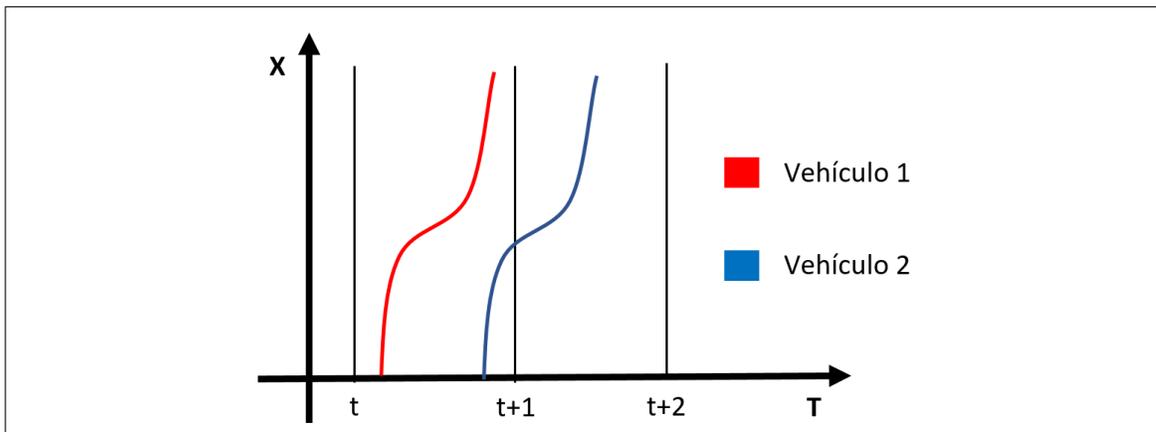


Figura 3.4. Selección de Horario

### **3.3. Consolidación de los datos**

Una vez finalizada la etapa de procesamiento de los pulsos GPS, los trayectos generados se agrupan según la vía por donde circularon. Por lo tanto, por cada calle paralela al interior del área de estudio se creará una base de datos que contenga los trayectos de este eje. Cada trayecto contenido en estas bases de datos contendrá toda la información del primer pulso GPS utilizado para su construcción y las siguientes variables que caracterizan el trayecto:

- Fecha: fecha y hora de ingreso al área de estudio en formato timestap
- Distancia: Distancia total recorrida entre el primer y último pulso GPS reportado por el vehículo al interior del área de estudio
- Sentido: Dirección de movimiento (Este, Oeste, Norte o Sur)
- Tiempo: Tiempo transcurrido entre el primer y último pulso reportado por el bus al interior del área de estudio
- Pulsos: Cantidad de pulsos que conforman el trayecto

## CAPÍTULO 4. METODOLOGÍA

La metodología planteada tiene como objetivo trabajar con promedios de datos que permitan obtener una buena estimación de tiempos de viaje, cuando la cantidad de información es limitada. Los promedios son los elementos básicos que se utilizarán para entrenar los modelos y contienen información sobre las condiciones de tráfico que enfrentaron los vehículos al circular en cierto horario y eje. Lo anterior implica que los vehículos serán agrupados de forma temporal a lo largo de una franja horaria y espacial dependiendo de la extensión de calle que sea seleccionada. Los promedios estarán contruidos por trayectos de vehículos que circulan por un eje. Con los trayectos agrupados, se obtendrán variables representativas de las condiciones de la zona en un determinado horario. De esta forma, se busca estimar el tiempo de viaje promedio a lo largo de un eje durante una franja horaria.

### 4.1. Variables modelo

Para los modelos planteados se utilizarán las mismas variables, las cuales provienen de datos históricos y datos en tiempo real. Además, se incorporarán datos de las calles paralelas para buscar mejoras en la estimación. Estas variables corresponden a:

$V_{hi}^{t+1}$  : Velocidad promedio histórica en el periodo  $t + 1$  de la calle de interés

$\sigma_{hi}^{t+1}$  : Desviación estándar histórica en el periodo  $t + 1$  de la calle de interés

$V_{ai}^t$  : Velocidad promedio actual en el periodo  $t$  de la calle de interés

$\sigma_{ai}^t$  : Desviación estándar actual en el periodo  $t$  de la calle de interés

$V_{hlk}^{t+1}$  : Velocidad promedio histórica en el periodo  $t + 1$  de la calle lateral  $k$

$\sigma_{hlk}^{t+1}$  : Desviación estándar histórica en el periodo  $t + 1$  de la calle lateral  $k$

$V_{alk}^t$  : Velocidad promedio actual en el periodo  $t$  de la calle lateral  $k$

$\sigma_{alk}^t$  : Desviación estándar actual en el periodo  $t$  de la calle lateral  $k$

$D_t$  : Dummy por cada horario del día. 1 si horario pertenece al periodo  $t$ . 0 e.o.c.

En base a estas variables, se estimará el tiempo de viaje en el periodo  $t+1$ , según los diferentes modelos. En el caso de las variables de datos históricos, estos serán calculados con información correspondiente a las 3 semanas anteriores a la ocurrencia del viaje.

Cada una de las variables utilizadas fue normalizada a media 0 y varianza 1 de acuerdo a la ecuación 4.1. Los valores de  $\mu$  y  $\sigma$  fueron calculados con los datos de entrenamiento y posteriormente fueron aplicados a los datos de testeo para mantener el mismo orden de transformación.

$$x_i^* = \frac{x_i - \mu}{\sigma} \quad (4.1)$$

#### 4.1.1. Diagrama Variables Vector Input

En la figura 4.1 se presenta un diagrama de la composición de datos que se utilizarán para la predicción. En verde se encuentra el tiempo de viaje promedio a estimar. Los datos en amarillo corresponden a información proveniente de la calle de interés y en azul aquellos proveniente de una calle lateral. En el caso de la información histórica, ésta se presenta como un triángulo con el color amarillo o azul según sea el caso de procedencia. Debido a que se busca determinar en qué situación los datos de calles paralelas tienen una mayor incidencia en la estimación de velocidad, se realizarán estimaciones utilizando el

vector de input completo o partes del él. Se considerarán cuatro situaciones: en la primera, se estimará utilizando sólo con información en tiempo real (rectángulos en color amarillo y azul); en la segunda situación se utilizará información en tiempo real, pero sólo de calles laterales (rectángulo azul). En la tercera situación se estimará solo con información histórica (triángulos amarillo y azul) y, finalmente, se estimarán modelos con toda la información disponible (vector input completo). Las cuatro situaciones planteadas representan las posibles situaciones que puede enfrentar un operador que requiera estimar tiempos de viaje y no disponga de una base de datos con la totalidad de información.

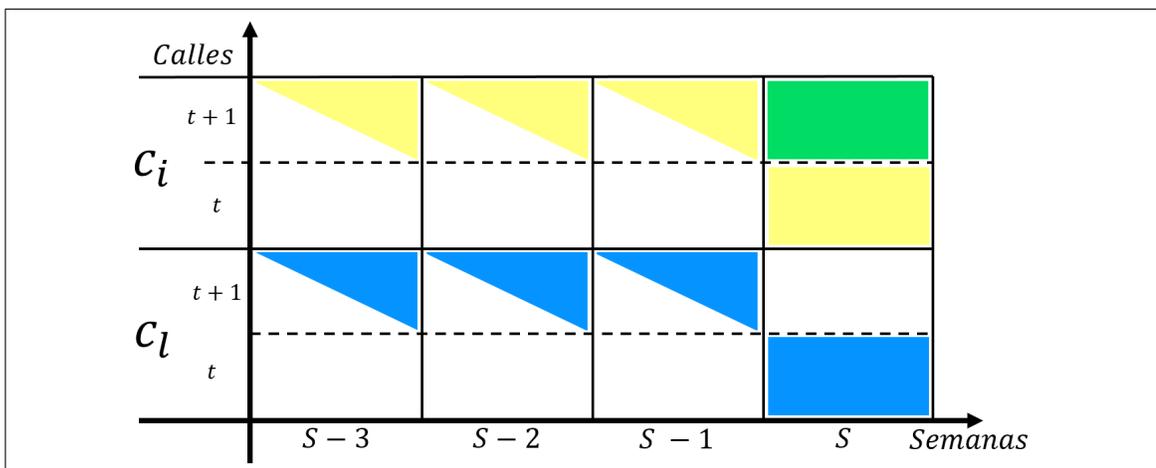


Figura 4.1. Diagrama Variables

## 4.2. Muestreo y Estructura

Como se mencionó anteriormente, se busca determinar si la adición de información de calles paralelas permite mejorar la calidad de la estimación de tiempo de viaje cuando se posee una baja cantidad de datos. Debido a que se está trabajando con una base de datos completa y que se posee una gran cantidad de vehículos en circulación por la zona de estudio, es necesario realizar un muestreo representativo, con el fin de simular una situación en que la cantidad de información es limitada. Para ello se realizará un muestreo aleatorio por calle y horario; esto permite mantener la proporción de vehículos, tanto desde

una perspectiva espacial como temporal. Por lo tanto, por cada calle se obtendrá una sub muestra que mantendrá la proporción de vehículos en cada periodo horario. Debido a que se trabajará con una muestra aleatoria de datos, se realizan diez réplicas de todo el proceso para evitar que singularidades de un muestreo aleatorio afecten los resultados finales de la modelación.

A continuación, se presenta un diagrama de la ejecución del código, donde *Situacion* corresponde a los datos que se utilizarán en el vector de input, *r* es el número de réplica y *Muestra* corresponde al momento en el cual se realiza el muestreo aleatorio de los datos:

---

**Algoritmo 1:** Secuencia ejecución

---

```
1 Secuencia de ejecución;
2 for Situacion  $\in$  {1, 4} do
3   | for r  $\in$  {1, 10} do
4   |   | Muestra;
5   |   | LR;
6   |   | SVR;
7   |   | NN;
8   |   | HA;
9   |   | LPA;
10  |   end
11 end
12 resultados;
```

---

En el diagrama de Algoritmo 1 se puede observar que para cada prueba se mantendrán los mismos datos. Esto implica que los 3 modelos y los 2 benchmarks de comparación (HA y LPA, ver sección 4.4), se usarán las mismas muestras de datos; por tanto, no existirán diferencias en los datos de inputs utilizados en el entrenamiento y calibración de cada uno de ellos.

### 4.3. Medidas de Desempeño

Se seleccionan dos medidas de desempeño que permiten comparar de forma válida los 5 tipos de estimaciones que se utilizarán para predecir la velocidad de los vehículos. El primero es el error cuadrático medio (MSE) (Ecuación 4.2) y el segundo es el error porcentual absoluto medio (MAPE) (Ecuación 4.3). Las medidas de desempeño se calcularán entre el valor estimado con la muestra los datos y el valor real obtenido con la totalidad de los datos.

$$MSE = \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n} \quad (4.2)$$

$$MAPE = \frac{\sum_{i=1}^n \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right|}{n} \quad (4.3)$$

En el caso de MSE, es un tipo de error que penaliza de mayor manera los errores de mayor magnitud. Es la medida de desempeño principal, ya que en la predicción de velocidades es preferible una mayor cantidad de errores pequeños a una baja cantidad de velocidades irreales, que distorsionen seriamente la planificación de rutas y la toma de decisiones. En cambio, MAPE, es un tipo de error que pondera de la misma forma distintas magnitudes de errores y al ser porcentual, permite observar directamente el desempeño de los modelos en relación con el valor que se quiere predecir.

Cada una de estas medidas de desempeño se calculó tanto para los datos de entrenamiento, como para los datos de validación y test. Además, en el caso de MSE, se calculó para la calibración de los hiper parámetros de los modelos SVR y NN con el objetivo de optimizar sus parámetros

#### 4.4. Calibración Modelos y Benchmark de comparación

En esta sección se detallará cómo se calibrarán y aplicarán los distintos modelos utilizados para estudiar el efecto de agregar información de calles paralelas y, también, los benchmarks de comparación que se utilizarán. Respecto a estos últimos, esto corresponde a la utilización de un promedio histórico y a un promedio en tiempo real.

##### 4.4.1. Promedio Histórico(Historical Average, HA)

Respecto a este punto de comparación, los valores utilizados para predecir se obtienen del promedio de velocidad de las últimas tres semanas en el mismo periodo y día de la semana (Figura 4.2). Es una estimación simple que no requiere de un sistema que permita contar con información en tiempo real de GPS de los vehículos en operación. Además, al no requerir información en tiempo real, permite calcular tiempos de viaje y rutas con anticipación.

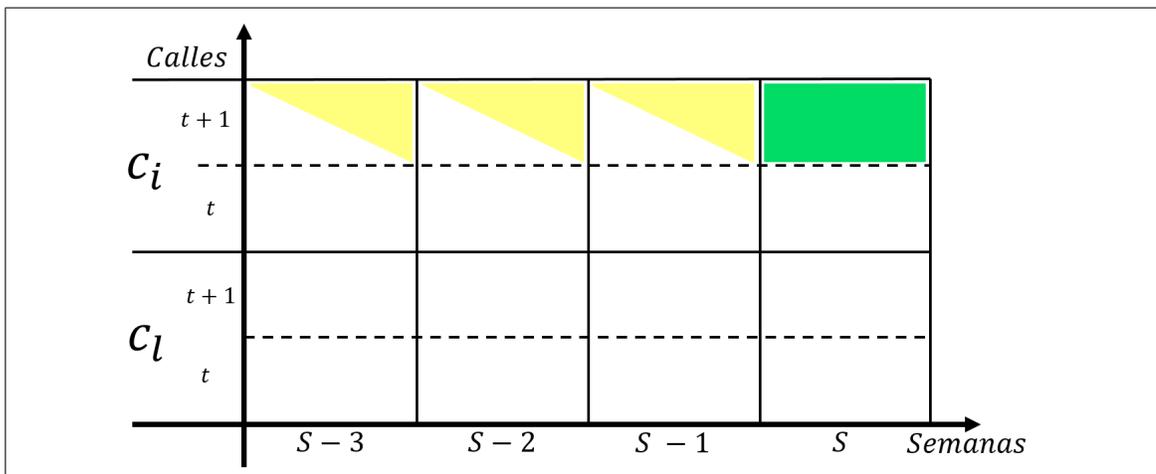


Figura 4.2. Diagrama HA

#### 4.4.2. Promedio Último Periodo (Last Period Average, LPA)

En este caso, los valores utilizados para predecir se obtienen del promedio de velocidad del periodo previo al observado (Figura 4.3); es decir, se utiliza la misma velocidad que se observa en el periodo actual. Es una estimación simple, pero requiere contar con información en tiempo real. Esto implica tener un sistema que recolecte constantemente la información de los vehículos en operación.

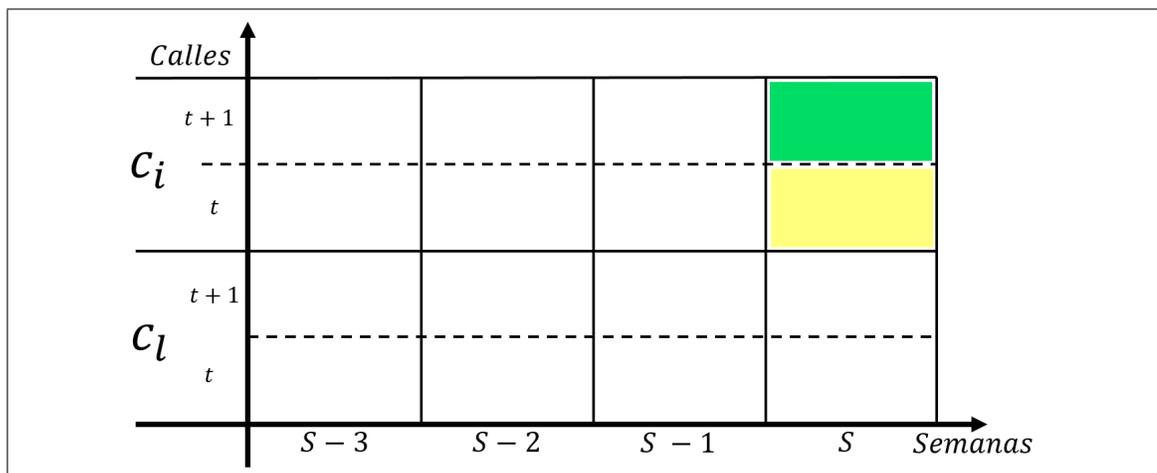


Figura 4.3. Diagrama LPA

#### 4.4.3. Regresión Lineal (LR)

Para el entrenamiento de los modelos de regresión lineal se utilizaron todas las variables que se disponían del eje de interés, así como de los ejes laterales, tal como se explicó en la sección anterior. Además, para mejorar la estimación de los parámetros  $\beta$  para cada variable, se utilizó validación cruzada con los datos de entrenamiento. Esto se realizó mediante Scikit-Learn 0.19.1 (Pedregosa et al., 2011)

#### **4.4.4. Regresión de Vectores de Soporte(SVR)**

Para el entrenamiento de los modelos de regresión de vectores de soporte se calibró en cada caso el parámetro de penalidad  $C$ . Esto implica, que para cada set de datos se seleccionó el mejor  $C$ . Para el parámetro  $C$  se probaron 5 posibles valores. Estos valores corresponden a 0.25, 0.5, 1, 2, 4. Esto se realizó mediante validación cruzada, utilizando cada semana como "unidad". En la siguiente sección se explicará en detalle la aplicación de la validación cruzada.

#### **4.4.5. Redes Neuronales(NN)**

Para los modelos de redes neuronales, se calibraron dos hiper parámetros: el número de neuronas y el parámetro de penalidad,  $\alpha$ . En el caso del número de neuronas, se realizó una calibración previa mediante validación cruzada y posteriormente se mantuvo fijo para los modelos finales. El valor obtenido corresponde a 8 nodos. Este valor fue el que más se repitió al minimizar el error cuadrático medio. El parámetro  $\alpha$  se calibró para cada modelo mediante validación cruzada. Los valores probados para el parámetro  $\alpha$  corresponden a 0.000001, 0.00001, 0.0001, 0.001, 0.01.

### **4.5. Validación Cruzada**

En esta sección se explicará la validación cruzada utilizada en la calibración de los hiper parámetros de los modelos SVR y NN.

Para este método, los datos fueron agrupados por semanas y éstas fueron utilizadas como unidades para la validación y testeo. Como se explicó anteriormente, cada uno de los vectores de input para los modelos cuenta con información histórica que está conformada por las tres semanas anteriores, al periodo para el cual se requiere estimar los tiempos de viaje. Por ello, al seleccionar una semana para validación, se deben descartar las tres semanas previas en el entrenamiento para así evitar utilizar los mismos datos de forma

indirecta. En la figura 4.4 se puede observar cómo se seleccionan las semanas en cada iteración considerando una semana de datos históricos.

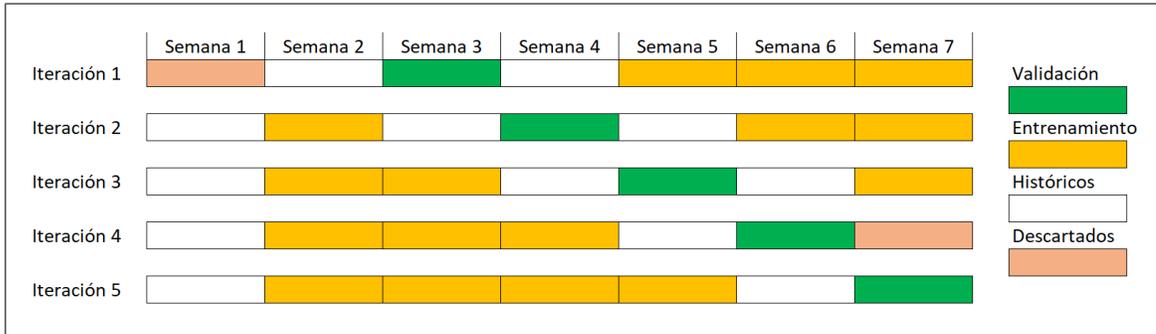


Figura 4.4. Diagrama Validación Cruzada

De esta forma, al seleccionar la semana de validación, se revisa que se tenga información de una semana previa para la información histórica y que esta semana de datos históricos no se utilice para el entrenamiento de los modelos. Además, existen ciertas semanas que deben ser descartadas, ya que no tienen una semana previa que permita generar sus datos históricos y que no interfiera con la semana de validación. Los modelos se entrenan para cada una de las iteraciones y se calcula el error cuadrático medio con la semana de validación. Este proceso se realiza para los 5 valores del hiper parámetro que se está calibrando y finalmente se selecciona aquel con el cual se obtiene el menor ECM.

## CAPÍTULO 5. CASO DE ESTUDIO

Para el estudio se consideró una zona en la cual existan calles paralelas en el mismo sentido que presenten una jerarquía similar. Con esto en consideración se seleccionó una zona comprendida en las comunas de Providencia y Ñuñoa. En esta zona existen 3 avenidas paralelas con sentido sur.

### 5.1. Zona

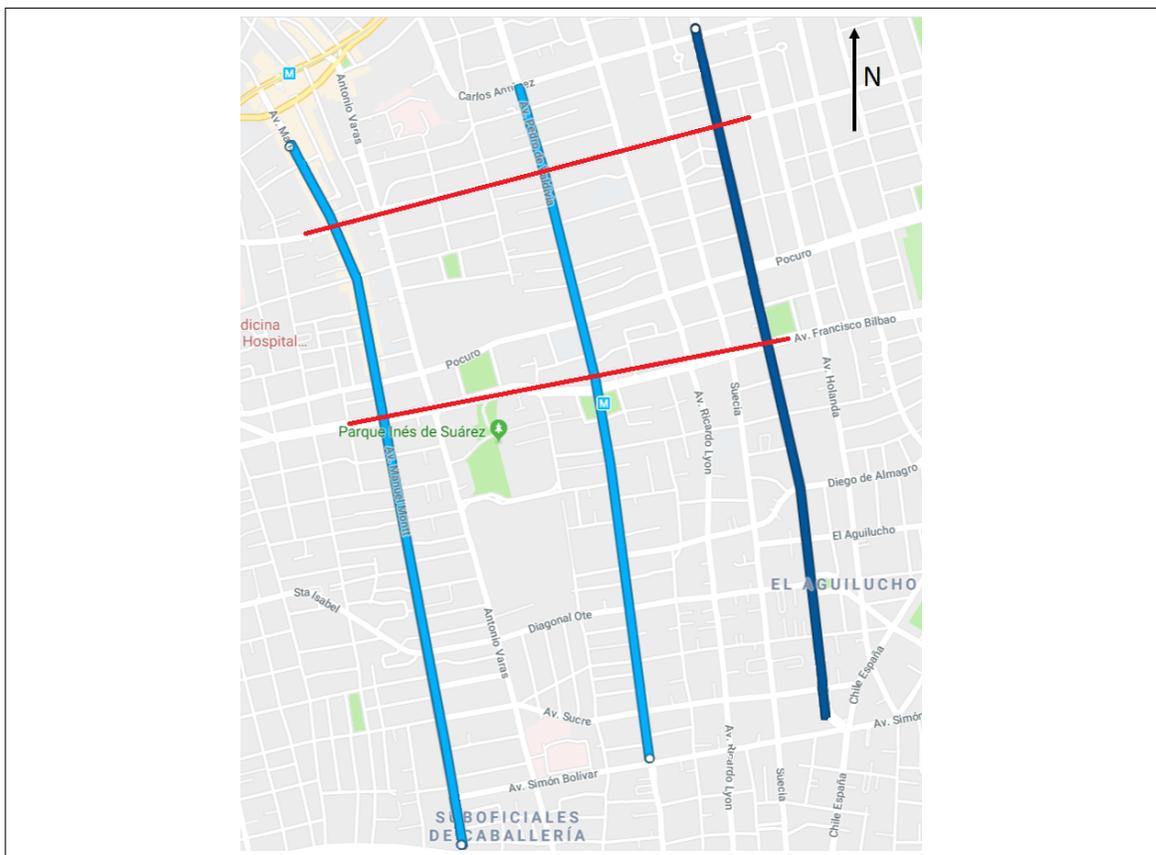


Figura 5.1. Zona Estudio

Para las tres avenidas en la zona de estudio, se consideró un cuadrante de 2.5 km entre Av. Nueva Providencia, por el norte, y Av. Simón Bolívar, por el sur. Además, se seleccionó un trayecto de 0.5 km entre Avenida Eliodoro Yañez y Pocuro, entre ambas bandas rojas indicadas en la figura 5.1. Los ejes considerados en esta zona corresponden, de derecha a izquierda en la figura 5.1, a Avda. Los Leones, Avda., Pedro de Valdivia y Avda. Manuel Montt. Pedro de Valdivia es la calle central y en la que se realizarán las predicciones. La separación entre Avda. Pedro de Valdivia con Avda. Los Leones corresponde a 650 metros, mientras que la distancia entre Avda. Manuel Montt y Avda. Pedro de Valdivia es de 750 metros. A continuación, se detallan las características de cada una de estas avenidas.

#### **5.1.1. Pedro de Valdivia(PdV)**

Avenida Pedro de Valdivia es el eje central de las tres avenidas en el área de estudio y es la calle en la que se estimará la velocidad. Es una avenida bidireccional de dos pistas por sentido. A diferencia de otras avenidas, las pistas tienen cubiertas de adoquines. Cuenta con 5 recorridos de transporte público (103, 114, 117, 212 ,515) que en conjunto operan durante las 24 horas del día. A continuación, en la tabla 5.1 se presenta la frecuencia por hora para el conjunto de recorridos de transporte público en el eje, el tiempo de viaje promedio y su desviación estándar, tanto para el tramo de 0.5 km, como para el tramo de 2.5 km. Como es de esperar la desviación estándar para el tramo de 0.5 km es mayor que para el tramo de 2.5 km.

Como se explicó en la sección de procesamiento de datos, los pulsos GPS se agruparon para construir trayectos que serán utilizados para los modelos. En el caso de Avenida Pedro de Valdivia, se cuenta con 53.400 trayectos para las 21 semanas en sentido Sur.

Tabla 5.1. Resumen Pedro de Valdivia

Hora	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00
Frecuencia (buses/hr)	8.93	3.65	2.1	1.94	1.76	9.81	25.76	24.94
Tiempo 0.5 (min/km)	2.13	1.79	1.81	1.84	1.99	2.16	2.72	4.10
Desviación 0.5 (min/km)	0.51	0.25	0.20	0.23	0.27	0.56	0.83	1.09
Tiempo 2.5 (min/km)	2.24	2.05	1.93	1.87	1.91	2.1	2.61	3.92
Desviación 2.5 (min/km)	0.38	0.28	0.2	0.16	0.14	0.35	0.48	0.94
Hora	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00
Frecuencia (buses/hr)	27.05	33.35	33.8	25.83	24.62	27.74	28.59	28.83
Tiempo 0.5 (min/km)	4.34	3.95	3.57	3.62	4.74	5.59	4.82	4.03
Desviación 0.5 (min/km)	1.05	1.11	0.84	0.91	1.34	1.67	1.37	1.20
Tiempo 2.5 (min/km)	4.56	4.01	3.4	3.65	4.05	4.27	3.94	3.87
Desviación 2.5 (min/km)	0.97	0.75	0.61	0.66	0.75	0.75	0.62	0.52
Hora	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Frecuencia (buses/hr)	27.46	26.17	23.13	28.7	33.93	24.48	20.66	15.34
Tiempo 0.5 (min/km)	4.51	6.11	8.19	8.04	6.04	3.25	2.87	2.37
Desviación 0.5 (min/km)	1.27	1.59	1.39	1.40	1.82	0.79	0.86	0.70
Tiempo 2.5 (min/km)	4.22	4.89	6.12	6.04	4.66	3.31	2.92	2.56
Desviación 2.5 (min/km)	0.57	0.65	0.74	0.78	0.93	0.53	0.53	0.43

### 5.1.2. Los Leones(LL)

Avenida Los Leones es una avenida bidireccional de dos pistas por sentido y se encuentra al Este de Pedro de Valdivia. A diferencia de la anterior, ésta cuenta con un ban-dejón central y las pistas están cubiertas de hormigón. Existe una pista sólo bus, pero en el sentido norte, por lo que no impacta en los datos analizados. Tiene un sólo recorrido de transporte público (104). A continuación, en la tabla 5.2, se presenta la frecuencia del recorrido, el tiempo de viaje promedio por el eje y su desviación estándar, tanto para el tramo de 0.5 km, como para el tramo de 2.5 km. Al igual que en el caso anterior, y en línea

con las expectativas, la desviación estándar para el tramo de 0.5 km es mayor que para el tramo de 2.5 km.

Tabla 5.2. Resumen Los Leones

Hora	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00
Frecuencia (buses/hr)	8.24	3.17	3.46	3.32	3.03	4.76	11.98	11.56
Tiempo 0.5 (min/km)	1.89	1.77	1.73	1.79	1.73	1.92	2.23	2.69
Desviación 0.5 (min/km)	0.62	0.28	0.29	0.22	0.22	0.45	0.86	1.26
Tiempo 2.5 (min/km)	2.29	2.1	2.03	1.98	1.97	2.1	2.6	3.69
Desviación 2.5 (min/km)	0.32	0.24	0.24	0.23	0.25	0.31	0.43	0.77
Hora	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00
Frecuencia (buses/hr)	12.19	18.4	17.54	12.24	14.47	14.93	15.11	15.84
Tiempo 0.5 (min/km)	2.42	3.40	2.84	3.00	3.60	4.02	3.84	3.89
Desviación 0.5 (min/km)	1.09	1.27	1.10	1.07	1.30	1.26	1.14	1.16
Tiempo 2.5 (min/km)	4.26	3.74	3.19	3.28	3.51	3.76	3.76	3.8
Desviación 2.5 (min/km)	0.83	0.56	0.41	0.36	0.4	0.43	0.42	0.41
Hora	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Frecuencia (buses/hr)	13.78	12.85	15.1	16.52	19.53	14.16	13.07	9.31
Tiempo 0.5 (min/km)	5.00	7.61	10.36	9.56	6.65	2.65	2.66	2.25
Desviación 0.5 (min/km)	1.97	2.38	2.16	2.11	2.46	1.08	0.94	0.82
Tiempo 2.5 (min/km)	4.1	4.95	6.35	6.55	4.82	3.25	2.84	2.66
Desviación 2.5 (min/km)	0.48	0.67	0.65	0.64	0.9	0.46	0.43	0.41

Respecto a la cantidad de trayectos, para Avenida Los Leones, se cuenta con 29.800 en sentido Sur.

### 5.1.3. Manuel Montt(MM)

Avenida Manuel Montt es una avenida unidireccional de 3 pistas en sentido Sur y se encuentra al oeste de Pedro de Valdivia. Al igual que en Avenida Pedro de Valdivia, las

pistas están cubiertas de adoquines. A diferencia de las dos avenidas anteriores, en la extensión de la zona de estudio, Avenida Manuel Montt presenta estacionamientos laterales en el costado oeste. 3 recorridos de transporte público (106,126, D09) recorren el eje dirección norte-sur. A continuación, en la tabla 5.3, se presenta la frecuencia para el conjunto de recorridos de transporte público en el eje, el tiempo de viaje promedio y su desviación estándar, tanto para el tramo de 0.5 km, como para el tramo de 2.5 km. Nuevamente, la desviación estándar para el tramo de 0.5 km es mayor que para el tramo de 2.5 km

Tabla 5.3. Resumen Manuel Montt

Hora	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00
Frecuencia (buses/hr)	7.43	1.86	2.01	2.24	2.04	3.72	11.95	16.86
Tiempo 0.5 (min/km)	2.18	2.14	2.12	2.06	2.19	2.22	2.48	3.74
Desviación 0.5 (min/km)	0.58	0.26	0.35	0.29	0.26	0.51	0.68	1.26
Tiempo 2.5 (min/km)	2.36	2.17	2.2	2.12	2.12	2.35	3.13	4.61
Desviación 2.5 (min/km)	0.47	0.14	0.22	0.22	0.21	0.37	0.52	0.86
Hora	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00
Frecuencia (buses/hr)	18.1	21.64	23.12	18.69	18.13	17.78	19.3	17.47
Tiempo 0.5 (min/km)	4.54	4.00	3.65	3.86	4.09	4.33	3.97	3.93
Desviación 0.5 (min/km)	1.37	0.97	0.80	0.91	0.95	1.08	0.95	0.94
Tiempo 2.5 (min/km)	4.94	4.42	3.97	4.12	4.17	4.31	4.21	4.3
Desviación 2.5 (min/km)	0.82	0.65	0.58	0.59	0.63	0.63	0.59	0.62
Hora	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Frecuencia (buses/hr)	17.19	15.54	15.88	19.97	18.34	13.32	12.27	10.7
Tiempo 0.5 (min/km)	4.29	4.49	6.53	7.49	4.32	3.37	3.01	2.52
Desviación 0.5 (min/km)	1.08	1.14	1.88	1.56	1.28	0.76	0.79	0.71
Tiempo 2.5 (min/km)	4.51	4.71	5.71	5.9	4.21	3.89	3.55	2.94
Desviación 2.5 (min/km)	0.63	0.65	0.82	0.81	0.74	0.65	0.59	0.59

Respecto a la cantidad de datos, se cuenta con 34.100 trayectos en el eje.

## 5.2. Comparación Calles

A continuación, se presenta la diferencia en los tiempos de viaje entre Avda. Pedro de Valdivia y las avenidas Los Leones y Manuel Montt, respectivamente. En la tabla 5.4 se presenta los tiempos de viaje para el tramo de 0.5km de Avda. Pedro de Valdivia para cada horario y la diferencia porcentual con las avenidas Los Leones y Manuel Montt. Es interesante notar que Avda. Manuel Montt tiene tiempos de viaje más parecidos a Avda. Pedro de Valdivia que Avda. Los Leones, salvo en algunos horarios nocturnos. Pese a presentar características viales similares, Avda. Los Leones presenta tiempos de viaje muy distintos en la mayoría de los horarios y hasta un 44 % de diferencia. Es de esperar que la información proveniente de Avda. Manuel Montt genere mayores mejoras en la estimación de los tiempos de viaje para Avda. Pedro de Valdivia que Avda. Los Leones. Sin embargo, para el tramo de 2.5km, en la tabla 5.5 se observa que existen mayores similitudes con Avda. Los leones, lo que es acorde con las características viales que presentan ambas calles.

Tabla 5.4. Comparación Tiempos de Viaje 0.5 km

Tiempo (min/km)	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00
Pedro de Valdivia	2,13	1,79	1,81	1,84	1,99	2,16	2,72	4,1
Los Leones	-11 %	-1 %	-4 %	-3 %	-13 %	-11 %	-18 %	-34 %
Manuel Montt	2 %	20 %	17 %	12 %	10 %	3 %	-9 %	-9 %
Tiempo (min/km)	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00
Pedro de Valdivia	4,34	3,95	3,57	3,62	4,74	5,59	4,82	4,03
Los Leones	-44 %	-14 %	-20 %	-17 %	-24 %	-28 %	-20 %	-3 %
Manuel Montt	5 %	1 %	2 %	7 %	-14 %	-23 %	-18 %	-2 %
Tiempo (min/km)	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Pedro de Valdivia	4,51	6,11	8,19	8,04	6,04	3,25	2,87	2,37
Los Leones	11 %	25 %	26 %	19 %	10 %	-18 %	-7 %	-5 %
Manuel Montt	-5 %	-27 %	-20 %	-7 %	-28 %	4 %	5 %	6 %

Tabla 5.5. Comparación Tiempos de Viaje 2.5 km

Tiempo (min/km)	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00
Pedro de Valdivia	2,24	2,05	1,93	1,87	1,91	2,1	2,61	3,92
Los Leones	2 %	2 %	5 %	6 %	3 %	0 %	0 %	-6 %
Manuel Montt	5 %	6 %	14 %	13 %	11 %	12 %	20 %	18 %
Tiempo (min/km)	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00
Pedro de Valdivia	4,56	4,01	3,4	3,65	4,05	4,27	3,94	3,87
Los Leones	-7 %	-7 %	-6 %	-10 %	-13 %	-12 %	-5 %	-2 %
Manuel Montt	8 %	10 %	17 %	13 %	3 %	1 %	7 %	11 %
Tiempo (min/km)	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Pedro de Valdivia	4,22	4,89	6,12	6,04	4,66	3,31	2,92	2,56
Los Leones	-3 %	1 %	4 %	8 %	3 %	-2 %	-3 %	4 %
Manuel Montt	7 %	-4 %	-7 %	-2 %	-10 %	18 %	22 %	15 %

## CAPÍTULO 6. RESULTADOS

Los resultados están divididos en subsecciones: primero se presentan los resultados de los modelos de comparación (HA y LPA), y luego los resultados asociados a los tres modelos propuestos (LR, SVR, NN). Para cada caso se presentará MSE y MAPE, dando prioridad a este último indicador por su interpretabilidad. Además, para cada uno de los modelos se presentan los resultados de cada configuración analizada (Histórico, Información Actual, Información Actual sin Calle Interés, Histórico más Información Actual). Finalmente se comparan las mejores opciones. Los resultados presentados corresponden a las dos extensiones estudiadas 0.5 km y 2.5 km. y consideran únicamente un 10 % de los buses en circulación, lo que corresponde a tener entre 1 a 3 buses por hora. Respecto al uso de porcentajes mayores al 10 % para contar con un mayor número de vehículos, se observó que el modelo es bastante robusto y no sufre variaciones al disminuir el tamaño de la flota de vehículos. En la tabla 6.1 se presenta el error MSE y MAPE para los tramos de 2.5 km y 0.5 km obtenido por modelos entrenados con información actual de Avda. Pedro de Valdivia para porcentajes de 50 %, 30 % y 10 % de los buses.

Tabla 6.1. Comparación Porcentaje de Datos

Porcentaje	50 %		30 %		10 %	
Modelo	MAPE	MSE	MAPE	MSE	MAPE	MSE
LR 2.5 km	0.064	0.071	0.063	0.070	0.062	0.070
SVR 2.5 km	0.066	0.078	0.066	0.079	0.069	0.086
NN 2.5 km	0.073	0.086	0.072	0.085	0.080	0.097
LR 0.5 km	0.143	0.520	0.136	0.492	0.147	0.516
SVR 0.5 km	0.136	0.487	0.134	0.480	0.135	0.483
NN 0.5 km	0.136	0.492	0.141	0.510	0.147	0.516

## 6.1. Puntos Comparación

Como se describió en la sección 3, los puntos de comparación son el promedio histórico (HA) y el promedio del último periodo (LPA). En tabla 6.2 se presenta el comportamiento de estas predicciones para las dos medidas de distancia estudiadas. Se puede observar que HA obtiene mejores resultados frente a LPA en ambas situaciones. Pese a que HA es una estimación más sencilla de realizar que LPA, debido a que el último requiere información actual o en tiempo real, tiene un comportamiento muy superior en las condiciones estudiadas. Sin embargo, el comportamiento de ambos es bastante inferior a los resultados obtenidos por los modelos que se presentarán en las siguientes subsecciones.

Tabla 6.2. Comparación Base

Distancia (km)	MAPE		MSE	
	2.5	0.5	2.5	0.5
HA	0.1222	0.2702	0.2804	1.6093
LPA	0.1959	0.3321	0.7844	2.6105

### 6.1.1. Información Histórica

En las tablas 6.3 y 6.4 se presentan los resultados de los modelos estimados con información histórica para los tramos de 2.5 km y 0.5 km respectivamente. En el caso de LR, se observa que la incorporación de información asociada a ejes paralelos produce ruido y empeora el resultado de la estimación para el tramo de 0.5 km. Para la sección de 2.5 km se mantienen los resultados al incorporar la información de cada eje paralelo por separado. Sin embargo, cuando se agrega la información asociada a ambos ejes en forma conjunta, los resultados empeoran. Al igual que en LR, los modelos SVR empeoran con la incorporación de información de calles paralelas y se presenta la misma situación para la sección de 2.5 km que en LR (es decir, la predicción mejora al incorporar la información de vías paralelas por separado, pero empeora al incorporar esta información de manera

conjunta). Respecto a NN, la información adicional no tiene grandes efectos para el tramo 2.5 km, pero en el caso de 0.5 km el desempeño empeora en más de un 50 %. Los mejores desempeños para esta situación se obtienen en el modelo LR para el tramo 2.5 km y de SVR para la sección de 0.5 km. En ambos casos sólo utilizando información asociada a la calle de interés.

Tabla 6.3. Histórico 2.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.0588	0.0629	0.0935	0.0625	0.0715	0.1189
PV + LL	0.0587	0.0627	0.0925	0.0626	0.0690	0.1147
PV + MM	0.0586	0.0626	0.0982	0.0605	0.0659	0.1268
PV + LL + MM	0.0608	0.0656	0.0960	0.0625	0.0673	0.1209

Tabla 6.4. Histórico 0.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.1433	0.1303	0.1867	0.5777	0.5315	0.7549
PV + LL	0.1483	0.1328	0.2524	0.6279	0.5558	1.1401
PV + MM	0.1468	0.1356	0.2170	0.5866	0.5207	0.8420
PV + LL + MM	0.1558	0.1340	0.2847	0.6614	0.5417	1.3619

### 6.1.2. Información Actual

En las tablas 6.5 y 6.6 se presentan los resultados de los modelos estimados con información actual para las secciones de 2.5 km y 0.5 km respectivamente. Para LR se observa que la información obtenida de calles paralelas produce mejoras marginales para el tramo de 2.5 km y empeora el resultado en la sección de 0.5 km. En el caso de SVR, se observa

que la incorporación de información adicional produce mejoras en las dos distancias estudiadas. NN produce, en líneas generales, resultados inferiores a los dos modelos anteriores y una degradación al incorporar la información de las calles paralelas. El mejor desempeño para la sección 2.5 km se obtiene con el modelo LR obteniéndose resultados similares en las 4 configuraciones. En el caso del tramo de 0.5 km, el mejor resultado se obtiene con SVR y utilizando información adicional de Avda. Manuel Montt. Como se mencionó en la sección Datos, Avda. Manuel Montt es el eje que presenta mayores similitudes con Avda. Pedro de Valdivia en lo referido a los tiempos de viaje en el sector estudiado.

Tabla 6.5. Información Actual 2.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.0625	0.0690	0.0798	0.0696	0.0864	0.0974
PV + LL	0.0625	0.0671	0.0907	0.0682	0.0806	0.1162
PV + MM	0.0615	0.0662	0.0758	0.0650	0.0780	0.0861
PV + LL + MM	0.0609	0.0662	0.0846	0.0631	0.0757	0.0999

Tabla 6.6. Información Actual 0.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.1396	0.1349	0.1473	0.5105	0.4832	0.5163
PV + LL	0.1392	0.1352	0.1805	0.5168	0.5012	0.6055
PV + MM	0.1405	0.1312	0.1824	0.5143	0.4579	0.6240
PV + LL + MM	0.1395	0.1320	0.2216	0.5182	0.4756	0.7895

### 6.1.3. Información Actual sin Calle Interés

En las tablas 6.7 y 6.8 se presentan los resultados de los modelos estimados con información en tiempo real, y sin información asociada a la calle de interés para los tramos

de 2.5 km y 0.5 km respectivamente. En ambas tablas se incorporaron los resultados de la estimación en tiempo real con información de Avda. Pedro de Valdivia. Para el modelo LR asociado a la sección de 0.5 km, se puede observar que el mejor resultado se obtiene al utilizar sólo la información proveniente de Avda. Los Leones. En cambio, el uso de esta información para el tramo de 2.5 km es la peor opción. En el caso de SVR, para el tramo de 0.5 km, los mejores resultados se obtienen al utilizar la información de Avda. Manuel Montt y para 2.5 km, se obtienen al utilizar los ejes paralelos de forma conjunta. Respecto al modelo NN, éste presenta los peores resultados en comparación a los otros dos modelos, pero también obtiene mejoras al utilizar información del eje Avda. Manuel Montt. Al igual que en los casos anteriores, el mejor resultado para la sección de 2.5 km se obtiene con el modelo LR y con SVR para el tramo de 0.5 km. Para ambos modelos, la información de Avda. Manuel Montt conduce a los mejores resultados en términos de las medidas de error.

Tabla 6.7. Información Actual Sin Calle Interés 2.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.0625	0.0690	0.0798	0.0696	0.0864	0.0974
LL	0.0641	0.0701	0.0955	0.0745	0.0894	0.1267
MM	0.0605	0.0676	0.0798	0.0666	0.0868	0.0985
LL + MM	0.0613	0.0657	0.0929	0.0642	0.0780	0.1141

Tabla 6.8. Información Actual Sin Calle Interés 0.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.1396	0.1349	0.1473	0.5105	0.4832	0.5163
LL	0.1462	0.1364	0.1834	0.5995	0.5403	0.7027
MM	0.1513	0.1276	0.1684	0.6109	0.4684	0.7341
LL + MM	0.1496	0.1366	0.2108	0.6179	0.4962	0.8846

#### 6.1.4. Histórico e Información Actual

En las tablas 6.9 y 6.10 se presentan los resultados de los modelos estimados con información histórica y actual para las secciones de 2.5 km y 0.5 km respectivamente. En esta situación los resultados obtenidos son levemente peores a los que se obtienen utilizando únicamente información histórica. En todas las situaciones la incorporación de datos adicionales de calles paralelas empeoran las estimaciones. Al igual que en los casos anteriores, para el tramo de 2.5 km los mejores resultados se obtienen con el modelo LR y con el modelo SVR para el tramo de 0.5 km.

Tabla 6.9. Histórico e Información Actual 2.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.0604	0.0635	0.0776	0.0611	0.0659	0.0864
PV + LL	0.0620	0.0642	0.1050	0.0648	0.0678	0.1321
PV + MM	0.0607	0.0621	0.0843	0.0585	0.0607	0.0916
PV + LL + MM	0.0605	0.0634	0.0894	0.0595	0.0645	0.1016

Tabla 6.10. Histórico e Información Actual 0.5 km

Configuración	MAPE			MSE		
	LR	SVR	NN	LR	SVR	NN
PV	0.1393	0.1301	0.2057	0.5089	0.4795	0.7760
PV + LL	0.1428	0.1347	0.2517	0.5462	0.5173	1.1985
PV + MM	0.1534	0.1404	0.2395	0.5577	0.5116	1.0332
PV + LL + MM	0.1495	0.1387	0.2859	0.5573	0.5195	1.6551

## **6.2. SVR por Hora**

Como se observa en los resultados anteriores, las mejores estimaciones se obtienen con los modelos SVR y LR, dependiendo de la extensión del trayecto seleccionado. Sin embargo, los modelos estimados con SVR tienen mejoras en su desempeño al incluir información asociada a ejes paralelos. Por esta razón, se selecciona el modelo SVR para realizar un análisis por hora y observar si la información de ejes paralelos genera mejoras importantes en horarios particulares. A continuación, se presentan resultados desagregados a nivel de horas de este modelo en términos de la medida de error MAPE para los trayectos de 2.5 km. Primero se presentan los resultados utilizando únicamente información actual y posteriormente aquellos que utilizan información histórica. En el Anexo A se pueden encontrar los resultados para 0.5 km con información actual y en el Anexo B los resultados para 0.5 km con información histórica.

### **6.2.1. Información Actual 2.5 km**

Para las estimaciones con información actual se considera aquellas que sólo utilizan este tipo de datos. En este caso, se tienen 7 configuraciones, donde las cuatro primeras columnas en las figuras 6.1, 6.2, 6.3 corresponden a estimaciones realizadas utilizando la información asociada a la calle de interés y las siguientes tres, a estimaciones realizadas únicamente con información de calles paralelas. Los gráficos se encuentran divididos en franjas de 8 horas con la misma escala.

En la franja horaria 00:00-08:00, figura 6.1, se observa que la incorporación de información de calles paralelas a la información de Avda. Pedro de Valdivia genera pequeñas mejoras en la estimación salvo en el horario entre 03:00 a 04:00. De las configuraciones probadas, la que presenta mejores resultados en cuanto mejoras, es la utilización de información asociada al eje Avda. Manuel Montt. En 7 casos se observan mejoras entre 0.5 % a 2.5 % y en el horario restante la estimación empeora en un 0.5 %. Respecto a la

utilización de información asociada al eje Avda. Los Leones, sólo existe un horario, donde se obtienen resultados significativamente superiores que utilizando información del eje Avda. Manuel Montt, disminuyendo en casi un 4 % el error en el horario de 04:00-05:00 (en el cual incorporar la información asociada al eje Avda. Manuel Montt se traduce en una mejora de 0.8 %). Además, en la mitad de los casos, incorporar la información del eje Avda. Los Leones produce peores estimaciones que utilizar únicamente la información de Avda. Pedro de Valdivia. Agregar la información de ambos ejes en forma simultánea genera resultados mínimamente superiores a utilizar sólo la información del eje Avda. Manuel Montt, pero aumenta la cantidad de horarios, en los cuales se obtienen peores resultados y aumenta más el error donde Avda. Manuel Montt obtiene peores estimaciones por si sola.

Respecto a las estimaciones sin información de Avda. Pedro de Valdivia, se observa que se pueden conseguir resultados similares a los obtenidos por los modelos que incluyen esta información e incluso mejoras. Al igual que en el caso anterior, los peores resultados se obtienen al utilizar información únicamente de Avda. Los Leones y los mejores resultados se obtiene usando únicamente la información de Avda. Manuel Montt.

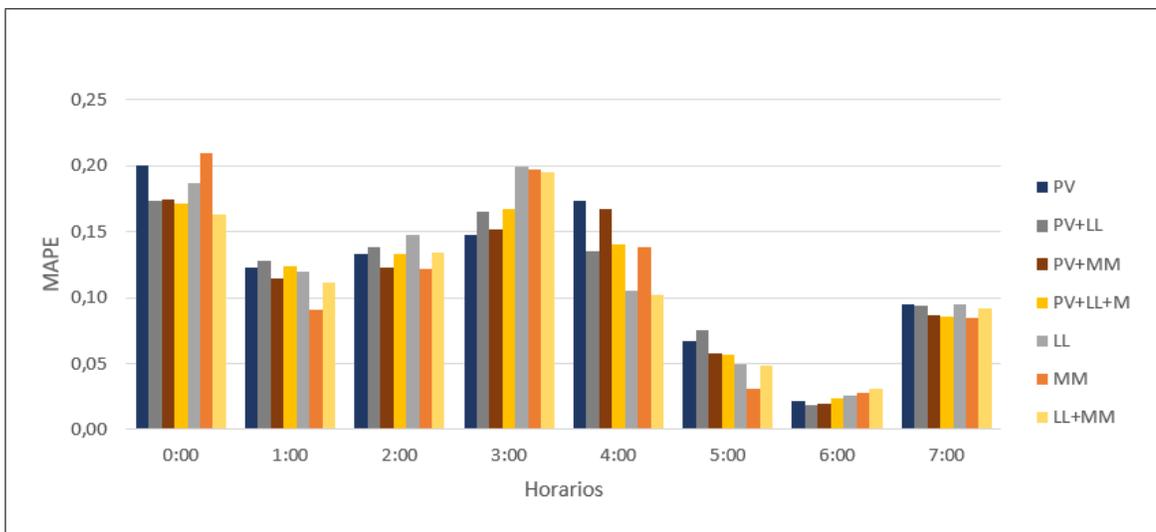


Figura 6.1. SVR MAPE 2.5 km Información Actual 00:00-08:00

En la franja horaria 08:00-16:00, figura 6.2, se observa que el error se encuentra en torno a un 4 % , 5 % lo que indica que los tiempos de viaje presentan menor varianza que en los horarios nocturnos y que posibles mejoras serán de menor magnitud que en el horario anterior. Al igual que para la franja horaria anterior, en la mayoría de los horarios se observan pequeñas mejoras en la estimación cuando se utiliza información de Avda. Pedro de Valdivia, más la de algún eje paralelo. Sin embargo, a diferencia de la franja horaria anterior, los mejores resultados se obtienen al utilizar la información de Avda. Los Leones y no de Avda. Manuel Montt. En 7 de los 8 horarios se pueden observar mejoras pequeñas del orden de 0.2 % a 1 % y en el horario restante se observa un deterioro del 1 %. El utilizar solamente información del eje Avda. Manuel Montt, conduce a los mejores resultados en sólo dos horarios, aunque en ambos casos las mejoras respecto a utilizar la información asociada al Avda. Los Leones son mínimas. Al utilizar la información de ambas calles paralelas de manera conjunta, se obtienen estimaciones similares al utilizar únicamente la información de Avda. Los Leones, pero existe un horario adicional, donde se empeora la estimación.

Para las estimaciones sin información de la calle de interés, se observa la misma situación anterior. Es decir, se obtienen resultados similares a los que se obtienen al utilizar la información del eje Avda. Pedro de Valdivia e incluso algunas mejoras. A diferencia del caso anterior, se observa que la mejor configuración corresponde a utilizar la información de ambas calles laterales.

En la franja horaria 16:00-00:00, figura 6.3, se observa que, al igual que en las dos franjas horarias anteriores, la información extra asociada a las calles paralelas permite, en alguna de sus configuraciones, obtener mejores estimaciones que utilizar únicamente la información de la calle de interés. En este caso se observa que al utilizar información adicional del eje Avda. Manuel Montt se obtienen mejoras de entre 0.1 % a 1.5 % en cinco horarios, en comparación con los resultados obtenidos al utilizar únicamente la información de Avda. Pedro de Valdivia. En los tres horarios restantes, en los cuales utilizar la

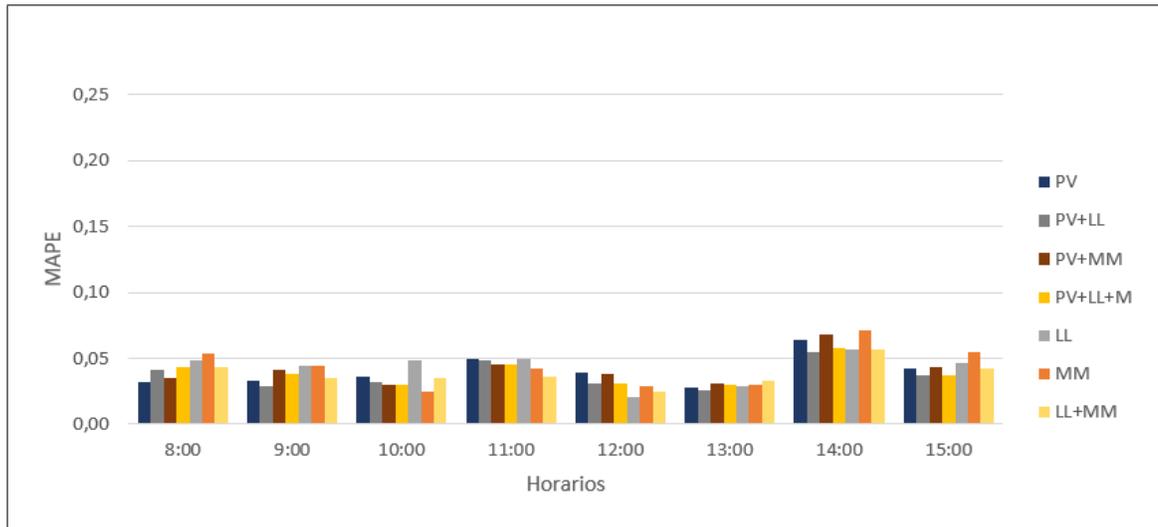


Figura 6.2. SVR MAPE 2.5 km Información Actual 08:00-16:00

información asociada al eje Avda. Manuel Montt resulta en un deterioro de la capacidad predictiva, el utilizar información proveniente de Avda. Los Leones produce una mejora u obtiene al menos el mismo error que al utilizar únicamente la información del eje Avda. Pedro de Valdivia. Pese a lo anterior, utilizar la información de ambos ejes paralelos en la estimación, empeora los resultados en 4 de los 8 horarios.

En lo referido a las estimaciones sin información de la calle de interés, se observa la misma situación anterior: se obtienen estimaciones similares a las que se obtienen utilizando únicamente la información de Avda. Pedro de Valdivia e incluso algunas mejoras. Al igual que el caso anterior, se observa que la mejor configuración corresponde a utilizar la información de ambas calles laterales, salvo en algunos horarios, en los cuales utilizar únicamente datos de Avda. Manuel Montt conduce a resultados superiores a utilizar la información de ambas calles en conjunto.

De los resultados de las estimaciones con datos con información actual se obtiene que en estos casos la información de calles paralelas genera mejoras en la estimación, pero se debe seleccionar con cuidado qué calles serán utilizadas. Como se observó anteriormente,

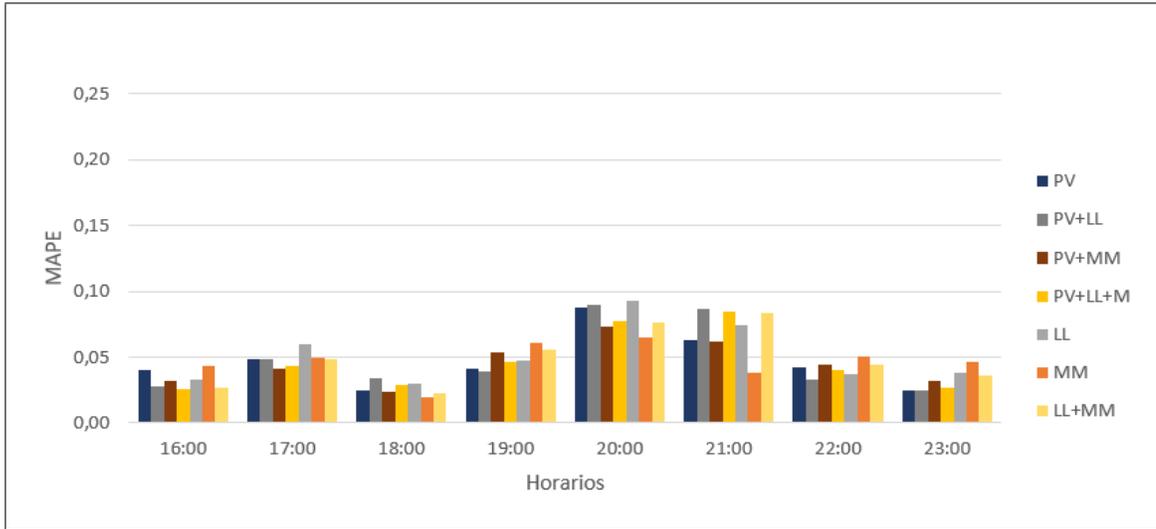


Figura 6.3. SVR MAPE 2.5 km Información Actual 16:00-00:00

Avda. Manuel Montt es el eje que permite, en líneas generales, obtener los mejores resultados en términos del nivel de error, pero existen casos en los que el uso de la información de Avda. Los Leones conduce a mejores modelos. También se observa que el uso de la información de más de una calle paralela no asegura una mejor estimación.

### 6.2.2. Información Histórica 2.5 km

En esta sección se consideran estimaciones que utilizan únicamente información histórica, así como las configuraciones que combinan el uso de información histórica con información en tiempo real. Por tanto, en este caso se consideran ocho configuraciones distintas, en las que los cuatro primeros resultados en las figuras 6.4, 6.5, 6.6 corresponden a estimaciones realizadas únicamente con información histórica, mientras que las siguientes cuatro se realizaron utilizando ambos tipos de datos (la información en tiempo real se indica con una A – actual). Los gráficos se encuentran divididos en franjas de ocho horas con la misma escala.

En la franja horaria 00:00-08:00, figura 6.4, se observa que el uso de información de calles paralelas no genera mejoras en las estimaciones, salvo en los dos últimos horarios de esta franja horaria. Respecto al uso de información actual, se observa que en la mitad de estos horarios (2:00, 3:00, 6:00,7:00), se obtiene mejoras en los resultados, pero en los restantes horarios, los resultados son substancialmente inferiores a los los obtenidos utilizando únicamente información histórica.

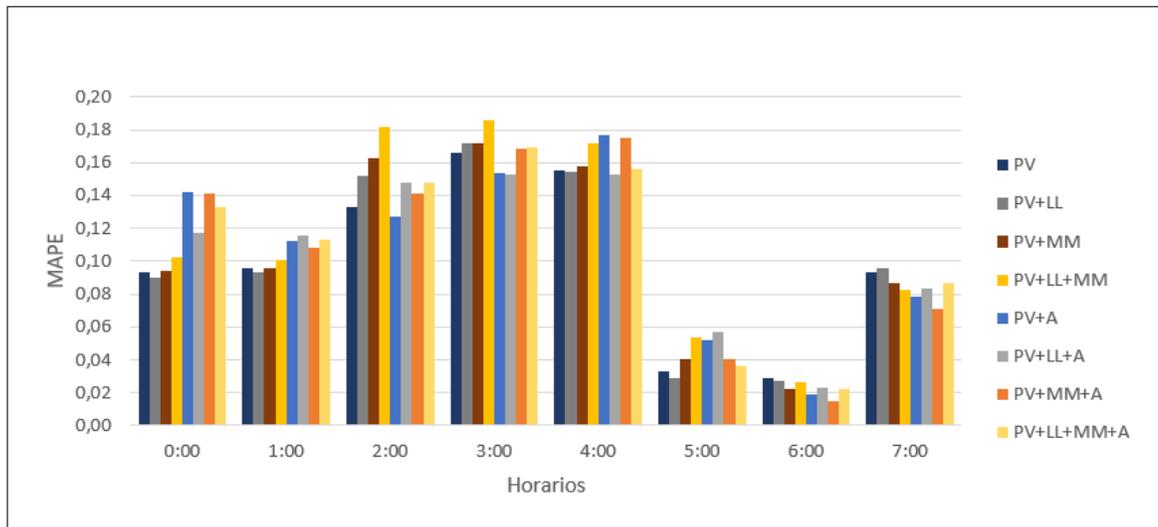


Figura 6.4. SVR MAPE 2.5 km Histórico 00:00-08:00

En la franja horaria 08:00-16:00, figura 6.5 se observa que, al utilizar únicamente información histórica, la información de calles paralelas no genera mejoras considerables, mientras que, al combinar el uso de información histórica y actual, existen pequeñas mejoras de alrededor de 0.5 %. A diferencia de la franja horaria anterior, en este caso las estimaciones con información actual obtienen resultados levemente superiores a los obtenidos utilizando únicamente datos históricos.

Finalmente, en la franja horaria 16:00-00:00, figura 6.6, se observa que la información de calles paralelas genera mejoras en alguna de sus configuraciones, tanto cuando se usa

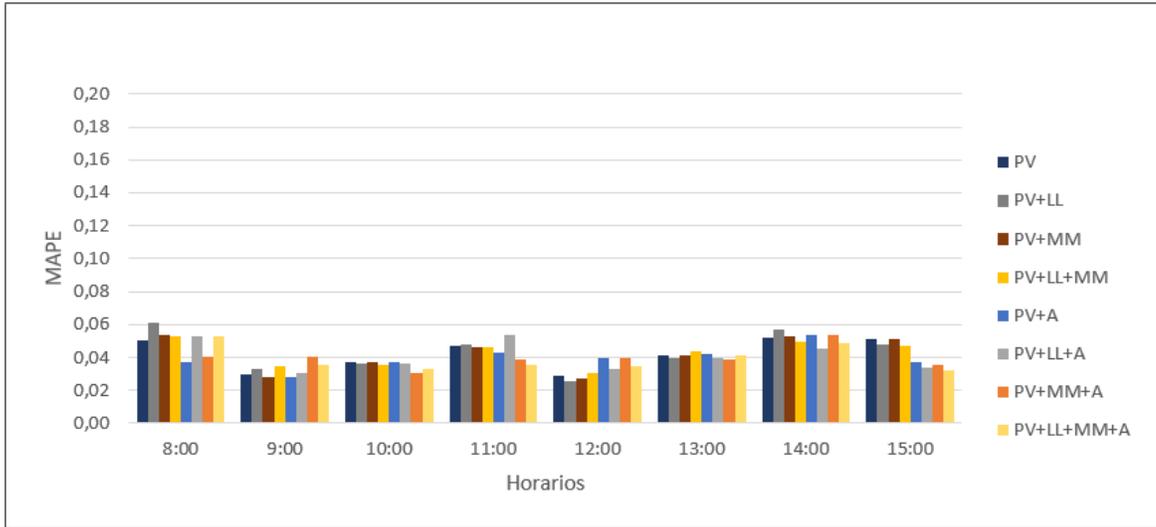


Figura 6.5. SVR MAPE 2.5 km Histórico 08:00-16:00

sólo información histórica como al utilizar información histórica y en tiempo real de manera combinada. Sólo en los horarios iniciados a las 9:00 y 12:00 se observa aumentos en el nivel de error. Al igual que en la franja horaria anterior, los resultados obtenidos con ambos tipos de información son levemente mejores a los obtenidos únicamente con información histórica.

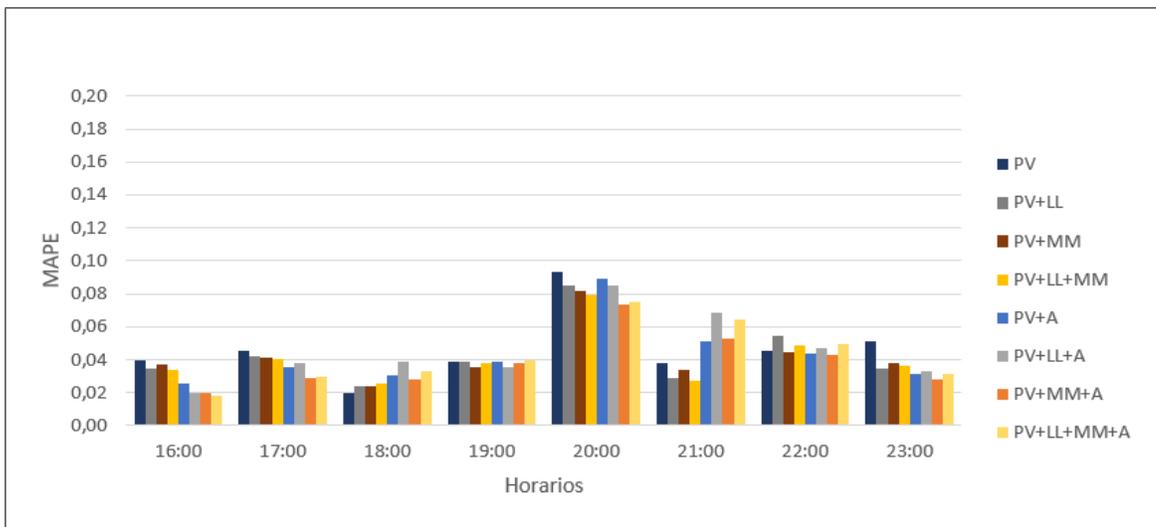


Figura 6.6. SVR MAPE 2.5 km Histórico 16:00-00:00

De los resultados de las estimaciones con datos con información histórica se obtiene que el uso de información de calles paralelas no genera mejoras evidentes en la estimación. Además, se observa que es preferible utilizar sólo la información histórica y no añadir la información actual.

## **CAPÍTULO 7. CONCLUSIONES**

A continuación, se presentan las conclusiones de este estudio. Esta sección se encuentra dividida en dos partes. En primer lugar, se presentan los resultados del trabajo y las conclusiones como tal, y luego, se presentan extensiones para posibles trabajos futuros.

### **7.1. Principales Contribuciones y Resultados**

El aspecto central de esta tesis era determinar si la utilización de información adicional proveniente de ejes paralelos, a un eje para el cual se pretende estimar el tiempo de viaje, permite obtener mejores resultados cuando se cuenta con una baja cantidad de información o directamente no se tiene datos actualizados para el eje de interés. En las pruebas realizadas se observó que este tipo de información puede generar mejoras, pero se debe realizar una selección adecuada de los ejes a utilizar, ya que de lo contrario su uso podría empeorar las predicciones.

En primer lugar, los resultados globales obtenidos indican que, en las mejores configuraciones, las mejoras obtenibles mediante el uso de información de calles paralelas son del orden de 0.3 % cuando el error es de alrededor de un 6 %, en la sección de 2.5km, y de un 0.5 % cuando el error es de alrededor de un 13 % en la sección de 0.5km. Sin embargo, estas disminuciones parecen ser marginales dado el bajo error que se obtiene al utilizar sólo información del eje de interés. Esta situación viene dada por la extensión del trayecto seleccionado, como se observa en los resultados, a mayor extensión, menor error en la predicción utilizando únicamente datos de la calle de interés. También este bajo error se explica por la extensión de los horarios seleccionados. La extensión de una hora simplifica la estimación de un promedio de tiempo de viaje, lo que disminuye el error que se obtiene con cualquiera de los modelos probados. El uso de franjas de menor duración aumentará el nivel de error, pero permitirían obtener estimaciones de mayor precisión para

cada vehículo en un horario específico.

Sin embargo, pese a que las mejoras son pequeñas a nivel global, los beneficios en ciertos horarios específicos son significativos. Al analizar cada uno de los horarios por separado, se observa que existen mejoras de hasta un 5 % cuando el nivel de error alcanza un 20 %. Estos valores dejan de ser marginales y permiten obtener estimaciones considerablemente mejores.

Otro punto importante al momento de analizar el aporte de la información asociada a calles paralelas es que la selección de qué calles utilizar puede variar dependiendo del horario analizado. De los resultados obtenidos, en líneas generales, el utilizar información del eje Avda. Manuel Montt para mejorar la modelación, generaba mayores beneficios que utilizar información de Avda. Los Leones. Pese a ello, en ciertos horarios el uso de información proveniente de este último eje conducía a mejores resultados. Incluso, existen horarios en los cuales el uso de información de Avda. Manuel Montt llevaba a una peor estimación. Esto sugiere que un uso adecuado de las calles seleccionadas, que complemente información faltante, podría conducir a modelos con mejor capacidad predictiva. Esto no implica necesariamente utilizar información de varios ejes a la vez, ya que como se observó en los resultados, esto no asegura mejorías.

En las configuraciones, en las cuales no se contaba con información de la calle de interés, las estimaciones realizadas únicamente con datos de calles paralelas permiten obtener predicciones muy similares en términos del nivel de error, a aquellas que sí contaban con información de la calle de interés. En estas configuraciones, la adecuada selección de los ejes que se utilizarán para la modelación es la única forma de obtener resultados adecuados. Por lo tanto, la selección de calles es fundamental para que este tipo de datos

adicionales sean una ayuda en las estimaciones del tiempo de viaje.

Respecto al uso de información actual o histórica, se obtuvo que las estimaciones realizadas únicamente con información histórica presentaban un menor nivel de error que aquellas realizadas con información actual, lo que viene dado por la extensión de los horarios. La información actual se vuelve menos relevante al aumentar el largo de cada ventana horaria. Sin embargo, en los casos en los cuales se utilizó información histórica, la información de calles paralelas no generó mejoras significativas en la estimación. Esto indica que el uso de información de calle paralelas es más promisorio cuando se realizan estimaciones con información actual o en tiempo real.

Finalmente, respecto a los modelos de aprendizaje considerados durante este estudio, los que exhiben un peor rendimiento son las redes neuronales (NN). Aun así, este resultado era esperable, ya que se trabajó con una baja cantidad de información, lo que en muchas ocasiones produce que este tipo de modelos se sobreentrene, produciéndose un sobreajuste, y, por tanto, observándose un comportamiento no adecuado. En el caso de la regresión lineal múltiple (LR) y de la regresión de vectores de soporte (SVR), los resultados obtenidos fueron muy similares. Sin embargo, el uso de información adicional asociada a calles paralelas tiende a presentar mejores resultados, en el caso de los modelos SVR. Por ello, si se pretende complementar información incompleta con información asociada a ejes paralelos, se recomienda el uso de modelos SVR.

## **7.2. Trabajo Futuro y Extensiones**

A continuación, se presentan posibles líneas de trabajo futuro y extensiones, en base a los resultados obtenidos en este trabajo:

- Analizar el uso de ventanas horarias de menor extensión (inferior a una hora) que disminuyan el espacio temporal entre la recolección de información y la predicción, sobre todo considerando que, en una ciudad como Santiago de Chile, el nivel de congestión puede presentar amplia variabilidad en el transcurso de una hora. De esta forma la información actual podría llegar a tener una mayor relevancia en la estimación, al ser utilizada como complemento de la información histórica.
- Realizar estimaciones por vehículo, en lugar de estimaciones de tiempo de viaje promedio en eje durante un periodo de tiempo.
- Determinar la forma de identificar secciones de ejes en vías paralelas, que permitan obtener la mejor información posible para la predicción en la calle de interés.
- Utilizar calles en más de un sentido. Determinar así, si flujos en sentidos opuestos o perpendiculares pueden aportar en la estimación.
- Considerar la inclusión de zonas alejadas al área de estudio en búsqueda de ejes que permitan mejorar la estimación en la calle de interés. Buscar zonas de la ciudad que puedan tener patrones similares al de estudio.
- Realizar un estudio con una flota de vehículos distinta. Buses de transporte público pueden tener un comportamiento muy homogéneo que no permite observar beneficios del uso de información de calles paralelas.
- Considerar posibles mejoras en la creación de trayectos y en la concatenación de pulsos GPS para cada uno de los vehículos utilizados en el cálculo de los promedios horarios.

## BIBLIOGRAFÍA

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

DTPM. (2017). *Informe de gestión dtpm 2015-2016* (Inf. Téc.). Directorio de Transporte Metropolitano.

Fellendorf, M. (1994). Vissim: A microscopic simulation tool to evaluate actuated signal control including bus priority. En *64th institute of transportation engineers annual meeting* (pp. 1–9).

Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Fusco, G., Colombaroni, C., Comelli, L., y Isaenko, N. (2015). Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models. En *Models and technologies for intelligent transportation systems (mt-its), 2015 international conference on* (pp. 93–101).

Gurmu, Z. K., y Fan, W. D. (2014). Artificial neural network travel time prediction model for buses using only gps data. *Journal of Public Transportation*, 17(2), 3.

Hornik, K., Stinchcombe, M., y White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366.

Jenelius, E., y Koutsopoulos, H. N. (2013). Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53, 64–81.

Julio, N. (2015). Aplicación de algoritmos de aprendizaje estadístico para predecir velocidades de buses con información en tiempo real.

Kriesel, D. (2007). A brief introduction on neural networks.

McCulloch, W. S., y Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115–133.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Nilsson, N. J. (1996). Introduction to machine learning. an early draft of a proposed textbook.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Rahmani, M., Jenelius, E., y Koutsopoulos, H. N. (2015). Non-parametric estimation of route travel time distributions from low-frequency floating car data. *Transportation Research Part C: Emerging Technologies*, 58, 343–362.

Ramezani, M., y Geroliminis, N. (2012). On the estimation of arterial route travel time distribution with markov chains. *Transportation Research Part B: Methodological*, 46(10), 1576–1590.

Shalev-Shwartz, S., y Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Tibshirani, R., James, G., Witten, D., y Hastie, T. (2013). *An introduction to statistical learning-with applications in r*. New York, NY: Springer.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V., y Chervonenkis, A. (1964). On a perceptron class. *Automation and Remote Control*, 25, 112–120.

Vapnik, V., y Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, 24, 774–780.

Wang, Y., Zheng, Y., y Xue, Y. (2014). Travel time estimation of a path using sparse trajectories. En *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 25–34).

Zhan, X., Hasan, S., Ukkusuri, S. V., y Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33, 37–49.

Zheng, F., y Van Zuylen, H. (2013). Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 31, 145–157.

## **ANEXO**

## A. RESULTADOS SVR POR HORA 0.5 KM CON INFORMACIÓN ACTUAL

Se consideran siete configuraciones, donde las primeras cuatro columnas presentadas en las figuras A.1, A.2, A.3 corresponden a estimaciones realizadas utilizando información de la calle de interés, mientras que las siguientes tres estimaciones son realizadas únicamente con información de calles paralelas.

En la franja horaria 00:00-08:00, figura A.1, se observa que incorporar información de calles paralelas genera mejoras en la mitad de los horarios con alguna de las configuraciones probadas. Los mejores resultados se obtienen con las configuraciones que utiliza ambas calles paralelas o Avda. Manuel Montt como única fuente de información de ejes paralelos. En los rangos horarios que inician en 0:00, 2:00, 4:00 y 7:00 se observan disminuciones del error de entre 1 % a 7 %. En el caso de las estimaciones sin información de la calle de interés, se observa que se obtienen estimaciones similares a las obtenidas de Avda. Pedro de Valdivia, salvo en el horario 00:00-01:00, en el cual se obtiene una mejora al utilizar únicamente Avda. Manuel Montt. Sin embargo, no se observa una tendencia clara de cuál de las tres configuraciones obtiene mejores resultados.

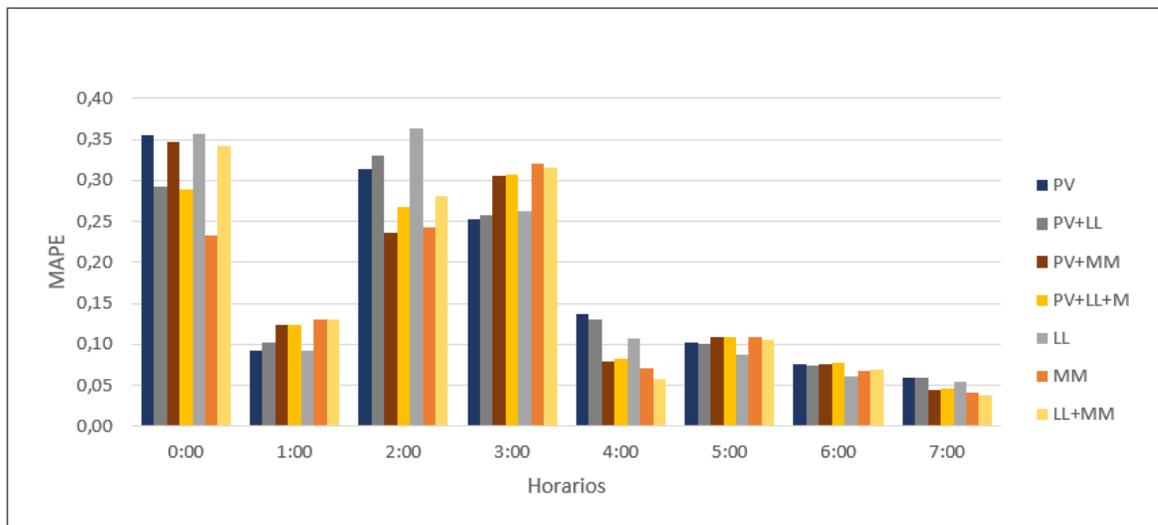


Figura A.1. SVR MAPE 0.5 km Información Actual 00:00-08:00

En la franja horaria de 08:00-16:00, figura A.2 se observa que los mejores resultados son obtenidos al utilizar únicamente información de Avda. Manuel Montt. Sin embargo, en los rangos horarios que inician a las 8:00, 12:00 y 15:00 se obtienen disminuciones de entre 1 % y 4 %. En los rangos horarios que inician a las 9:00, 13:00 y 14:00 aumenta el error una magnitud similar a las mejoras obtenidas en los rangos con disminuciones del error. Al utilizar información de Avda. Los Leones, se observa el mismo comportamiento que en la franja horaria anterior, es decir, estimaciones con un nivel de error similar al obtenido al utilizar únicamente información de Avda. Pedro de Valdivia, pero aumentos en el error en la mayoría de los horarios.

En el caso de las estimaciones sin información de la calle de interés, se observa que los resultados son similares a los obtenidos utilizando información completa. Sin embargo, no se observa una tendencia clara.

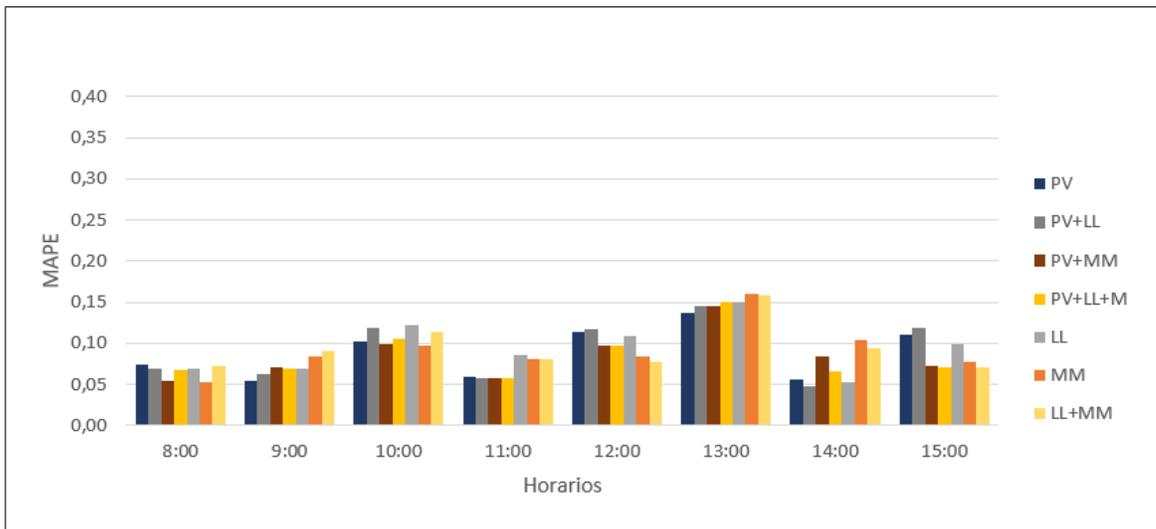


Figura A.2. SVR MAPE 0.5 km Información Actual 08:00-16:00

En la franja horaria de 16:00-00:00, figura A.3, se observa que, en los horarios de mayor congestión (18:00-22:00), utilizar información de calles paralelas empeora la estimación respecto a los resultados obtenidos utilizando únicamente Avda. Pedro de Valdivia.

En los cuatro rangos horarios restantes, que inician a las 16:00, 17:00, 22:00 y 23:00, se observa que las calles paralelas generan disminuciones en el error de estimación. Es importante notar que en los casos en que las estimaciones obtenidas utilizando la calle de interés y la información de un eje paralelo obtienen un aumento del error, es decir información de Avda. Pedro de Valdivia con Avda. Los Leones o Avda. Manuel Montt, el utilizar la información de ambos ejes paralelos en forma conjunta aumenta más el nivel de error, como se observa en los rangos horarios que inician a las 18:00, 19:00 y 21:00. De la misma forma, cuando el error disminuye al añadir uno de los ejes paralelos, el uso en forma conjunta lleva a disminuciones mayores como se observan en los rangos horarios que inician a las 16:00, 17:00 y 22:00. Esta situación también se observa cuando se realiza la estimación sin información de la calle de interés.

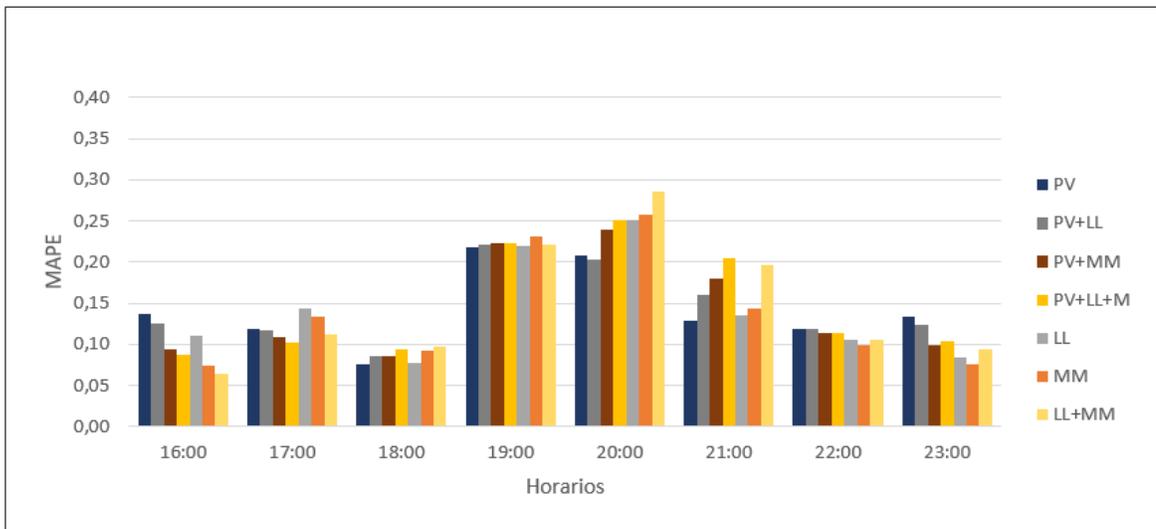


Figura A.3. SVR MAPE 0.5 km Información Actual 16:00-00:00

## B. RESULTADOS SVR POR HORA 0.5 KM CON INFORMACIÓN HISTÓRICA

Se consideran ocho configuraciones, donde las primeras cuatro columnas presentadas en las figuras B.1, B.2, B.3 corresponden a estimaciones realizadas únicamente con información histórica y las siguientes cuatro se realizaron utilizando conjuntamente información histórica e información actual (indicada con una A).

En la franja horaria 00:00-08:00, figura B.1 se observa que en general la incorporación de información de calles paralelas resulta en un deterioro de la capacidad predictiva de los modelos al utilizar únicamente información histórica. Cuando la información histórica se complementa con información actual, la información asociada al eje paralelo de Avda. Los Leones genera leves mejoras en la estimación. De los ocho horarios considerados, en cuatro (0:00, 2:00, 4:00, 7:00) se observa una disminución del error de entre 1 % a 6 % , mientras que en los restantes cuatro el error aumenta entre 1 % a 2.5 %.El uso de información actual no ayuda a mejorar la estimación en comparación con el uso único de datos históricos.

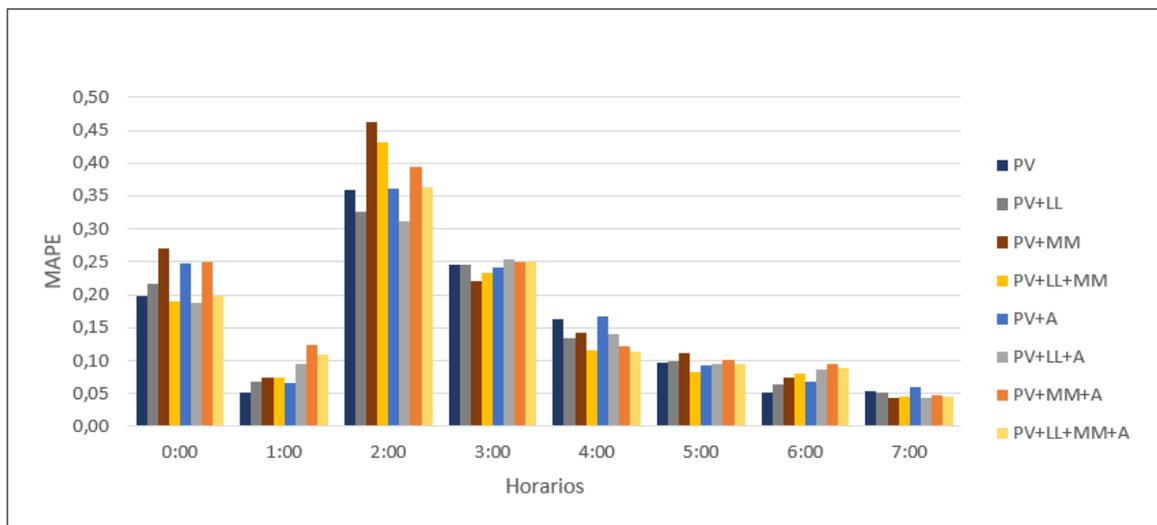


Figura B.1. SVR MAPE 0.5 km Histórico 00:00-08:00

En la franja horaria 08:00-16:00, figura B.2, se observa que, en la mayoría de los casos, la información de calles paralelas no conduce a mejoras en los resultados, cuando se usa solo información histórica. Cuando se incorpora la información actual, el uso de información de calles paralelas empeora la estimación en la mayoría de los horarios.

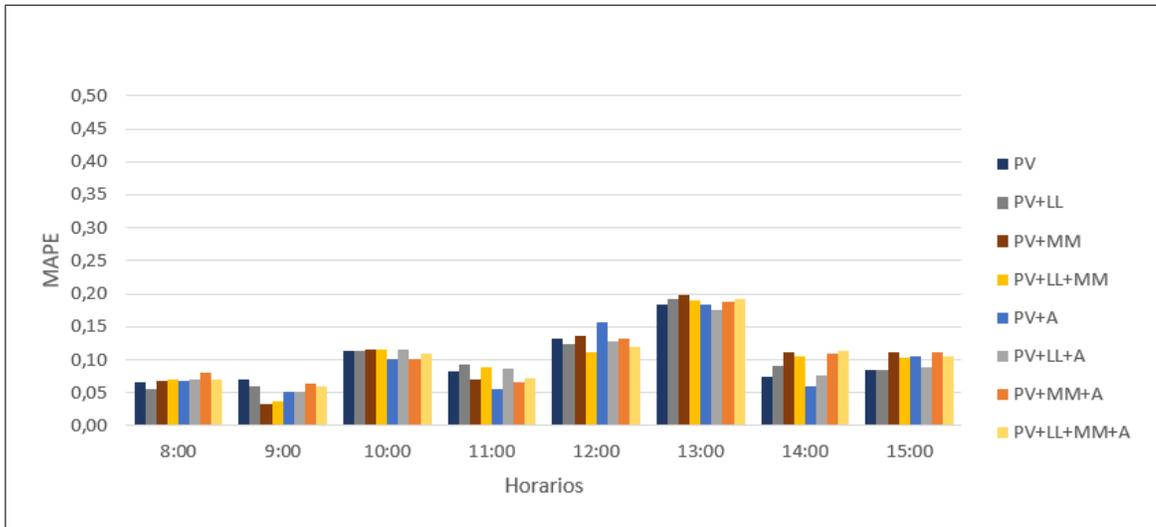


Figura B.2. SVR MAPE 0.5 km Histórico 08:00-16:00

En la franja horaria 16:00-00:00, figura B.3, cuando se utiliza únicamente información histórica, la información de calles paralelas asociada a Avda. Manuel Montt, resulta en disminuciones del nivel de error de entre 1 % y 3 % en seis de los horarios, y que en los dos restantes se observan aumentos de 0.5 % y 1 %. En cambio, cuando se utiliza información histórica y actual, la incorporación de datos de calles laterales empeora la estimación en seis horarios y en los dos horarios restantes se obtiene un error similar.

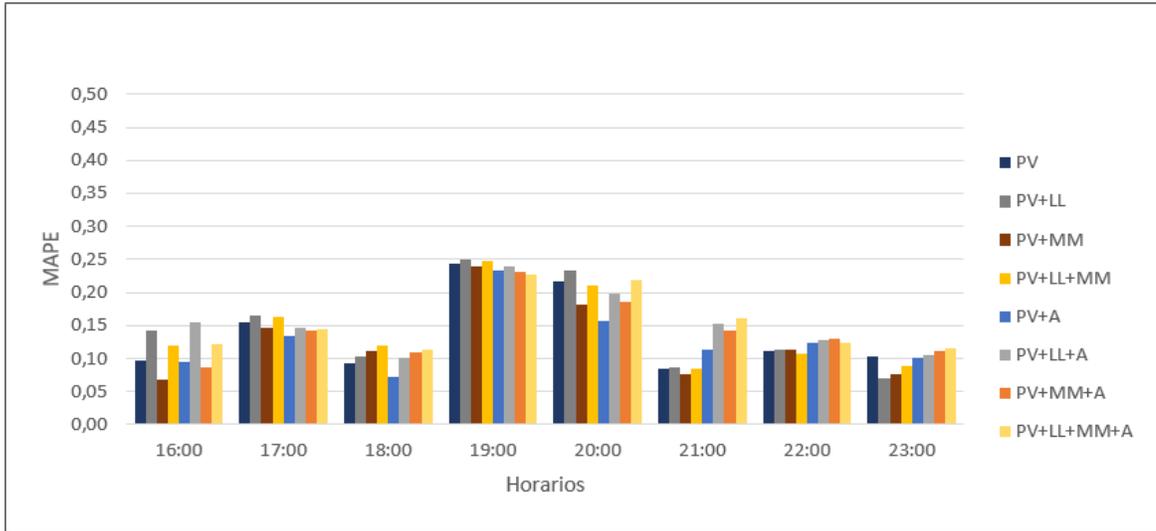


Figura B.3. SVR MAPE 0.5 km Histórico 16:00-00:00