PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MODELOS BASADOS EN DATOS GPS PARA DETERMINAR LA NO DETENCIÓN DE BUSES EN PARADEROS

NICOLÁS ANDRÉS GARCÍA GERBELLA

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

Profesor Supervisor:
JUAN CARLOS HERRERA MALDONADO

Santiago de Chile, abril 2017

© 2017, Nicolás García Gerbella

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MODELOS BASADOS EN DATOS GPS PARA DETERMINAR LA NO DETENCIÓN DE BUSES EN PARADEROS

NICOLÁS ANDRÉS GARCÍA GERBELLA

Tesis presentada a la Comisión integrada por los profesores:

JUAN CARLOS HERRERA MALDONADO

HANS LOBEL DÍAZ

ANTONIO GSCHWENDER

ALFREDO SERPELL BLEY

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

Santiago de Chile, abril 2017

A mi familia

AGRADECIMIENTOS

Quiero agradecer la oportunidad de poder realizar esta investigación, tanto a la escuela de ingeniería, como al departamento de Transporte y Logística. Además, agradecer a mi profesor supervisor, Juan Carlos Herrera. Su dedicación y preocupación, fueron clave para el desarrollo de esta tesis.

También me gustaría agradecer a los demás profesores del departamento, tanto por generar un ambiente grato de trabajo, como su apoyo cuando ha sido necesario. Agradecer en particular al profesor Hans Lobel, que me ayudó a descubrir procedimientos y metodologías nuevas para mí. Incluir al profesor Karim Pichara, que me aconsejó acerca que material de apoyo hubiese sido útil para la investigación. No puedo dejar de incluir a Owen Bull, su investigación y trabajo son lo que inspiran esta tesis, y darme la posibilidad de continuar su investigación ha sido una experiencia satisfactoria.

Por otro lado, agradecer a todas las entidades y personas que hicieron que fuera posible obtener datos reales de buses. Agradecer al Centro de Desarrollo Urbano Sustentable (CEDEUS, CONICYT/FONDAP 15110020). A Dictuc por los dispositivos GPS proporcionados y a los miembros de TransitUC por hacer posible trabajar con datos online de SONDA y Transantiago. Además, agradecer al Directorio de Transporte Público Metropolitano, a su director Guillermo Muñoz y a todas las personas que estuvieron involucradas de forma directa o indirecta en la obtención de datos para esta investigación. En especial agradecer a Sebastián Tamblay, que hizo posible el enlace con el DTPM y dedicó parte de su tiempo a resolver dudas acerca de esta investigación.

También agradecer a todos los funcionarios, administrativos, y compañeros del Departamento de Ingeniería de Transporte y Logística, ya que de alguna u otra manera me apoyaron en la realización de mi magíster.

Por último, agradecer a mi familia, ellos me han entregado todo el apoyo necesario para que pueda desarrollarme en lo que me gusta. Gracias por su paciencia, ánimo y amor incondicional.

TABLA DE CONTENIDOS

AGRADECIMIENTOS iii

ÍNDICE DE TABLAS vi

ÍNDICE DE FIGURAS viii

RESUMEN x

ABSTRACT xii

1 INTRODUCCIÓN 1

1.1 Objetivos 3

1.2 Contenidos 3

2 MARCO TEÓRICO 5

2.1 Modelo de Bull y Herrera (2015) 5

2.2 Métodos de aprendizaje supervisado 13

2.2.1 Máquinas de vectores de soporte (SVM) 13

2.2.2 Análisis discriminante lineal (LDA) 16

2.3 Indicadores y test estadísticos 17

2.3.1 Validación cruzada 17

2.3.2 Matriz de confusión 18

3 METODOLOGÍA 21

3.1 Recolección de datos 21

3.1.1 Base de datos 1 – Buses de Transantiago 21

3.1.2 Base de datos 2 – Datos particulares 24

3.2 Análisis de modelos basados en cinemática 26

3.2.1 Validación Bull y Herrera 27

3.2.2 Variaciones del modelo de Bull y Herrera 29

3.2.3 Análisis de sensibilidad del modelo Bull y Herrera 36

3.3 Modelos basados en aprendizaje supervisado 39

3.4 Comparación de modelos 42

3.5 Análisis respecto a la periodicidad de datos 42
3.5.1 Modelo basado en cinemática 43
3.5.2 Modelos basados en aprendizaje supervisado 44
4 RESULTADOS 45
4.1 Modelos basados en cinemática 45
4.1.1 Validación y análisis de error modelo Bull y Herrera 45
4.1.2 Cambios realizados al modelo Bull y Herrera 48
4.1.3 Análisis de sensibilidad del modelo 50
4.2 Modelos Basados en aprendizaje supervisado 58
4.2.1 Máquinas de vectores de soporte (SVM) 58
4.2.2 Análisis de Discriminante Lineal (LDA) 62
4.3 Comparación de modelos 65
4.4 Análisis periodicidad datos GPS 66
4.4.1 Modelo basado en cinemática 66
4.4.2 Modelos basados en aprendizaje supervisado 69
5 CONCLUSIÓN 76
5.1 Respecto a los objetivos 76
5.2 Extensiones 77
BIBLIOGRAFÍA 79
A N E X O S 81
Anexo 2.1 – Casos 2 y 3 – Modelo Bull y Herrera 2015 82
Caso 2 – Un log GPS en el área de estacionamiento 82
Caso 3 – Más de un log GPS en el área de estacionamiento 85
Anexo - Programación del modelo 87
Anexo 3.1.3 - Preparación de rutas y paraderos 91

ÍNDICE DE TABLAS

Tabla 2.1 Resumen de restricciones Modelo Bull y Herrera 12
Tabla 3.1 Valores que entregan los datos GPS 22
Tabla 3.2 Resumen de observaciones obtenidos con datos GPS propios 26
Tabla 3.3 Velocidad máxima – Servicio AA2001 27
Tabla 3.4 Aceleración y desaceleración promedio, junto al promedio en valor absoluto del Servicio AA2001 28
Tabla 3.5 Valores de h1 y h2 al variar el largo del paradero 37
Tabla 3.6 Valores de h1 y h2 para distintas ubicaciones de paraderos 39
Tabla 3.7 Variables descriptivas para modelos basados en aprendizaje supervisado 40
Tabla 3.8 Observaciones con variables para modelos de aprendizaje supervisado 41
Tabla 3.9 Numero de validaciones para los distintos valores de τ 43
Tabla 4.1 Indicadores clasificación modelo Bull y Herrera 45
Tabla 4.2 Observaciones y error por Caso 46
Tabla 4.3 Comparación de indicadores entre modelo original y variación aceleración 48
Tabla 4.4 Comparación de indicadores entre modelo original y al utilizar distancias 49
Tabla 4.5 Comparación de indicadores entre modelo original y variación con velocidad intantánea 49
Tabla 4.6 Comparación de indicadores entre modelo original y variación al relajar restricciones 50
Tabla 4.7 Valor del indicador FPR al realizar análisis de sensibilidad 51
Tabla 4.8 Valor del indicador FNR al realizar análisis de sensibilidad 51
Tabla 4.9 Valor de la Precisión al realizar análisis de sensibilidad 52
Tabla 4.10 Indicadores para distintos largos de paradero 57
Tabla 4.11 Una cruz marca las variables predictivas que se utilizan en el modelo SVM 59
Tabla 4.12 Indicadores modelos SVM – Clasificación 61
Tabla 4.13 Variables predictivas que utiliza el modelo LDA 63
Tabla 4.14 Indicadores modelos SVM – Clasificación 63
Tabla 4.15 Observaciones y error por casos para τ =10 68

Tabla 4.16 Variables predictivas utilizadas en los modelos SVM.....69
Tabla 4.17 Variables predictivas utilizadas en los modelos de LDA 73
Tabla 4.18 Indicadores modelos LDA– Clasificación.....74
Tabla 7.1 Resumen de restricciones Modelo Bull y Herrera.....89

ÍNDICE DE FIGURAS

Figura 1.1 El porcentaje de reclamos por no dentención en paradero ha ido al alza desde el año 2011 respecto a los reclamos totales que recibe el DTPM 1

Figura 2.1 El Primer acercamiento a la formulación del modelo es determinar cuánto tiempo le toma al bus recorrer una distancia L a velocidad máxima, para así determinar el valor de τs. 6

Figura 2.2 La diferencia horizontal entre la trayectoria a velocidad máxima y un proceso de aceleración hasta velocidad máxima, determina el tiempo extra que toma este proceso, definiéndolo como τa. 8

Figura 2.3 Siendo la franja celeste el área de estacionamiento, los Casos 1, 2, y 3 representan: No hay log GPS en SA, hay un log GPS en SA, o más de un log GPS en SA. 9

Figura 2.4 Ejemplo - Caso 1 presenta la relajación respecto al supuesto inicial, donde el log GPS XD sugiere que el proceso de aceleración no ocurre por completo. 10

Figura 2.5 Representación gráfica de una transformación a espacio de características donde las clases son separables linealmente. 14

Figura 2.6 Se muestra un mismo set de datos separados por dos hiperplanos distintos, el de la derecha logra maximizar el margen entre los datos de las distintas categorías, siendo Tmax el margen óptimo 15

Figura 2.7 La matriz de confusión presenta los resultados de una clasificación bianria, separando los resultados en cuatro categorías. 18

Figura 2.8 Matriz de confusión que represnta la clasificación que se realiza en esta investigación, clasificando el Error Tipo I como los buses que se detienen y el modelo predice que no se detuvo. 19

Figura 3.1 Diagrama de ubicación del paradero PE113 en A.v Santa Rosa al sur de Alborada, y los servicios que se observan desde el paradero. 23

Figura 3.2 Se pueden generar distintos sets de datos a partir de una muestra, seleccionando datos a una periodicidad determinada como los presentados por las distintas tonalidades de azul. 25

Figura 3.3 Visualización de los valores h1 y h2 en un paradero 38
Figura 4.1 Clasificación Modelo Bull y Herrera –Caso 1 al graficar las ditancias de los GPS antes y después del paradero 47
Figura 4.2 FPR y su desviación estándar – Vmax 53
Figura 4.3FNR y su desviación estándar - Vmax 53
Figura 4.4 FPR y su desviación estándar – Caso aceleración.....54
Figura 4.5 FNR y su desviación estándar - Caso aceleración.....55
Figura 4.6 Valores de indicadores al variar largo de paradero 56
Figura 4.7 Valores de indicadores al variar la ubicación del paradero.....57
Figura 4.8 Base de datos 1 por clase al graficar las variables v_inicial y v_paradero.....60
Figura 4.9 Datos de la clasificación del modelo SVM sobre los datos de entrenamiento....61
Figura 4.10 Datos de la clasificación del modelo LDA sobre los datos de entrenamiento ..64
Figura 4.11 Indicadores de los distintos modelos de predicción- Base de datos 1 65
Figura 4.12 Valores de indicadores al variar espaciamiento – Modelo B&H 67
Figura 4.13 Variables v_inicial y v_paradero para τ=10 70
Figura 4.14 Variables v_inicial y v_paradero para τ=60 70
Figura 4.15 Valores de indicadores al variar espaciamiento – Modelo SVM 71
Figura 4.16 Clasificación para modelo SVM τ=30.....72
Figura 4.17 Valores de indicadores al variar espaciamiento – Modelo LDA.....74
Figura 4.18 Clasificación para modelo LDA τ=30 75
Figura 7.1 Ejemplo - Caso 2 82
Figura 7.2 Ejemplo 2 - Caso 2 83
Figura 7.3 Funcionamiento Algoritmo Bull y Herrera 90
Figura 7.4 Referenciación de paraderos a la ruta.....93

RESUMEN

En el sistema de transporte público de Santiago ocurre con frecuencia que los buses no se detienen en paradas establecidas cuando hay pasajeros esperando por subir. Este comportamiento es muy dañino para los usuarios, ya que aumenta su tiempo de espera y reduce la confiabilidad en el sistema. El Directorio de Transporte Público Metropolitano (DTPM), recibe reclamos a diario reportando este tipo de comportamiento. Sin embargo, desde un punto de vista legal, estos reclamos no son suficientes para generar una multa a los conductores, ya que es necesario comprobar que el bus efectivamente no se detuvo.

Con este propósito, los dispositivos GPS instalados en los buses pueden ser de uso. Desafortunadamente, estos dispositivos no proveen la trayectoria exacta del bus, sino que registran su posición cada 30 segundos. Un estudio previo (Bull & Herrera, 2015) desarrolla un modelo que usa GPS cada 30 segundos, para determinar si el bus se detuvo en la parada. Sin embargo, esta metodología no ha sido validada.

El objetivo principal de esta tesis es definir un modelo que logre determinar, con la mayor precisión posible, cuándo los buses no se detienen en las paradas establecidas. Para lograr esto, primero se evalúa y valida el modelo de Bull y Herrera (2015) utilizando datos reales. Basándose en estos resultados, mejoras al modelo son propuestas. En segundo lugar, modelos de aprendizaje supervisado son desarrollados utilizando dos métodos: Máquinas de vectores de soporte y Análisis de discriminante lineal.

Para este proceso se utilizan dos bases de datos distintas. La primera utiliza datos proporcionados por el Directorio de Transporte Público Metropolitano, con un espaciamiento de 30 segundos entre registros GPS; esta se utiliza para calibrar y validar los modelos. El segundo conjunto considera datos GPS cada un segundo, que permite realizar un análisis de sensibilidad respecto a la periodicidad de los datos, y cómo afecta la precisión de los modelos.

El desempeño de los distintos modelos bordea el 89% de precisión. Sin embargo, se concluye que el modelo a recomendar depende principalmente del objetivo del modelador, ya que se puede lograr una mejor precisión con ciertos modelos, pro por otro lado se puede disminuir el Error Tipo I con otro.

ABSTRACT

In Santiago's (Chile) public transport system, buses skip a stop even though there are passengers waiting at the stop. This behaviour is very damaging for the public transport system since it increases user's waiting times, and it reduces user's reliability and confidence in the system, attributes that are very hard to regain. At the Metropolitan Public Transport Department (agency in charge of Santiago's public transport system), several complaints reporting this behaviour are received daily. However, from the legal point of view, these complaints are not enough to issue a fine to the bus drivers, since it needs to be proved that the bus skipped the stop. For this purpose, GPS devices installed on every bus of the system could be of use. Unfortunately, these devices do not provide the exact trajectory of each bus but only its position every 30 seconds.

Previous research (Bull and Herrera, 2015) developed a kinematic based model that uses 30 sec GPS data to determine whether a bus skipped a formal stop or not. The main goal of this study is twofold. First, it assesses and validates this model using real data. Based on the results of this assessment, model's improvements are proposed to achieve better results. Second, machine learning models are also developed and their performances are compared to those of the first model. The models tried are: Support Vector Machine and Linear discriminant analysis.

For the assessment, two different datasets are used. The first one considers data provided by the Metropolitan Public Transport Department, and has a time resolution of 30 seconds between GPS logs. The second dataset uses GPS data reported every second, which allows us to study the impact of using different than 30 sec time intervals between consecutive GPS logs.

The performance of the different models show an accuracy around 89%. However, it is concluded that the model to be recommended depends on the purpose being achieved, since one model can achieve a higher precision, while other model can achieve a lower Type I Error.

1 INTRODUCCIÓN

El Directorio de Transporte Público Metropolitano (DTPM) recibe reclamos a diario reportando distintas situaciones. Desde el inicio de Transantiago la mayor cantidad de denuncias se ha concentrado en dos motivos: la frecuencia de los buses y la no detención de buses en paraderos. Durante el 2016, se registraron cerca de 28 mil reclamos, y más de 11 mil de ellos se deben a la no detención de buses, lo que corresponde al 40 por ciento del total de los reclamos y denuncias (DTPM, 2016). La Figura 1.1 grafica la evolución de los dos principales tipos de reclamos a lo largo de los años.

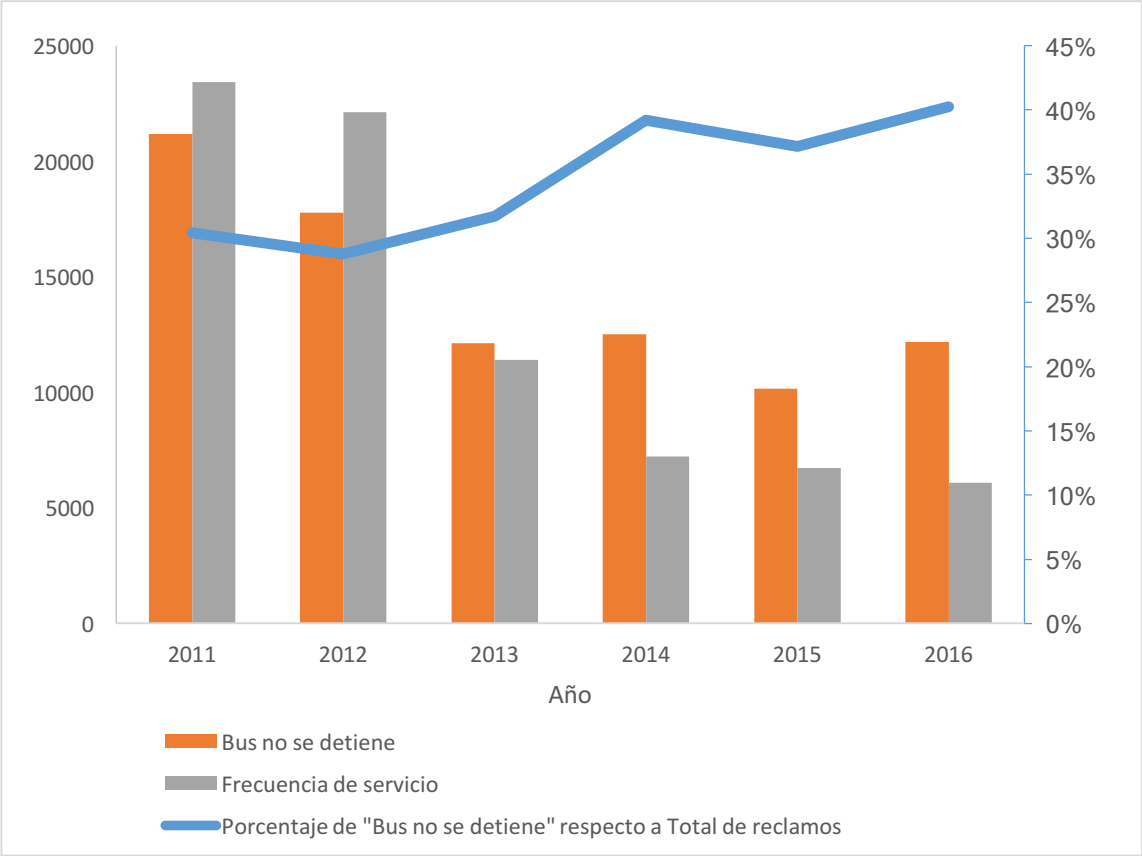

Figura 1.1 El porcentaje de reclamos por no dentención en paradero ha ido al alza desde el año 2011 respecto a los reclamos totales que recibe el DTPM

Fuente: Elaboración propia – Datos Directorio de Transporte Público Metropolitano (2016)

Para un usuario, la no detención de un bus se traduce en una peor calidad de servicio ya que disminuye su confiabilidad en el sistema, atributo cuyo nivel es difícil de recuperar. Esto se refleja en la evaluación que realizan los usuarios al sistema de transporte público, obteniendo cinco veces consecutivas una nota de 4,7 (DTPM, 2015).

Por lo anterior, y para evitar la no detención de buses en paraderos aun cuando hay usuarios esperando por subir, independiente de su causa, se ejecuta una multa al momento que se observa este tipo de comportamiento. Esta multa se le realiza directamente chofer del bus, y puede llegar hasta los 40 mil pesos (Transantiago, 2016).

No obstante, un reclamo por parte de un usuario no es sinónimo de una multa. Las multas solo se generan por parte de fiscalizadores cuando observan estas infracciones *in-situ*. Durante el 2015 se realizaron 1.200 fiscalizaciones, que resultaron en 138 conductores multados (Transantiago, 2016). Esta cifra, en comparación con los 10.145 reclamos que se recibieron el mismo año, parece insuficiente (DTPM, 2016).

Por esto, se plantea la necesidad de contar con una metodología que logre determinar la no detención de buses. Bull y Herrera (2015) proponen un modelo basado en información GPS que busca determinar la detención de buses en paraderos establecidos.

Se plantea utilizar tecnología GPS porque ya se encuentra presente en todos los buses, y no es necesario invertir recursos en ella. A pesar de los beneficios asociados a esta tecnología, existen ciertas limitaciones. La principal es que el dispositivo GPS entrega su ubicación cada 30 segundos, por lo que no se obtiene la trayectoria exacta de los buses y se debe reconstruir a partir de estos datos. A medida que el espaciamiento temporal entre datos GPS se reduce, la reconstrucción de la trayectoria se realiza con mayor precisión, facilitando la tarea de determinar la detención o no detención del bus. No obstante, una mayor frecuencia de datos implica un aumento en el consumo de recursos del dispositivo GPS y un costo para la empresa operadora. Además, sólo un cambio de contrato entre el Ministerio de Transporte y Telecomunicaciones y los operadores, como el establecido en la Resolución 286 (2012) les exigiría a los operadores aumentar esta frecuencia.

El modelo de Bull y Herrera (2015) está basado en cinemática y reconstrucción de trayectorias. Sin embargo, no ha sido validado, por lo tanto, la precisión del modelo es desconocida. Un análisis de este modelo es fundamental para generar una propuesta que entregue resultados con la mayor precisión posible.

1.1 Objetivos

Esta investigación tiene como objetivo principal definir un modelo que, en base a los datos reportados por dispositivos GPS cada 30 segundos, logre determinar la no detención de buses en paradas establecidas.

Los objetivos específicos son los siguientes:

- Validar el modelo propuesto por Bull y Herrera (2015) y proponer mejoras al mismo.
- Investigar el uso de métodos de aprendizaje supervisado para determinar la no detención de buses.
- Determinar la relación entre la frecuencia de datos GPS y la precisión de los modelos evaluados.

1.2 Contenidos

Esta tesis está dividida en cinco capítulos, partiendo por esta introducción. El segundo capítulo presenta el marco teórico, que incluye una revisión bibliográfica acerca del estado del arte de los temas a tratar en los capítulos siguientes. El tercer capítulo explica la metodología de esta investigación, la cual incluye la toma de datos y los pasos que se realizan para lograr los objetivos mencionados en la sección 1.1.

El cuarto capítulo muestra los resultados de la investigación. Se divide en cuatro partes, haciendo alusión a los objetivos de esta tesis. La primera parte valida el modelo de Bull y Herrera y realiza un análisis de sensibilidad respecto a los parámetros

involucrados. Además, en base a los resultados expuestos, se proponen cambios al modelo y se validan. La segunda parte presenta los modelos de aprendizaje supervisado, el proceso de calibración y validación de éstos. La tercera parte de este capítulo, hace una comparación entre los modelos formulados, para determinar el modelo a recomendar, , mientras que la cuarta parte hace un análisis respecto a la relación entre la frecuencia o periodicidad de los datos GPS, y la precisión de los modelos elaborados.

Por último, en el quinto capítulo se presentan las principales conclusiones e implicancias de esta investigación, así como también recomendaciones para investigaciones futuras.

2 MARCO TEÓRICO

Al realizar una revisión del estado del arte respecto a la no detención de buses en paraderos, la información que se encuentra es casi nula, dejando en evidencia que no se ha presentado esta problemática en otros países.

Al no encontrarse experiencia internacional acerca del tema, esta investigación se enmarca a los demás tópicos atingentes a ella. Durante este capítulo se abordarán estos temas, dividiéndolos en tres sub capítulos: Modelo de Bull y Herrera, Aprendizaje supervisado, e Indicadores y *test* estadísticos.

2.1 Modelo de Bull y Herrera (2015)

El modelo ideado por Bull y Herrera es el único acercamiento que se ha realizado hacia esta problemática. Sin embargo, su investigación queda enmarcada en la teoría, y no llega a realizarse una etapa de validación.

Comienza a gestionarse el 2014, a partir de la preocupación por parte del DTPM acerca del tema. Esta última, cuenta con los datos GPS de los buses, y son entregados cada 30 segundos. Los datos GPS consisten en: patente, línea operada, fecha, hora, y posición. En base a estos, los autores construyen su modelo.

Una consideración importante del modelo es que ningún bus que se haya detenido sea acusado de lo contrario. Para estar lo más seguro de esto, el modelo se formula de forma muy exigente. Esto quiere decir que cuando el modelo acusa a un bus de no detenerse, éste debe estar muy seguro de ello. El acusar erróneamente una no detención se considera Error Tipo I, y el dejar de acusar una será referido como Error Tipo II.

Debido a que los datos GPS son entregados cada cierto tiempo, la trayectoria del bus no es conocida en su totalidad, por lo que se debe realizar algún tipo de procesamiento para determinar si un bus se detuvo o no. No obstante, la reconstrucción

exacta de la trayectoria no es necesaria, ya que el modelo lo único que busca determinar es si el bus se detuvo entre un par de datos.

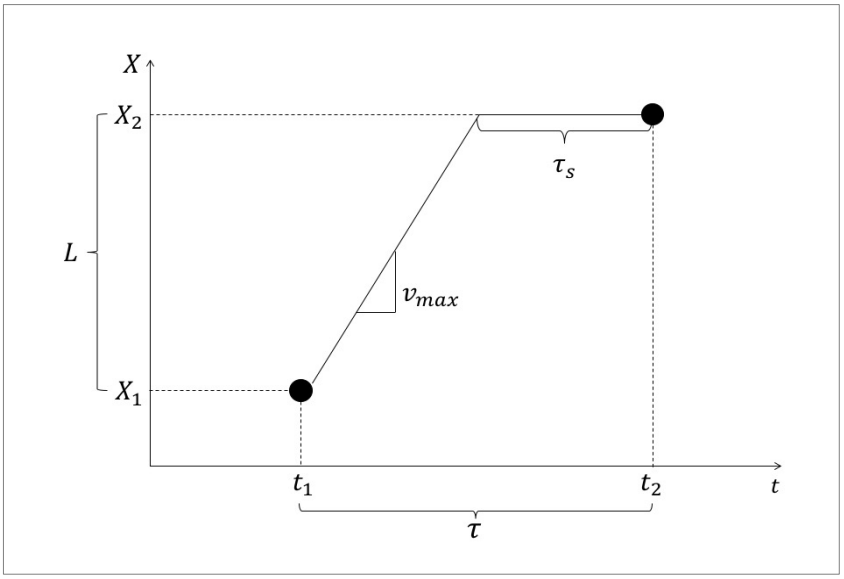

Figura 2.1 El Primer acercamiento a la formulación del modelo es determinar cuánto tiempo le toma al bus recorrer una distancia L a velocidad máxima, para así determinar el valor de τ_s. Fuente: Elaboración propia

Para realizar esto, los autores recurren a la cinemática, basándose en el modelo de Sun y Ban (2013). En la Figura 2.1 cada dato (también llamado *log* GPS) corresponde a los círculos negros. Se observa cómo un bus recorre una distancia L entre datos tomados cada τ segundos, y se muestra cuánto tiempo le hubiese tomado recorrer la misma distancia L, a velocidad máxima v_{max}. La diferencia entre τ y este tiempo, es expresada como τ_s.

El tiempo τ_s muestra la demora del bus. Un alto valor de τ_s sugiere que el bus circuló a una baja velocidad o se pudo haber detenido. De forma contraria, entre menor sea su valor, mayor es la probabilidad de que el bus no se haya detenido.

El modelo mide si durante el lapso de tiempo τ el bus alcanza a realizar una parada completa. Esto quiere decir que el bus alcance a desacelerar desde v_{max} a cero, y luego acelerar desde cero a v_{max}. Esta medida es muy exigente ya que no es necesario

que todo el proceso de aceleración y desaceleración ocurra durante τ. Además, no es obligatorio que el bus circule a v_{max}, exigiendo que el proceso de detención tome más tiempo. Este tipo de medidas apuntan a reducir al mínimo el Error Tipo I.

El modelo de Bull y Herrera asume una tasa de aceleración igual a la tasa de desaceleración, en valor absoluto, y se denomina a. No se especifica la razón de este supuesto, pero resulta válido al analizar el comportamiento de un bus.

Se define τ_{acc} como el tiempo necesario para acelerar desde una velocidad cero hasta v_{max} , es decir, $\tau_{acc} = v_{max}/a$. Durante τ_{acc} el bus logra avanzar una distancia L_a, cuyo valor se deriva de una de las ecuaciones básicas de cinemática, y se muestra a continuación.

$$L_a = \frac{\tau_{acc}^2 \cdot a}{2} = \frac{v_{max}^2}{2a} \tag{2.1}$$

Para hacer comparable el valor de τ_{acc} con τ_s, es necesario descontar el tiempo que le tomaría al bus recorrer esta distancia L_a a v_{max}. Este tiempo se denomina τ_l y se expresa en la ecuación (2.2).

$$\tau_l = \frac{L_a}{v_{max}} = \frac{v_{max}}{2a} \tag{2.2}$$

Se denomina τ_a al tiempo de demora asociado a una aceleración. Este valor se presenta en la ecuación (2.3) y es esencial en el modelo, ya que todas las restricciones se hacen respecto a él. La Figura 2.2 busca representar de forma gráfica este valor, para su mejor comprensión.

$$\tau_a = \tau_{acc} - \tau_l = \frac{v_{max}}{2a} \tag{2.3}$$

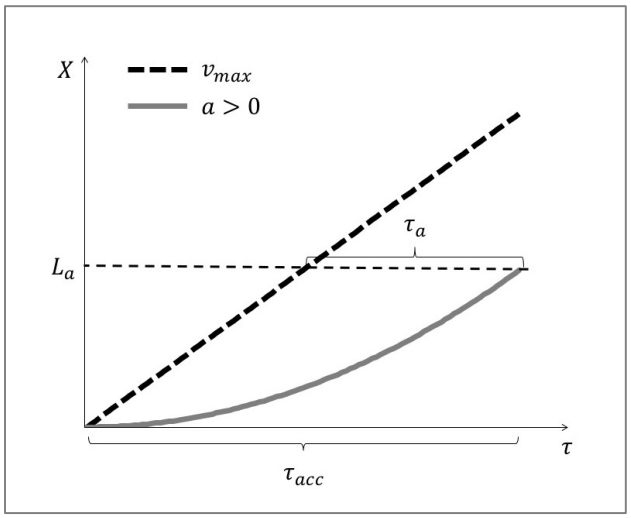

Figura 2.2 La diferencia horizontal entre la trayectoria a velocidad máxima y un proceso de aceleración hasta velocidad máxima, determina el tiempo extra que toma este proceso, definiéndolo como τ_a.
Fuente: Elaboración propia

La línea continua representa la trayectoria de un bus en un proceso de aceleración, donde comienza con una velocidad cero y termina a v_{max}. La línea punteada representa la trayectoria que recorrería el bus a v_{max}. La diferencia horizontal entre estas curvas, muestra el tiempo extra que está asociado a un proceso de aceleración. Esto es análogo al proceso de desaceleración.

Este modelo le asigna a cada paradero un área de estacionamiento, reconociendo que un paradero tiene un largo físico y que no todos los buses se detienen exactamente en frente al paradero. El largo de esta área de estacionamiento dependerá de cada paradero. Un bus recorriendo esta área puede o no mandar su ubicación. Un bus que transita más despacio es más probable que envíe su *log* GPS, ya que le toma más tiempo atravesar esta área.

El área de estacionamiento se denota por SA (por su nombre en inglés, *Stop Area*). Sus límites aguas arriba y aguas abajo se denotan por X_{SA}^{U} y X_{SA}^{D} respectivamente. El modelo selecciona los datos GPS fuera del SA pero más cercanos a ella, tanto aguas arriba como aguas abajo. La ubicación de estos datos se denota por X_U y X_D.

Además de estos dos datos se utilizan todos aquellos datos que se encuentren dentro del área de estacionamiento. Dependiendo de la cantidad de ellos, los autores dividen los casos posibles en tres: no hay *log* GPS en SA, un *log* GPS en SA, y más de un *log* GPS en SA. Cada uno de estos casos se comporta de forma distinta, y los autores generan restricciones para cada uno de ellos. La Figura 2.3 muestra los tres posibles casos existentes. El área celeste representa SA, y los círculos negros representan los datos o *logs* GPS.

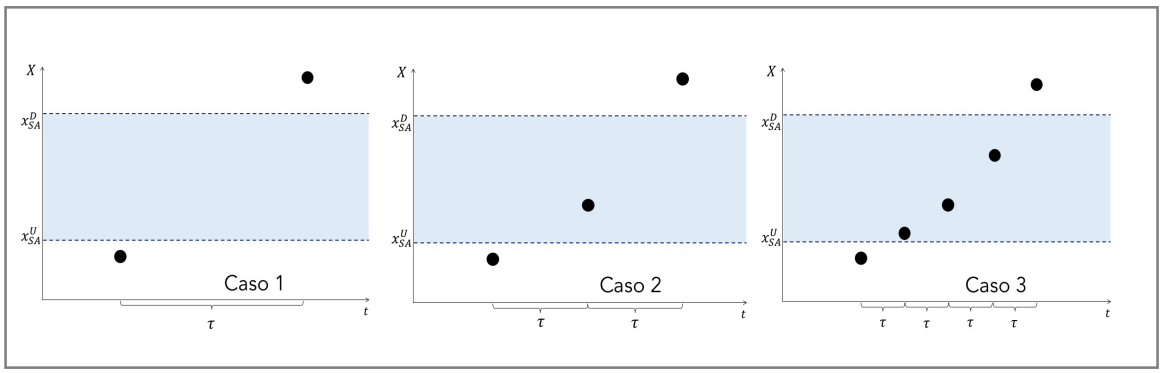

Figura 2.3 Siendo la franja celeste el área de estacionamiento, los Casos 1, 2, y 3 representan: No hay log GPS en SA, hay un log GPS en SA, o más de un log GPS en SA.
Fuente: Elaboración propia

A continuación, se explica la metodología a seguir para el Caso 1. Los demás casos siguen la misma lógica, y para mayor detalle se recomienda ir a la fuente original (Bull y Herrera, 2015).

Caso 1: No hay log GPS en el área de estacionamiento

Debido al análisis anterior, queda claro que si $\tau_s \geq 2\tau_a$, el bus se pudo haber detenido en el área de estacionamiento. Aunque existen otras razones del por qué el bus se pudo haber demorado, no es posible determinar que éste no se detuvo. La formulación que se realiza en este subcapítulo, consiste en relajar esta restricción, y así determinar el mayor valor de τ_s en el que todavía se puede concluir que el bus no se detuvo.

Para realizar esto, los autores relajan el supuesto $\tau_s \geq 2\tau_a$, dando pie a la posibilidad que la detención no ocurra de forma completa entre un par de datos GPS. Sin embargo, esto se debe realizar de tal forma que el Error Tipo I no aumente.

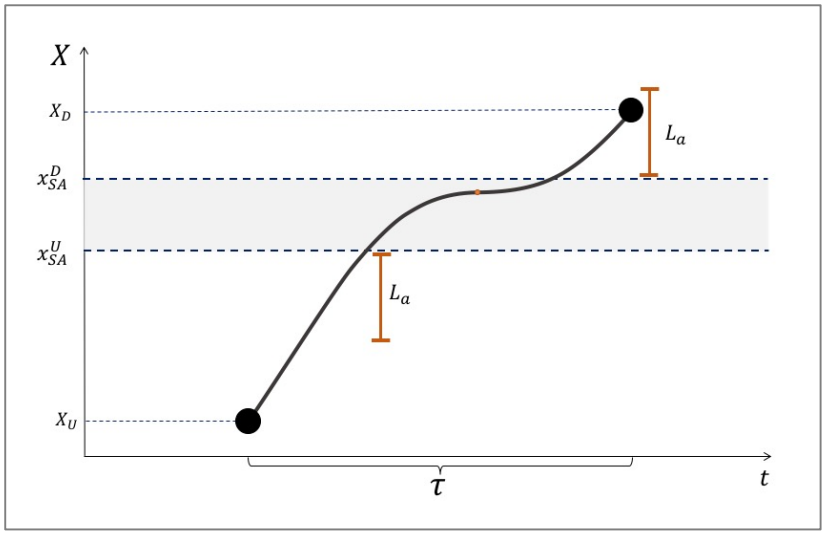

Figura 2.4 Ejemplo - Caso 1 presenta la relajación respecto al supuesto inicial, donde el log GPS X_D sugiere que el proceso de aceleración no ocurre por completo.
Fuente:Elaboración propia

En la Figura 2.4, X_D se encuentra cerca de X_{SA}^D, lo que sugiere que el proceso de aceleración no alcanza a realizarse por completo. Por tanto, los autores sugieren que solo una fracción α del proceso ocurre en este tramo. En consecuencia, el tiempo asociado a esta fracción del proceso de aceleración corresponde a $\alpha\tau_a$.

Durante un proceso de aceleración completo se recorre una distancia L_a. Esto sugiere que si $X_D - X_{SA}^D < L_a$, solo una fracción del proceso de aceleración ocurrió en el tramo. Mientras que si $X_D - X_{SA}^D \geq L_a \implies \alpha = 1$. Esto mismo se puede realizar para el proceso de desaceleración, cuando X_U se encuentra muy cercano a X_{SA}^U. Para realizar la restricción, el modelo elige la menor de estas tres distancias.

$$\Delta = \min\{X_D - X_{SA}^D;\ X_{SA}^U - X_U;\ L_a\} \tag{2.5}$$

Para mantener al margen el Error Tipo I se considera que solo uno de los procesos, entre la aceleración y la desaceleración, no ocurra de forma completa. Existe el caso en que ninguno de estos procesos ocurra de forma completa, pero no se considera para no relajar aún más la hipótesis inicial.

Para relacionar el largo Δ con τ_s, se debe obtener el tiempo que toma recorrer este largo al desacelerar o acelerar a una tasa constante a. A éste lo llamaremos τ'_{acc}, y se obtiene sustituyendo el valor de L_a por Δ en la ecuación 2.1.

$$\tau'_{acc} = \sqrt{2\Delta/a} \tag{2.6}$$

Para determinar el valor del tiempo extra que produce el proceso de aceleración o desaceleración al recorrer Δ, se sustituye el valor de de L_a por Δ en la ecuación 2.2, a este valor se le denomina τ'_l.

$$\tau'_l = \Delta/v_{max} \tag{2.7}$$

Siendo la diferencia entre estos valores el valor del tiempo extra, aportado por una fracción del proceso de aceleración o desaceleración. Es posible concluir que el bus no se detuvo en el paradero si:

$$\tau_s < \sqrt{2\Delta/a} - \Delta/v_{max} + \tau_a \tag{2.8}$$

Caso 2 y Caso 3, uno o más log GPS en el área de estacionamiento

Los Casos 2 y 3 siguen la misma lógica de este primer caso, donde se realizan ecuaciones respecto a τ_s para determinar si el bus se pudo haber detenido. Estas ecuaciones y formulaciones son más complejas porque incluyen la información de los log GPS dentro del área de estacionamiento. Las restricciones de los Casos 2 y 3 buscan determinar si el bus se pudo haber detenido entre tres datos GPS.

Si existen n datos GPS dentro de SA, donde la ubicación de cada uno de ellos se denota por X_i, donde i toma el valor desde 1 hasta n. En este caso el modelo formula restricciones adicionales para ver si la detención ocurre dentro o fuera de SA. Para esto

se introducen los términos γ_U y γ_D, que son la fracción máxima del proceso de detención contenido en los tramos τ_{U1} y τ_{1D} correspondientemente.

Como las restricciones están formuladas para tres logs GPS, en el Caso 3 los autores diferencian los casos donde el trio a analizar contienen un log fuera del área de estacionamiento, al caso en el que el trio de log se encuentre dentro del área de estacionamiento. Estas restricciones se presentan en la Tabla 2.1.

Tabla 2.1 Resumen de restricciones Modelo Bull y Herrera

Fuente: Elaboración propia a partir de Bull y Herrera (2015)

CASO	RESTRICCIONES
CASO 1	$\Delta = \min\{X_D - X_{SA}^D;\ X_{SA}^U - X_U;\ L_a\}$ $\tau_s < \sqrt{2\Delta/a} - \Delta/v_{max} + \tau_a$ (1)
CASO 2	***Si*** $\boldsymbol{\tau_{sUD} \geq 2\tau_a}$ $min\{(2\tau_a)\gamma_D;\ \tau_{s1D}\} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a$ (2) ***Si*** $\boldsymbol{\tau_{sUD} < 2\tau_a}$ $\tau_{sU1} < \sqrt{\frac{2(X_{SA}^U - X_U)}{a}} - \frac{(X_{SA}^U - X_U)}{v_{max}} + \tau_a$ (3) $\tau_{s1D} < \sqrt{\frac{2(X_D - X_{SA}^D)}{a}} - \frac{(X_D - X_{SA}^D)}{v_{max}} + \tau_a$ (4)
CASO 3	**Dos log GPS dentro del área de estacionamiento y uno fuera** $\tau_{sU1} < \sqrt{\frac{2(X_{SA}^U - X_U)}{a}} - \frac{(X_{SA}^U - X_U)}{v_{max}} + \tau_a \ \wedge\ \tau_{s12} + \min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a$ (5) $\tau_{snD} < \sqrt{\frac{2(X_n - X_{SA}^D)}{a}} - \frac{(X_n - X_{SA}^D)}{v_{max}} + \tau_a \ \wedge\ \tau_{s(n-1)n} + min\{(2\tau_a)\gamma_D;\ \tau_{snD}\} < 2\tau_a$ (6) **Tres log GPS dentro del área de estacionamiento** $\tau_{s(i-1)i} + \tau_{si(i+1)} < 2\tau_a$ (7)

Si se requiere mayor información acerca de cómo se formulan estas restricciones, se recomienda ir a la fuente directa Bull y Herrera (2015).

2.2 Métodos de aprendizaje supervisado

El aprendizaje supervisado consiste en generar modelos de predicción o clasificación en base a un conjunto de datos de entrenamiento, estos deben contar tanto con una variable objetivo como con variables predictivas. La variable objetivo es el resultado de la clasificación o predicción que se busca realizar, en el marco de esta investigación es la detención o no detención del bus en un paradero. Mientras que las variables predictivas son un conjunto de características asociadas a este dato u objeto.

En base a estos datos se entrena y construye el modelo de predicción, que logra determinar qué características están asociadas a los distintos resultados de la variable objetivo. De esta forma, al enfrentarse a un conjunto de datos no visto, se puede predecir su resultado. Se le llama aprendizaje supervisado por la presencia de la variable objetivo para guiar el proceso de clasificación (Friedman et al, 2001).

Las variables predictivas tienden a tener rangos que varían considerablemente entre ellas. Esta diferencia afecta la predicción en ciertos métodos. Para algunos algoritmos, estas diferencias pueden generar una tendencia a priorizar las variables con mayor rango, generando un efecto no deseado en los resultados. Por esta razón es importante estandarizar los valores de las variables a utilizar.

Existen varios algoritmos dentro de la clasificación de datos. En esta investigación se utilizan dos de ellos: Máquinas de vectores de soporte (SVM) y Análisis discriminante lineal (LDA). Se eligen estos debido a que son simples, rápidos de procesar e implementar, y presentan buenos resultado, esto último los hace muy utilizados en distintos rubros. A continuación, se revisarán en detalle cada uno de ellos.

2.2.1 Máquinas de vectores de soporte (SVM)

Las SVM se originan por Vapnik y Chervonenkis en 1964, pero es durante los años 90’ que se desarrollan a la forma en que se utilizan hoy en día (Boser et al., 1992, Cortes y Vapnik, 1995). Aunque comenzaron como un método propuesto para

clasificación binaria, hoy se utilizan en regresión, agrupación, multi-clasificación, y en clasificación de proteínas.

Las SVM buscan estimar hiperplanos o separadores lineales de forma que cada categoría quede lo más separada posible del resto. Esto se puede realizar en el espacio original donde se encuentran los datos, para casos separables o cuasi-separables. En caso que los datos no sean separables linealmente, se usa un término de penalización que castiga los ejemplos o datos que no pudieron ser separados de forma lineal. Lo que se hace en este caso es transformar el espacio de entradas a un espacio de mayor dimensionalidad, de tal forma de tener un espacio donde los datos sean mejor separados por un hiperplano. La búsqueda de estos separadores lineales en un espacio transformado se realiza con las funciones Kernel. En la Figura 2.5, se puede observar la aplicación de esta función, donde el espacio de características generado por la función Kernel es representado por Φ.

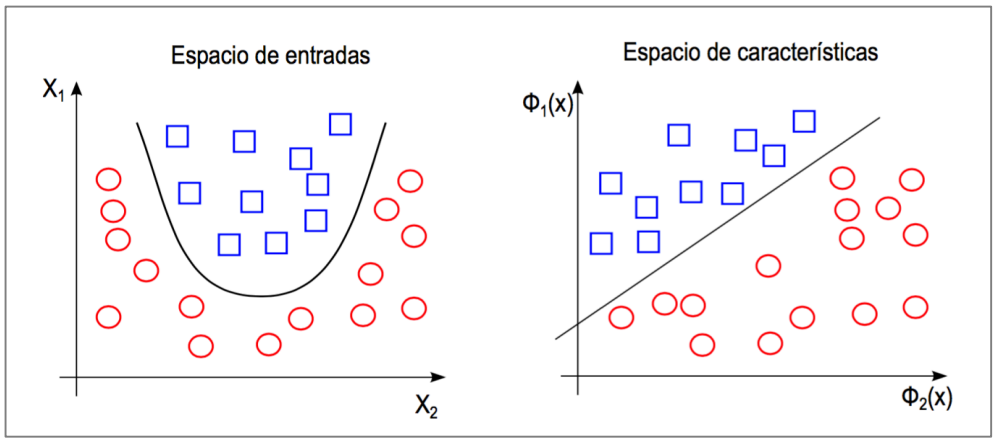

Figura 2.5 Representación gráfica de una transformación a espacio de características donde las clases son separables linealmente.
Fuente: Carmona (2014)

Existen distintas funciones Kernel, las más frecuentemente usadas son: lineales, polinomiales y gaussianas. Este se elige generalmente por el dominio y problema, ya que no hay funciones Kernel mejores que otras, si no que estas funcionan de mejor o peor manear de acuerdo al número de variables, datos, y separabilidad de las clases. Shawe-

Taylor y Christianini (2004) explican cómo son utilizadas cada una de éstas en diferentes aplicaciones. Los modelos de clasificación de datos, en general, buscan minimizar el riesgo empírico, mientras que las SVM buscan minimizar el riesgo estructural empírico, buscando el modelo más simple. Esto se realiza induciendo un hiperplano equidistante a los ejemplos más cercanos de cada clase, de forma de maximizar el margen entre los ejemplos y el hiperplano. Los datos que caen justo en la frontera del margen, son los vectores de soporte. Este margen ayuda a tener una buena capacidad de generalización, evitando el sobreajuste. La Figura 2.6 grafica este proceso.

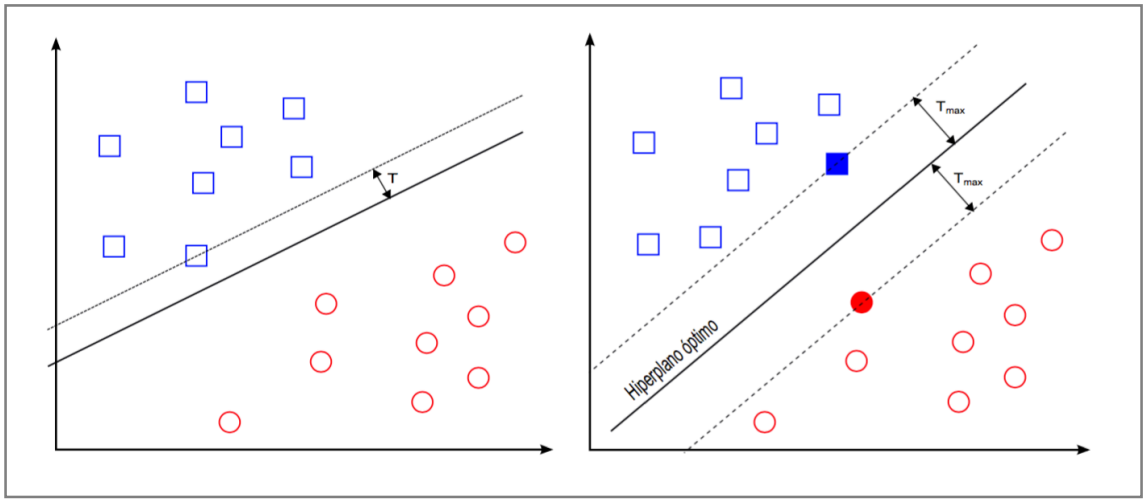

Figura 2.6 Se muestra un mismo set de datos separados por dos hiperplanos distintos, el de la derecha logra maximizar el margen entre los datos de las distintas categorías, siendo Tmax el margen óptimo
Fuente: Carmona (2014)

En el caso de que este margen no se cumpla, existe un costo asociado en la función objetivo. Este costo está asociado a una variable de holgura, relacionada con la distancia hasta el margen, y una constante C. A mayor valor de esta constante, mayor peso se le otorga a este error, mejorando la clasificación. Sin embargo, un aumento en el valor de C tiene consecuencias en el sobreajuste, perdiendo generalidad al momento de predecir. Para una descripción más profunda, junto al planteamiento matemático, se recomienda revisar Smola & Schölkopf (1998).

2.2.2 Análisis discriminante lineal (LDA)

El LDA es un método simple, que se puede utilizar como un clasificador lineal o para reducir la dimensionalidad de una matriz o set de datos. No tiene hiperparámetros que se deban calibrar o configurar, lo que facilita su implementación.

Este modelo trabaja con variables independientes, y a partir de éstas busca una combinación lineal que caracterice las clases. Izenman (2013) define estas nuevas variables como *class label*, que significa 'etiqueta de clase', ya que son estas las que buscan caracterizar las distintas clases mediante una combinación lineal de las variables predictivas.

Las *class labels* son creadas en función de maximizar la distancia de separación entre clases y así evitar el sobreajuste. De esta forma al presentar una observación sin etiqueta, el modelo predice la clase a la que pertenece mediante las *class labels*. La exactitud de esta predicción dependerá de la información provista en los datos de entrenamiento y de la naturaleza del problema. Cuando las variables predictivas logran separar o caracterizar de forma correcta cada una de las clases, se refleja en la combinación lineal de estas.

Cuando existen dos clases distintas, se necesita sólo un *class label*, mientras que si existen más de dos clases se requieren por lo menos dos, y se puede generar un máximo de $K - 1$ clasificadores, siendo K el número de clases.

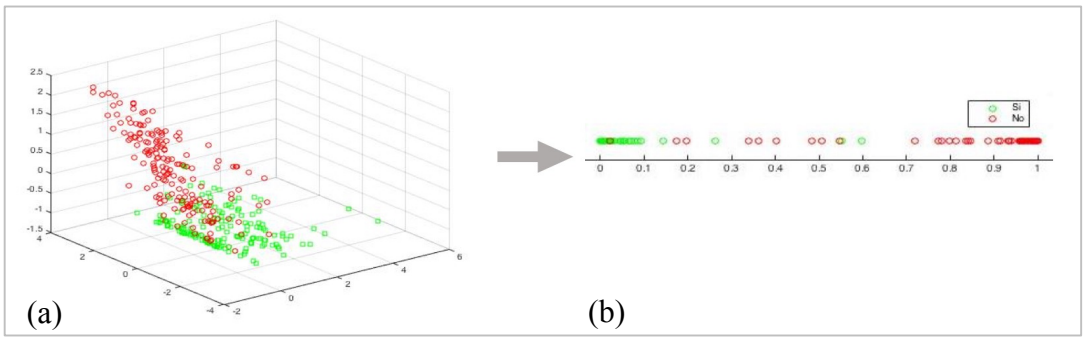

Figura 2.7 La imagen (a) presenta dos clases de datos respecto a tres variables predictivas. Mientras que la imagen (b) presenta un class label construido con un modelo LDa respecto a estos mismos datos.

Fuente: Elaboración propia

Para esto, utiliza la regla de Bayes donde busca minimizar el cociente entre la varianza de datos de diferentes clases sobre la varianza de datos de la misma clase. Para mayor información de cómo funciona el análisis de discriminante lineal, se recomienda dirigirse a Izenman (2013).

2.3 Indicadores y test estadísticos

Para identificar las limitaciones y ventajas que tiene cada uno de los modelos, se utilizan un par de indicadores y test estadísticos. Se presenta a continuación una breve introducción a cada uno de ellos.

2.3.1 Validación cruzada

La validación cruzada es un método que se utiliza en el aprendizaje de máquinas para compensar la tasa de error optimista que puede resultar de la varianza de los datos (Cawley, 2006). Los métodos de clasificación a utilizar funcionan con dos etapas, donde con una porción de los datos los métodos realizan sus modelos de clasificación y con el resto de los datos se realiza una validación.

La validación cruzada divide la muestra en *n* secciones, donde las máquinas de aprendizaje realizan la primera etapa con *n-1* secciones, y la segunda etapa con la sección restante. Este proceso se itera *n* veces, cambiando la sección con la que se realiza la validación. Así, evita el sesgo que se produce por la muestra que se selecciona para validar, ya que se valida con toda la muestra en algún punto.

A partir de este proceso, el indicador de precisión de la validación, resulta ser un promedio de todas las validaciones. Así, se pueden obtener otros indicadores, que no resultan ser de una validación en particular, pero de un proceso de validación, como un rango de precisión y la desviación estándar de esta.

2.3.2 Matriz de confusión

La matriz de confusión permite observar el desempeño de un modelo de precisión. Se divide en filas y columnas, donde las columnas representan las predicciones hechas por el algoritmo, mientras que las filas representan las instancias reales de las clases. Facilita ver donde el modelo se está confundiendo, o qué clase se determina de forma correcta.

Si existen dos clases, los positivos (P) y los negativos (N), la matriz de confusión mostrará cuatro celdas con datos. La Figura 2.8 muestra una matriz de confusión.

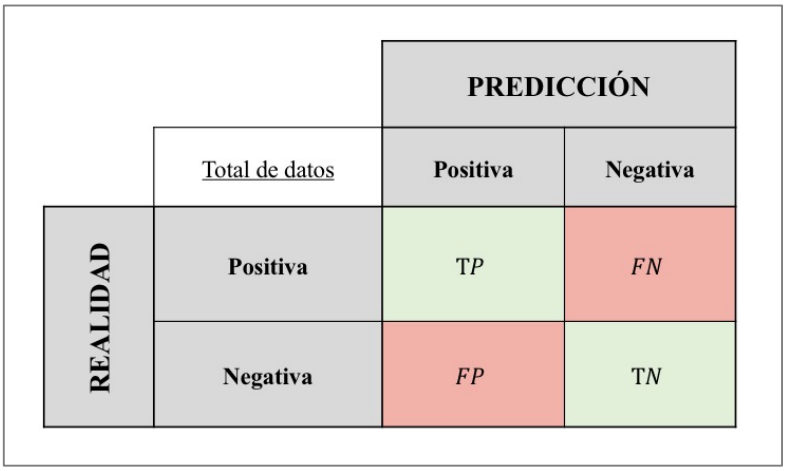

Figura 2.8 La matriz de confusión presenta los resultados de una clasificación bianria, separando los resultados en cuatro categorías.
Fuente: Elaboración propia

En la Figura 2.8, las celdas de colores representan los cuatro casos posibles. Las celdas verdes representan los datos que debiesen tener mayores valores cuando el modelo funciona de forma adecuada, mientras que las celdas rojas representan datos no deseados o errores que comete el modelo. A continuación, se presentan cada uno de estos casos.

Verdaderos positivos (TP): Datos pertenecientes a la clase positiva, que fueron clasificados correctamente como pertenecientes a esta clase.

Verdaderos negativos (TN): Datos pertenecientes a la clase negativa que fueron clasificados de forma correcta.

Falsos positivos (FP): Datos que pertenecen a la clase negativa, pero el algoritmo los clasificó como clase positiva, estos también son conocidos como el error tipo I.

Falsos negativos (FN): Datos positivos que el algoritmo definió de forma errónea como negativos, estos son lo que se conoce como error tipo II.

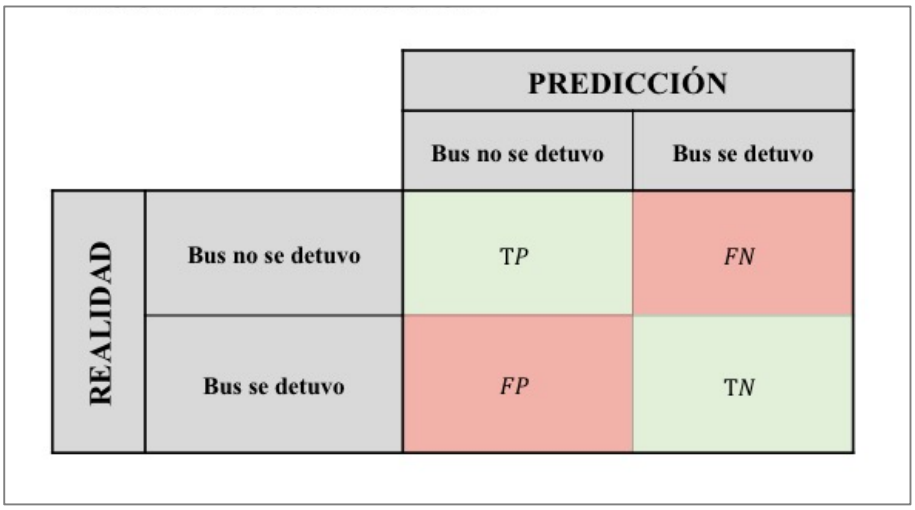

Figura 2.9 Matriz de confusión que represnta la clasificación que se realiza en esta investigación, clasificando el Error Tipo I como los buses que se detienen y el modelo predice que no se detuvo. Fuente: Elaboración propia

Esta matriz facilita cuantificar los siguientes indicadores estadísticos, que se utilizarán para comparar los algoritmos y modelos a utilizar.

Tasa de falsos positivos (FPR): Es la proporción de casos negativos que el algoritmo detecta como positivos. Indicador que muestra la tasa de error tipo I respecto a su clase.

$$\frac{FP}{FP + TN}$$

Tasa de falsos negativos (FNR): Es la proporción de casos positivos que el algoritmo detecta como negativos. Indicador que muestra la tasa de error tipo II respecto a su clase.

$$\frac{FN}{FN + TP}$$

Precisión: Es la proporción de los casos detectados de forma correcta.

$$\frac{TP + TN}{P + N}$$

3 METODOLOGÍA

En esta investigación se utilizan datos reales para la calibración y evaluación de los modelos propuestos. En este capítulo se presentan primeramente los datos recolectados y la preparación de éstos para ser utilizados en los modelos. Posteriormente se describe cómo se abordará cada uno de los objetivos específicos de esta investigación.

3.1 Recolección de datos

Esta investigación tiene dos tipos de datos asociados: las observaciones y los log GPS. Una observación corresponde al paso de un bus por un paradero, estas pueden corresponder a una detención o a una no detención. Cada observación tiene asociado una serie de log GPS de ese bus en las inmediaciones del paradero.

Los log GPS, pueden provenir directamente de los buses de Transantiago, a través del DTPM, o de un GPS particular. Esto logra separar el conjunto de datos en dos, los que se denominan Base de datos 1 y Base de datos 2 respectivamente.

Cabe destacar que en ambos casos fue necesario un pre procesamiento de los datos a fin de ajustar las posiciones obtenidas mediante el GPS a la ruta real del recorrido. Este proceso se conoce como *map-matching* y se describe en más detalle en el Anexo 3.1. A continuación se explica cómo se generaron las dos bases de datos.

3.1.1 Base de datos 1 – Buses de Transantiago

Estos datos GPS son con los que trabaja la DTPM, y provienen directamente de los buses que opera Transantiago. La información que entregan los dispositivos GPS incluye las variables que se muestran en la Tabla 3.1.

Tabla 3.1 Valores que entregan los datos GPS

Fuente: Elaboración propia

VARIABLES	DESCRIPCIÓN
Patente	La patente del bus. Funciona como identificación
Longitud	Representa la ubicación longitudinal respecto al meridiano de Greenwich
Latitud	Representa la ubicación latitudinal respecto al Ecuador
Tiempo GPS	Tiene la información de: año, mes, día, hora, minuto, y segundo, del dato GPS
Ruta	Servicio al cual pertenece, con su dirección correspondiente, ida o regreso
Velocidad	Velocidad instantánea con la que circula el bus
Ignición	Se refiere si el motor del bus se encuentra operando o está detenido

Las observaciones a las que corresponden estos datos se realizan en terreno, determinando si el bus se detiene o no en el paradero. Primero se decidió el lugar donde se realizó la medición. Se decide que el mejor lugar para desarrollar este experimento es un corredor dado su alto flujo de buses, lo que entrega una mayor cantidad de datos en una menor ventana de tiempo. Además, existen muchos recorridos *Expreso*, por lo que no todos los buses que circulan por el corredor se detienen en el paradero a escoger. De esta forma, es posible observar fácilmente tanto detenciones como no detenciones.

Se consideró como detención todo bus que llega a una velocidad cero en el área considerada por el paradero. Esto incluye a los buses que se detuvieron debido a la presencia del semáforo ubicado 5 metros aguas abajo del paradero, a pesar que ningún usuario se bajó o subió del bus.

Un lugar que cumple las condiciones antes expuestas es el corredor de Santa Rosa, al sur de su intersección con Américo Vespucio. Las observaciones se realizan durante la punta mañana ya que el flujo de buses es mayor. Finalmente, se establece que el paradero a observar es el PE113, que es un paradero donde no se detienen servicios expresos. En este paradero se pueden observar 14 servicios de Transantiago, de los

cuales 5 son expresos. Durante la punta mañana estos recorridos suman una frecuencia agregada de 120 buses/hora.

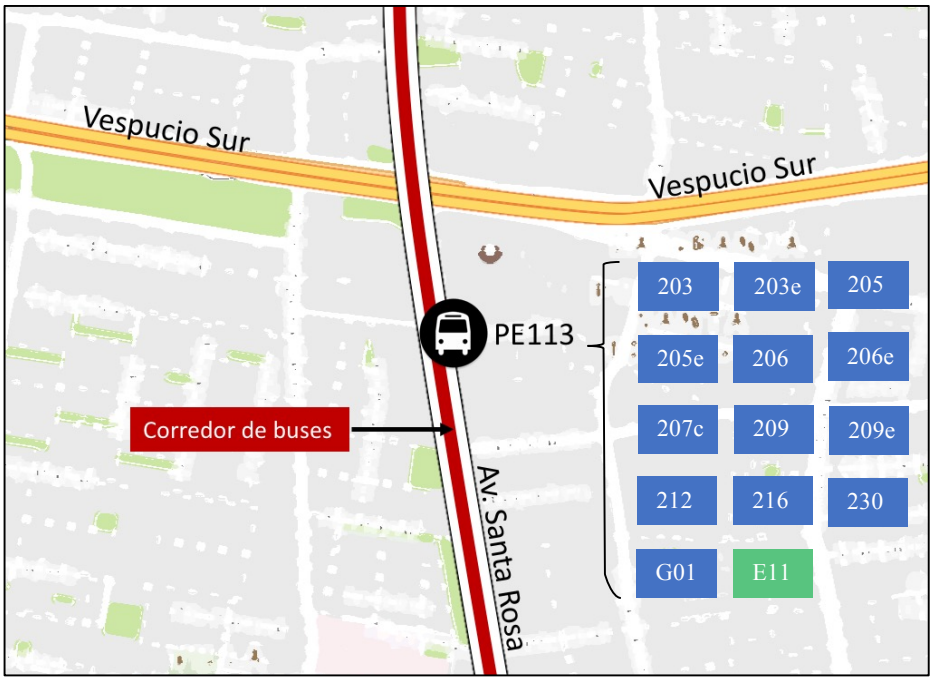

Figura 3.1 Diagrama de ubicación del paradero PE113 en A.v Santa Rosa al sur de Alborada, y los servicios que se observan desde el paradero.
Fuente: Elaboración propia via Google Maps

Durante el periodo observado, se utiliza una cámara de video para grabar lo que ocurre en el paradero. Esto es útil para verificar patentes y hora en que los buses circulan. Sin embargo, como es un corredor cuyo flujo de buses es muy alto, no se logra capturar las patentes de todos los buses ya que algunos obstruyen la visibilidad de otros o la velocidad a la que circulan hace indetectable su patente para la cámara. No obstante, al generar la grabación se utilizó también el audio del video para generar registro acerca de los buses que circulaban. El observador presente logra capturar verbalmente patente, hora, y detención y no detención de los buses. Se priorizó los buses que no se detienen, ya que estos son los más difíciles de capturar por la cámara. Los datos se tomaron en dos instancias, el día 9 de septiembre y 10 de octubre de 2016.

La primera muestra se realizó desde las 6:50 hasta las 9:00 horas para abarcar el mayor tramo posible de buses, y se observaron 262 buses pasar, de los cuales 124 se detuvieron y 138 no se detuvieron en el paradero. De estos buses, existen 2 que no se logra obtener su recorrido GPS, ya que aparecen en el sistema como buses en tránsito, y su ubicación geográfica no se conoce para la hora deseada. Esto deja un total de 260 observaciones, 124 detenciones, y 136 no detenciones.

La segunda muestra se realizó desde las 7:30 hasta las 8:30 horas debido a que solo se podía acceder a una hora de información, y este es el horario con mayor flujo de buses. Se observaron 115 buses, de estos existen cuatro para los que no se logra obtener su recorrido GPS, ya que aparecen en el sistema como buses en tránsito. Esto deja un total de 111 observaciones, 45 detenciones, y 66 no detenciones. Las dos muestras en su conjunto suman 371 datos, donde 169 son detenciones y 202 corresponden a no detenciones.

3.1.2 Base de datos 2 – Datos particulares

Uno de los objetivos de esta tesis es estudiar la relación entre la frecuencia con que se reciben datos GPS y la precisión para predecir la no detención de buses. Por esto, se debe realizar un análisis con datos que permita hacer una variación en la frecuencia.

Se utilizó un dispositivo GPS que entrega longitud, latitud, hora y velocidad instantánea, cada un segundo. Esta periodicidad permite analizar todas las posibles frecuencias de datos. Además, hace posible generar distintos sets de datos para una misma observación, seleccionando datos a un intervalo de tiempo de τ. La Figura 3.2 grafica esta situación.

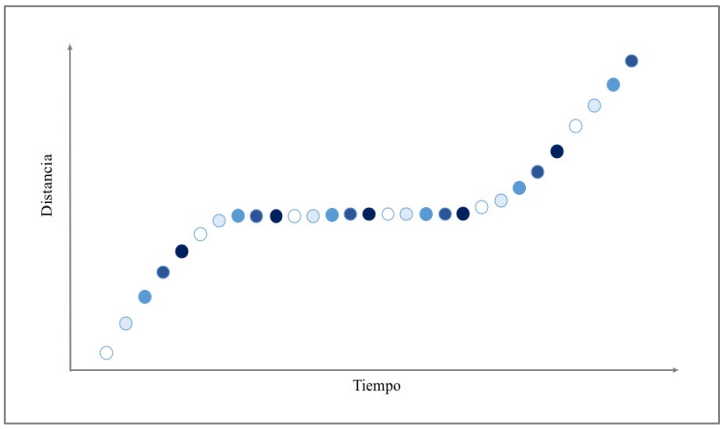

Figura 3.2 Se pueden generar distintos sets de datos a partir de una muestra, seleccionando datos a una periodicidad determinada como los presentados por las distintas tonalidades de azul.
Fuente: Elaboración propia

La Figura 3.2 muestra un set de datos con una de periodicidad de un segundo por dato, donde cada dato es representado por un círculo. Si utilizamos la situación hipotética que el τ a evaluar es 5 segundos, con estos datos podemos generar cinco sets distintos. Los círculos más oscuros, representan a los datos de los segundos 5, 10, 15, 20, y 25. Seleccionando solo estos datos se genera un set de datos con un espaciamiento de 5 segundos. Los otros cuatro sets de datos, se pueden generar de forma análoga seleccionando los círculos de las otras tonalidades. Este proceso hace posible generar cinco clasificaciones o predicciones distintas, en base a una sola observación. Este proceso es equivalente para cualquier valor de τ, donde se generan τ sets de datos.

El dispositivo GPS se utilizó en un servicio de Transantiago, y mientras se utilizaba se anotaron los paraderos en los que se detuvo y en los que no se detuvo el bus. Se consideró como detención el que el bus llegara a una velocidad cero en el área considerada por el paradero. Para tener una mayor cantidad de observaciones de 'No Detenciones', se utilizaron paraderos ficticios. Estos son paraderos que no corresponden al servicio, pero sí se encuentran en su ruta. De hecho, generalmente se encuentran cerca de un paradero 'real' del servicio, por lo que la detención en ese paradero podría afectar la predicción de lo ocurrido en el paradero ficticio. Se tienen muestras a distintas horas del día, tanto para punta mañana, punta tarde y fuera de punta.

Tabla 3.2 Resumen de observaciones obtenidos con datos GPS propios

Fuente: Elaboración propia

	PARADEROS RECORRIDOS	NO DETENCIONES	DETENCIONES	Nº DATOS GPS
Muestra 1	43	22	21	2412
Muestra 2	42	22	20	2230
Muestra 3	27	11	16	1654
Muestra 4	25	20	5	1433
Muestra 5	23	8	15	1576
Muestra 6	24	15	9	1508
TOTAL	**184**	**98**	**86**	**10813**

Todas las muestras de datos fueron recogidas al utilizar el servicio 426 de Transantiago. Se escoge este servicio ya que es un recorrido largo que permite muchas observaciones en una muestra. Como se observa en la Tabla 3.2, existen muestras de distinto largo dependiendo del tramo recorrido. En total, las seis muestras tomadas observan 184 paraderos, donde el bus se detuvo 86 veces. Para información más detallada acerca de las muestras tomadas se puede acceder al Anexo computacional DatosBaseDeDatos1.xls.

La limitación de estos datos es que no son los mismos con los que trabaja Transantiago, por lo que pueden existir diferencias en la toma de datos o satélites que utilizan los GPS.

3.2 Análisis de modelos basados en cinemática

Esta sección se divide en cuatro partes. La primera parte presenta la validación del modelo de Bull y Herrera (2015), donde se explica la programación del modelo, su construcción y funcionamiento. La segunda parte explica cómo se realiza el proceso de

proponer e implementar cambios a la metodología. La tercera explica la comparación de modelos. La cuarta parte calibra los parámetros, además de detallar y explicar los distintos valores que se utilizan en el análisis de sensibilidad para cada uno de ellos.

3.2.1 Validación Bull y Herrera

Primeramente, se definen los valores de los parámetros del modelo, v_{max} y a. Para esto se utilizan los datos provenientes de la Base de Datos 2 debido a su menor periodicidad.

En el caso de $\boldsymbol{v_{max}}$, en un principio se asume como 60 km/h, ya que es el límite de velocidad en la zona urbana del país. Para validar este supuesto se toman las velocidades instantáneas de la Base de datos 2. Estos datos se pueden observar en el anexo computacional BaseDatos2.csv Los tres valores máximos de la velocidad instantánea se presentan en la Tabla 3.3.

Tabla 3.3 Velocidad máxima – Servicio AA2001

Fuente: Elaboración propia

1º	$\boldsymbol{61,65\ m/s}$
2º	$60{,}81\ m/s$
3º	$59{,}99\ m/s$

La Tabla 3.3 valida la decisión de elegir una velocidad máxima igual a 60 km/h, dado que la mayoría de los datos se encuentran bajo esa velocidad.

La aceleración, $\boldsymbol{a}$, es una variable que se encuentra en la mayoría de las restricciones del modelo. Además, muchas de las otras variables se calculan respecto a ella. Para calibrarla, se utilizó nuevamente la Base de Datos 2. Es decir, los datos recolectados con un dispositivo GPS propio dentro de un servicio Transantiago se utilizaron para obtener la aceleración y desaceleración promedio del bus. Estos datos se

pueden observar en el anexo computacional Servicio-AA-2001.xls. Un resumen de los valores promedio se presenta en la Tabla 3.4.

Tabla 3.4 Aceleración y desaceleración promedio, junto al promedio en valor absoluto del Servicio AA2001
Fuente: Elaboración propia

Aceleración	$\mathbf{0,93\ m/s^2}$
Desaceleración	$-1{,}16\ m/s^2$
Promedio Valor Absoluto	$\mathbf{1,04\ m/s^2}$

Considerando que el modelo asume una misma tasa de aceleración y desaceleración (en magnitud), se decide utilizar $a = 1\ m/s^2$.

Con los valores anteriores, la validación del modelo se realiza con los datos provenientes de la Base de datos 1. Una vez realizada la validación del modelo, se procede a hacer un análisis de los resultados, donde se utilizan los valores de precisión, FPR, y FNR. Se observan los resultados para cada una de las observaciones, tratando de hacer una distinción donde el modelo falla.

Este modelo se realiza en C# mediante Visual Studio. El programa está pensado para recibir las observaciones y corroborarlos mediante los datos GPS. Como dato de entrada, el programa recibe las observaciones. Esta lista se presenta como un archivo de texto, donde cada línea contiene: Código de servicio, dirección, patente, hora, fecha, y paradero.

El programa selecciona los datos GPS de cada observación abarcando un rango de 20 minutos. Esto quiere decir que si el archivo indica que la observación ocurre a las 8:10, el programa selecciona los datos GPS desde las 8:00 hasta las 8:20 para el bus correspondiente.

En el archivo de salida se escribe el valor de los parámetros y el número del *set* de datos que se evalúa. Luego escribe “Si”, si el modelo indica que bus se pudo haber detenido, o “No” si indica lo contrario.

El Anexo 7.2, muestra cómo funciona de forma más detallada el programa, haciendo alusión a las fórmulas explicadas en el capítulo 2. Para mayor detalle de cómo funciona la programación del modelo, se puede acceder al Anexo computacional bajo la carpeta **Programa Modelo Bull y Herrera.**

3.2.2 Variaciones del modelo de Bull y Herrera

En esta subsección se presentan cuatro posibles cambios a la metodología. En primer lugar, se considera el caso en que la magnitud de la aceleración y la desaceleración no tengan los mismos valores. En segundo lugar, se propone el uso de la distancia entre log GPS como restricción frente a una no detención. En tercer lugar, respecto a los resultados de la sección anterior, se sugiere hacer avances con el uso de la velocidad instantánea que proporciona el log GPS. Por último, se sugiere relajar las restricciones asociadas al Caso 2, de forma de disminuir el valor de FNR.

La validación de estos cambios propuestos se realiza con los datos de la Base de datos 1, para así compararlos con los resultados obtenidos en la sección anterior.

3.2.2.1 Diferenciación entre aceleración y desaceleración

Lo descrito en la sección 3.2.1 sugiere que la aceleración y desaceleración no se comportan de forma idéntica. Aunque tienen valores similares, 0,9 m/s^2 y 1,1 m/s^2, no son del todo iguales. Esta observación hace que sea interesante plantear este cambio en la metodología.

Lo primero que se realiza es introducir la notación a utilizar. Ya no existe el valor a en el modelo, sino que se utiliza a_1 y a_2 para identificar la aceleración y desaceleración respectivamente.

Esto sugiere que tampoco existen los valores τ_a y L_a, ya que ahora existe un tiempo y largo asociado a la aceleración y otro a la desaceleración. Estos serán τ_{a1}, τ_{a2}, L_{a1}, y L_{a2}. Para no elaborar todo el modelo nuevamente, simplemente se actualizarán las restricciones para el Caso 1, Caso 2, y Caso 3.

***Caso 1**:* Se reformula la restricción que representa la ecuación (2.8), para esto se debe primero encontrar el mínimo entre los largos asociados.

$$\Delta = \min\{X_D - X_{SA}^D;\ X_{SA}^U - X_U;\ L_{a1}, L_{a2}\} \tag{4.1}$$

El mínimo entre estos largos, puede estar asociado a la aceleración o desaceleración, las restricciones serán distintas para cada caso.

Si $\Delta = X_D - X_{SA}^D$ *o* L_{a2}

$$\tau_s < \sqrt{2\Delta/a_2} - \Delta/v_{max} + \tau_{a1} \tag{4.2}$$

Si $\Delta = X_{SA}^U - X_U$ *o* L_{a1}

$$\tau_s < \sqrt{2\Delta/a_1} - \Delta/v_{max} + \tau_{a2} \tag{4.3}$$

***Caso 2**:* Se reformulan las restricciones de este caso presentadas en la Tabla 2.1, por las ecuaciones (4.6), (4.7), y (4.8) respectivamente.

Primero se debe formular la fracción de la aceleración o desaceleración que pudo haber ocurrido en este caso, las cuales son presentadas en las fórmulas (4.4) y (4.5).

$$(\tau_{a1} + \tau_{a2})\gamma_U = \sqrt{\frac{2(X_1 - X_{SA}^U)}{a_1}} - \frac{(X_1 - X_{SA}^U)}{v_{max}} + \tau_{a2} \tag{4.4}$$

$$(\tau_{a1} + \tau_{a2})\gamma_D = \sqrt{\frac{2(X_{SA}^D - X_1)}{a_2}} - \frac{(X_{SA}^D - X_1)}{v_{max}} + \tau_{a1} \tag{4.5}$$

Para que una observación sea catalogada como 'No Detención' se deben cumplir las restricciones (4.6), (4.7), y (4.8) en conjunto. La primera representa que el bus no se haya detenido entre X_U y X_D, mientras que las ecuaciones (4.7) y (4.8) se cercioran que si hubo una detención, esta no pueda ocurrir dentro del área de estacionamiento.

$$min\{(\tau_{a1} + \tau_{a2})\gamma_D;\ \tau_{s1D}\} + min\{(\tau_{a1} + \tau_{a2})\gamma_U;\ \tau_{sU1}\} < \tau_{a1} + \tau_{a2} \tag{4.6}$$

$$\tau_{sU1} < \sqrt{\frac{2(X_{SA}^{U} - X_U)}{a_2}} - \frac{(X_{SA}^{U} - X_U)}{v_{max}} + \tau_{a1} \quad (4.7)$$

$$\tau_{s1D} < \sqrt{\frac{2(X_D - X_{SA}^{D})}{a_1}} - \frac{(X_D - X_{SA}^{D})}{v_{max}} + \tau_{a2} \quad (4.8)$$

***Caso 3**:* La primera restricción asociada al Caso 3, presentada en la Tabla 2.1 se remplaza por las ecuaciones (4.9) y (4.10) en conjunto.

$$\tau_{sU1} < \sqrt{\frac{2(X_{SA}^{U} - X_U)}{a_2}} - \frac{(X_{SA}^{U} - X_U)}{v_{max}} + \tau_{a1} \quad (4.9)$$

$$\tau_{s12} + min\{(\tau_{a1} + \tau_{a2})\gamma_U;\ \tau_{sU1}\} < \tau_{a1} + \tau_{a2} \quad (4.10)$$

Mientras que la segunda restricción se remplaza por las ecuaciones (4.11) y (4.12) en conjunto.

$$\tau_{snD} < \sqrt{\frac{2(X_D - X_{SA}^{D})}{a_1}} - \frac{(X_D - X_{SA}^{D})}{v_{max}} + \tau_{a2} \quad (4.11)$$

$$\tau_{s(n-1)n} + min\{(\tau_{a1} + \tau_{a2})\gamma_D;\ \tau_{snD}\} < \tau_{a1} + \tau_{a2} \quad (4.12)$$

Finalmente, la restricción asociada a tres log GPS dentro del área de estacionamiento se remplaza por (4.13) en el modelo.

$$\tau_{s(i-1)i} + \tau_{si(i+1)} < \tau_{a1} + \tau_{a2} \quad (4.13)$$

3.2.2.2 Uso de distancias para distinguir la no detención

Utilizar las distancias de forma directa podría aportar a determinar la no detención entre dos log GPS y reducir el valor de FNR. Una restricción adicional que determine si

la distancia entre dos log GPS consecutivos es menor a la necesaria para el proceso de aceleración o desaceleración, puede determinar que el bus no se pudo haber detenido.

La nueva restricción se muestra en la ecuación (4.31)

$$X_D - X_U < L_a \tag{4.31}$$

En el modelo se utiliza esta restricción como una restricción inicial, que es independiente a las demás restricciones. Si ésta se cumple, clasifica a la observación como una no detención y no se prosigue con el modelo.

3.2.2.3 Incorporación de velocidades instantáneas

La introducción de esta variable podría ser muy beneficiosa para todos los casos en que se encuentre un log GPS dentro del área de estacionamiento, ya que será observable de forma directa si el bus se encuentra detenido.

Se utiliza v_i para representar la velocidad instantánea de un log GPS posicionado en X_i. Al agregar la velocidad instantánea se deben crear nuevas restricciones. Estas se presentan para el Caso 1, Caso 2, y Caso 3.

En el caso X_Uy X_D, sus velocidades instantáneas pueden ser útiles para no exigirle una desaceleración o aceleración completa al modelo, sino que una desaceleración desde v_U o una aceleración hasta v_D. Para esto, se crea un valor τ_a relacionado con la velocidad instantánea: τ_{aU} asociado a la desaceleración y τ_{aD} asociado a la aceleración. Se realiza el proceso análogo al explicado en las ecuaciones (2.1), (2.2), y (2.3), pero definimos τ_{acc}^i como el tiempo necesario para acelerar desde una velocidad igual a cero hasta v_i.

$$\tau_{acc}^i = v_i/a \tag{4.14}$$

Durante τ_{acc}^i el bus logra avanzar una distancia asociada, la llamaremos L_a^i.

$$L_a^i = \frac{\tau_{acc}^{i\ 2} \cdot a}{2} = \frac{v_i^2}{2a} \tag{4.15}$$

El tiempo que toma recorrer esta distancia a v_{max}, al que nos referiremos como τ_l^i, debe ser descontado de τ_{acc}^i.

$$\tau_l^i = \frac{L_a^i}{v_{max}} = \frac{v_i^2}{2a \cdot v_{max}} \tag{4.16}$$

$$\tau_{ai} = \tau_{acc}^i - \tau_l^i = \frac{v_i}{a} - \frac{v_i^2}{2a \cdot v_{max}} = \frac{2 \cdot v_i \cdot v_{max} - v_i^2}{2 \cdot a \cdot v_{max}} \tag{4.17}$$

Además de este uso, la velocidad instantánea puede ser beneficiosa para datos GPS dentro del área de estacionamiento. Si la velocidad instantánea dentro del paradero es cero resulta evidente que el bus se detuvo, mientras que, si la velocidad es elevada esta sugiere que no ocurre una detención dentro de esta área. Para asegurar que el bus no se haya podido detener, se debe cumplir que el largo asociado a la desaceleración debe ser más grande que la distancia del log GPS a los bordes del paradero. Esto se formula con la ecuación (4.18).

$$L_a^1 > \min\{X_1 - X_{SA}^U, X_{SA}^D - X_1\} \tag{4.18}$$

Caso 1: Se reformula la restricción que representa la ecuación (2.8)

$$\Delta = \min\{X_D - X_{SA}^D;\ X_{SA}^U - X_U;\ L_a\} \tag{4.19}$$

Si $\Delta = X_D - X_{SA}^D$

$$\tau_s < \tau_{aD} + \tau_a \tag{4.20}$$

Si $\Delta = X_{SA}^U - X_U$

$$\tau_s < \tau_{aU} + \tau_a \tag{4.21}$$

Si $\Delta = L_a$

$$\tau_s < 2\tau_a \tag{4.22}$$

***Caso 2**:* Se agrega la restricción (4.18), si ésta se cumple, el bus no se detuvo. Si ésta no se cumple, se evalúan las ecuaciones formuladas a continuación.

La ecuación (4.23) representa que el bus no se haya detenido entre X_U y X_D.

$$min\{(2\tau_a)\gamma_D;\ \tau_{s1D}\} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a \tag{4.23}$$

Las ecuaciones (4.25) y (4.26) se cercioran que si hubo una detención en τ_{s1D}, esta no pueda ocurrir dentro del área de estacionamiento.

$$\Delta = \min\{X_D - X_{SA}^D;\ L_a\} \tag{4.24}$$

Si $\Delta = X_D - X_{SA}^D$

$$\tau_{s1D} < \tau_{aD} + \tau_a \tag{4.25}$$

Si $\Delta = L_a$

$$\tau_{s1D} < \sqrt{2\Delta/a} - \Delta/v_{max} + \tau_a \tag{4.26}$$

En cambio, las ecuaciones (4.28) y (4.29) se cercioran que si hubo una detención en τ_{sU1}, ésta no pueda ocurrir dentro del área de estacionamiento.

$$\Phi = \min\{X_{SA}^U - X_U;\ L_a\} \tag{4.27}$$

Si $\Phi = X_{SA}^U - X_U$

$$\tau_{sU1} < \tau_{aU} + \tau_a \tag{4.28}$$

Si $\Phi = L_a$

$$\tau_{sU1} < \sqrt{2\Delta/a} - \Delta/v_{max} + \tau_a \tag{4.29}$$

***Caso 3**:* Se agrega la restricción (4.30). Si ésta se cumple, el bus no se detuvo. Si ésta no se cumple, se evalúan las ecuaciones asociadas a este caso en la Tabla 2.1.

$$L_a^n > \min\{X_n - X_{SA}^U, X_{SA}^D - X_n\} \tag{4.30}$$

Donde *n* representa el número del log GPS dentro del área de estacionamiento a evaluar. Esta ecuación se debe cumplir para cada valor de *n* posible.

3.2.2.4 Relajar restricciones

Si se analizan las restricciones asociadas al modelo, se observa que, a pesar de no exigir una detención completa en todas las restricciones, si lo realiza en algunas. Además, cuando no exige una detención completa, sí exige un proceso completo de aceleración o desaceleración.

En esta propuesta se busca relajar estas restricciones, de tal forma que se ajusten a la distancia que se encuentren los GPS del área de estacionamiento. Las restricciones reformuladas para el Caso 1, Caso 2, y Caso 3, se presentan a continuación.

Caso 1*:* Permite tanto una aceleración como una desaceleración incompleta

$$\Delta = \min\{X_D - X_{SA}^D;\ L_a\} \tag{4.32}$$

$$\varphi = \min\{X_{SA}^U - X_U;\ L_a\} \tag{4.33}$$

$$\tau_s < \sqrt{2\Delta/a} - \Delta/v_{max} + \sqrt{2\varphi/a} - \varphi/v_{max} \tag{4.34}$$

Caso 2*:* Se utilizan los valores de las ecuaciones (4.32) y (4.33) para generar las restricciones.

$$\tau_{sU1} < \sqrt{\frac{2(X_1 - X_{SA}^U)}{a}} - \frac{(X_1 - X_{SA}^U)}{v_{max}} + \sqrt{2\varphi/a} - \varphi/v_{max} \tag{4.35}$$

$$\tau_{s1D} < \sqrt{\frac{2(X_{SA}^D - X_1)}{a}} - \frac{(X_{SA}^D - X_1)}{v_{max}} + \sqrt{2\Delta/a} - \Delta/v_{max} \tag{4.36}$$

Caso 3*:*

$$\tau_{sU1} < \sqrt{2\varphi/a} - \varphi/v_{max} + \sqrt{\frac{2(X_1 - X_{SA}^U)}{a}} - \frac{(X_1 - X_{SA}^U)}{v_{max}} \tag{4.37}$$

$$\tau_{snD} < \sqrt{\frac{2(X_n - X_{SA}^D)}{a_1}} - \frac{(X_n - X_{SA}^D)}{v_{max}} + \sqrt{2\Delta/a} - \Delta/v_{max} \tag{4.38}$$

$$\tau_{s(i-1)i} + \tau_{si(i+1)} < \sqrt{\frac{2(X_{i+1}-X_i)}{a}} - \frac{(X_{i+1}-X_i)}{v_{max}} + \sqrt{\frac{2(X_i-X_{i-1})}{a}} - \frac{(X_i-X_{i-1})}{v_{max}} \quad (4.39)$$

3.2.3 Análisis de sensibilidad del modelo Bull y Herrera

A fin de determinar la influencia de los parámetros v_{max} y a en el modelo, se procede a realizar un análisis de sensibilidad. Además de evaluar la sensibilidad del modelo a cambios en estos dos parámetros, se procedió a evaluar la sensibilidad del largo del paradero y de su ubicación. Es decir, interesa determinar el impacto de no contar con la información correcta de estas dos variables.

Durante este proceso se utilizan los datos de la *Base de datos 1*, ya que corresponden a datos de Transantiago. Para cuantificar la sensibilidad del modelo respecto a cada uno de los parámetros o datos, se utilizan los indicadores FPR, FNR, y la precisión. A continuación, se presentan los valores utilizados en el análisis de sensibilidad tanto de los parámetros como de los datos a evaluar.

3.2.3.1 Velocidad máxima

El valor utilizado en la validación es 60 km/h. Las variaciones respecto a esta variable se hicieron en descenso hasta llegar a 45 km/h. Esto resulta coherente con lo que ocurre en la realidad, ya que un bus al detenerse periódicamente debido a los paraderos y semáforos, la mayoría de las veces no alcanza a circular a 60 km/h entre detenciones. Así, se evaluó el modelo en: 45 km/h, 48 km/h, 51 km/h, 54 km/h, 57, y 60 km/h.

3.2.3.2 Aceleración

El valor utilizado en la validación es $a = 1\ m/s^2$. Considerando los valores encontrados a partir de la Base de Datos 2 (ver sección 3.2.1), este parámetro se varió en $0{,}1\ m/s^2$, desde $0{,}8\ m/s^2$ hasta $1{,}2\ m/s^2$ para realizar el análisis de sensibilidad.

3.2.3.3 Largo del paradero

En teoría, cada paradero tiene su propio largo dependiendo de la cantidad de bahías con que cuente, por lo que no requiere calibración. Por ejemplo, paraderos que se encuentran en la periferia de Santiago, donde pasan uno o dos servicios de buses, tienen en general, un largo de 20 metros. Mientras que los paraderos ubicados en intersecciones de avenidas o en estaciones intermodales, su largo puede llegar hasta los 50 metros.

Sin embargo, en la práctica este largo depende del modo de operar de los conductores. Si el conductor abre las puertas del bus en el segundo sitio, para que los usuarios desciendan o asciendan al bus, el paradero a pesar de demarcar un largo de 20 metros, este operaría como uno de 50. Por lo tanto, es necesario realizar este análisis de sensibilidad para determinar la importancia de la exactitud en la calibración de este dato.

Como los paraderos pueden tener un largo desde 20 hasta 50 metros, las variaciones se hicieron en esta misma línea, evaluando tres posibles largos: 20, 30, y 40 metros. Los parámetros h_1 y h_2, indican el largo en metros que tiene el paradero aguas abajo y aguas arriba del punto que define su ubicación geográfica. Estos parámetros sirven para determinar los valores de X_{SA}^{D} y X_{SA}^{U}. La Tabla 3.5 se muestra el valor de h_1 y h_2 para los tres largos de paradero a evaluar.

Tabla 3.5 Valores de h1 y h2 al variar el largo del paradero
Fuente: Elaboración propia

Largo de paradero	***Parámetro***	***Valor***
20 metros	h_1	5 metros
	h_2	15 metros
30 metros	h_1	5 metros
	h_2	25 metros
40 metros	h_1	5 metros
	h_2	35 metros

Como se observa en la Tabla 3.5, se decide fijar el valor de h_1 en 5 metros, e ir variando únicamente el valor de h_2. Se toma esta decisión considerando que la ubicación geográfica de los paraderos suele estar cerca de un semáforo, y el aumentar el valor de h_1 no refleja lo que ocurre en la realidad.

Además, la ubicación teórica puntual del paradero corresponde al extremo aguas abajo, donde usualmente se ubica la señal y donde se detiene la primera puerta del bus. La Figura 3.3 muestra esta situación, donde la estrella refleja la georreferenciación del paradero.

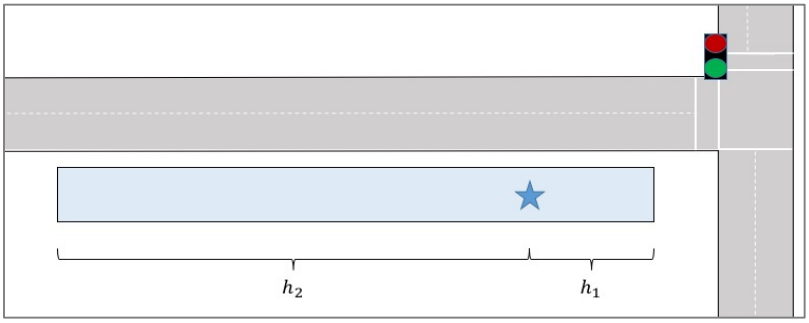

Figura 3.3 Visualización de los valores h1 y h2 en un paradero

Fuente: Elaboración propia

3.2.3.4 Ubicación del paradero

La georreferenciación consiste en ubicar en un punto específico cada paradero, este determina en qué calle y entre qué cuadras se encuentran cada uno de ellos. Sin embargo, un pequeño fallo de cálculo puede localizar al paradero a 20 o 40 metros de su ubicación real. Este error es comparable con localizar al paradero justo en una intersección en vez de 10 metros antes de ella. Este genera incertidumbre acerca de la ubicación real de los paraderos.

Para analizar el efecto que produciría este error en el modelo, se crean tres escenarios posibles. El primero es el escenario original, que es ubicar el paradero donde se encuentra georreferenciado. El segundo escenario, ubica al paradero 40 metros aguas

arriba en la ruta, mientras que el tercer escenario ubica al paradero 40 metros aguas abajo de ella. Se elige este largo porque genera un cambio sustantivo en la ubicación geográfica del paradero, siguiendo la línea de lo mencionado anteriormente.

Para hacer efectivo este cambio, simplemente se utilizaron los valores de h_1 y h_2 del modelo. La Tabla 3.6 muestra un resumen de los valores que toman los parámetros para cada uno de los escenarios.

Tabla 3.6 Valores de h1 y h2 para distintas ubicaciones de paraderos

Fuente: Elaboración propia

Ubicación	***Parámetro***	***Valor***
Punto actual	h_1	5 metros
	h_2	35 metros
40 metros aguas abajo	h_1	-35 metros
	h_2	75 metros
40 metros aguas arriba	h_1	45 metros
	h_2	-5 metros

3.3 Modelos basados en aprendizaje supervisado

En esta sección se explica qué métodos se utilizan para generar modelos basados en aprendizaje supervisado y cómo se calibra y valida cada uno de estos modelos. Se utilizan dos métodos distintos: Máquinas de soporte de vectores y Análisis discriminante lineal. Para este proceso se utiliza la *Base de datos 1*, que permite percibir cómo se comportan los modelos al utilizar datos de Transantiago.

Estos datos se dividen en un 70 y 30 por ciento, donde la primera parte se utiliza para calibrar los distintos modelos, y la parte restante se utiliza para validar y comparar los modelos entre sí. Se toma una muestra aleatoria de estos datos, generando dos grupos de 260 y 111 observaciones correspondientemente.

Tabla 3.7 Variables descriptivas para modelos basados en aprendizaje supervisado

Fuente:Elaboración propia

VARIABLES	DESCRIPCIÓN
Nº datos	Muestra el número de log GPS recogido, siendo 2 el mínimo.
Tipo	Variable que toma el valor de 1 si no existen log GPS en el área de estacionamiento, 2 si posee un log, y 3 si posee más de uno.
v1	La velocidad instantánea que entrega el GPS justo antes de *SA*
vf	La velocidad instantánea que entrega el primer GPS justo después de *SA*.
distTotal	La distancia en ruta, entre el último log recogido y el primero ($X_D - X_U$).
vpromedio	Velocidad promedio entre el último log recogido y el primero
vparadero	Velocidad estimada dentro del paradero
$X_D - X_D^{SA}$	Distancia desde el último log recogido hasta el paradero
$X_U^{SA} - X_U$	Distancia desde el primer log recogido hasta el paradero

Los modelos basados en aprendizaje supervisado requieren de un proceso de dos etapas, calibración y clasificación. Durante el proceso de calibración, en base a un *set* de datos, el modelo aprende de estos datos y busca patrones en él. En la clasificación, toma otros datos y aplica el modelo creado sobre ellos, prediciendo la clase a la que pertenecen. En base a esta última etapa se obtienen indicadores acerca del rendimiento del modelo.

Para generar estos modelos, el primer paso es generar variables predictivas a partir de los datos GPS de los buses. En total se generan nueve variables que están presentes en todas las observaciones (ver Tabla 3.7). Las variables predictivas se construyen a partir del log GPS justo aguas arriba (X_U) y justo aguas abajo (X_D) del paradero, además de todos los datos dentro del área de estacionamiento (SA).

La variable **vparadero** se estima de forma distinta dependiendo del valor de Nºdatos. Si $\mathbf{Nºdatos} = 2$, **vparadero** = **vpromedio** ya que no hay más información

para obtener un dato más específico. Mientras que si **N** > 2, la variable **vparadero** toma el valor promedio de las velocidades instantáneas de los datos GPS dentro de SA.

La variable explicativa corresponde a la detención o no detención del bus en el paradero. Esta variable binaria tiene el valor "Si" si la observación corresponde a una detención y el valor de "No" en el caso contrario. A esta variable se le llama **Realidad**.

El resultado de este proceso es una tabla donde cada fila representa una observación. La Tabla 3.8 muestra un ejemplo de este caso, a partir de la planilla Excel donde se encuentran los datos.

Tabla 3.8 Observaciones con variables para modelos de aprendizaje supervisado

Fuente: elaboración propia via Excel

Patente	Paradero	Realidad	Tipo	XUSA-Xu	XD-XDSA	v1	vfinal	Ndatos	DistTotal	vpromedio	vparadero
BJFD-88	PE112	Si	2	176.30	212.11	0	0	3	428.40	25.70	46.86
BJFF-14	PE112	No	2	176.42	236.01	2.96	53.89	3	452.43	27.15	0.00
BJFF-14	PE113	No	1	189.76	189.91	53.89	27.97	2	419.67	50.36	50.36
BJFG-61	PE112	Si	2	104.42	133.34	0	43.89	3	277.76	16.67	14.82
BJFG-61	PE113	No	1	292.42	4.04	43.89	0	2	336.46	40.37	40.37
BJFG-77	PE112	No	1	168.02	221.32	58.89	34.82	2	429.33	51.52	51.52
BJFG-77	PE113	No	1	204.45	189.01	34.82	33.89	2	433.46	52.02	52.02
BJFG-82	PE112	No	1	226.00	167.07	35.93	42.97	2	433.07	51.97	51.97
BJFG-82	PE113	No	1	138.94	234.45	35.93	21.85	2	413.40	49.61	49.61
BJFG-84	PE112	Si	1	29.96	209.37	35.93	0	2	279.33	33.52	33.52
BJFG-84	PE113	No	1	18.89	52.17	30.93	28.89	2	111.06	13.33	13.33
BJFH-71	PE113	No	2	213.23	175.59	24.82	0	3	428.82	25.73	28.89
BJFH-84	PE112	Si	2	101.26	86.15	21.85	38.89	3	227.41	13.64	20.93
BJFH-84	PE113	No	1	339.61	1.94	38.89	10.93	2	381.55	45.79	45.79
BJFJ-34	PE113	No	1	365.24	90.23	50.93	57.97	2	495.48	59.46	59.46
BJFJ-50	PE113	No	1	96.48	265.85	42.97	9.82	2	402.33	48.28	48.28
BJFJ-60	PE112	No	1	195.03	130.77	17.96	45.93	2	365.80	43.90	43.90
BJFJ-60	PE113	No	1	124.19	237.42	36.86	16.85	2	401.61	48.19	48.19
BJFJ-68	PE113	No	1	222.14	70.41	0	60.93	2	332.55	39.91	39.91

Como es explicado en el capítulo 2, antes de generar los modelos se deben estandarizan los valores para cada una de las variables, de tal forma que la media sea cero y la desviación estándar sea uno. Una vez realizado esto, se toman las observaciones reservadas para la calibración de modelos y se utiliza la herramienta *Classification Learner* de MATLAB. En ésta se calibra el modelo utilizando validación cruzada, que divide el 70 por ciento de los datos destinados a esta etapa en cinco partes. De esta forma, el modelo se calibra con cuatro partes y se evalúa con la parte restante, turnándose entre ellas de tal forma que se evalúa con todos los datos una vez.

Para los métodos a evaluar, se realiza una búsqueda exhaustiva, intentando todas las combinaciones que permiten las 9 variables predictivas, desde utilizarlas todas hasta

utilizar solo una de las variables. Además, se varían los costos asociados a el Error Tipo I y Tipo II en la función objetivo, se varían en números enteros desde uno hasta diez. De todas las posibles combinaciones generadas se toman en consideración solo aquellas que presenten un FPR menor o igual al registrado con el modelo Bull y Herrera con los datos de clasificación, y de estos se elige el que presente una mayor precisión.

Una vez calibrados los modelos, se toma el 30% de los datos reservados para la clasificación y se clasifican estos datos. Esto permite corroborar el proceso de calibración, viendo si el modelo se ajusta a datos con los que no fue calibrado. Se utiliza el programa MATLAB para este proceso, y los resultados de estos se presentan a través de los indicadores FPR, FNR, y precisión.

3.4 Comparación de modelos

Una vez formulados los modelos basados en aprendizaje supervisado, se procede a realizar una comparación entre modelos. Para esto se utiliza el modelo mejor evaluado entre el modelo de Bull y Herrera y sus modificaciones, y los modelos SVM y LDA.

Como estos últimos han sido formulados con el 70% de la *Base de datos 1* se comparan los resultados del 30% restante, tanto del modelo basado en cinemática como de los basados en métodos de aprendizaje supervisado. Esto se hace utilizando los indicadores FPR, FNR, y precisión.

3.5 Análisis respecto a la periodicidad de datos

Una gran limitación que presentan los datos GPS de los buses del Transantiago es su periodicidad, la cual ha sido referida como τ a lo largo de esta investigación. Con una menor periodicidad sería posible representar de mejor manera la trayectoria de los buses. Aunque no es posible determinar de forma exacta la trayectoria del bus, un mayor número de datos nos entrega una mayor información.

Debido a esto, se realiza el análisis de cómo afecta la periodicidad entre datos GPS (o su inverso, la frecuencia) al funcionamiento del modelo. Se analizan cuatro periodicidades distintas: 10 segundos, 15 segundos, 30 segundos, y 60 segundos.

Esta sección se divide en dos, evaluando en primera instancia el modelo basado en cinemática y luego los modelos basados en aprendizaje supervisado. Los resultados se presentan en base a los indicadores FPR, FNR, y precisión.

3.5.1 Modelo basado en cinemática

Durante este análisis, se analiza el modelo mejor evaluado entre el modelo de Bull y Herrera y las variaciones propuestas. La *Base de datos 2* posee datos cada un segundo, lo que permite generar sets de datos con diferentes periodicidades. Como se menciona en la sección 3.1.2, es posible generar τ sets de datos distintos a partir de una misma observación. Esto quiere decir que es posible evaluar τ veces la detención o no detención de un bus en un paradero con la misma observación. La Tabla 3.9 muestra el número de validaciones que es posible realizar para la *Base de datos 2* con cada valor de τ.

Tabla 3.9 Numero de validaciones para los distintos valores de τ
Fuente: Elaboración propia

τ	No detenciones	Detenciones	Totales
10 segundos	980	860	1840
15 segundos	1470	1290	2760
20 segundos	1960	1720	3680
30 segundos	2940	2580	5520
45 segundos	4410	3870	8280
60 segundos	5880	5160	11040

3.5.2 Modelos basados en aprendizaje supervisado

Los modelos basados en aprendizaje supervisado descritos en la sección 3.3, corresponden a modelos calibrados con datos GPS cada 30 segundos. Las variables predictivas a utilizar dependen de la periodicidad, por ejemplo: número de datos, distancia total, y distancias desde o hasta el paradero, se ven afectadas por el valor de τ. Es por esta razón, que se deben calibrar y crear modelos para los distintos valores de τ utilizando los métodos SVM y LDA.

En esta etapa se utiliza solo un set de datos por observación para no generar un sobreajuste en los modelos, ya que, aunque los *sets* de datos son diferentes, al tener solo un segundo de diferencia pueden presentar patrones que generen un modelo sobre ajustado.

El primer paso es generar las variables predictivas para cada valor de τ, para esto se tendrán cuatro tablas distintas. En cada una de estas tablas se deben estandarizar los datos de cada una de las variables. Cada una de las tablas tiene 184 observaciones, las que se dividen de forma aleatoria en un 70 y 30 por ciento para las etapas de calibración y clasificación.

La Base de datos 2, cuenta con 184 observaciones, por lo que en la etapa de calibración y clasificación se utilizan 129 y 55 observaciones respectivamente. De esta forma para los distintos valores de τ, se calibran los modelos con las 129 observaciones destinadas para esto, y se validan con las 55 observaciones restantes.

Luego se utiliza el mismo procedimiento descrito en la sección 3.3 para calibrar los distintos modelos, para finalmente evaluar con la porción de los datos reservados para la clasificación.

4 RESULTADOS

Este cápitulo se divide en cuatro partes. La primera se enfoca en los resultados de la validación de los modelos basados en cinemática. La segunda parte presenta los resultados del proceso de calibración y validación de los modelos basados en aprendizaje supervisado. La tercera realiza una breve comparación entre los modelos. Mientras que la cuarta y última parte presenta los resultados del efecto que tiene la periodicidad de datos sobre los indicadores.

4.1 Modelos basados en cinemática

Esta sección se divide en tres partes: Validación de modelo Bull y Herrera y análisis de error, variaciones al modelo, y análisis de sensibilidad.

4.1.1 Validación y análisis de error modelo Bull y Herrera

Durante esta etapa se utiliza la Base de datos 1, esto corresponde a 371 observaciones que han sido clasificadas por el modelo utilizando el valor de los parámetros $v_{max} = 60\ km/h$ y $a = 1\ m/s^2$. Los resultados se presentan en la Tabla 4.1.

Tabla 4.1 Indicadores clasificación modelo Bull y Herrera

Fuente: Elabroación propia

Indicador	Valor
Precisión	85,98%
FPR	11,24%
FNR	16,34%

Para ser capaces de observar cómo se comportan las restricciones del modelo, se analizan por separado los siguientes casos: No hay log GPS en el área de estacionamiento, hay un log GPS en el área de estacionamiento y hay más de un log GPS en el área de estacionamiento. A estos tres casos se les llamará Caso 1, Caso 2 y Caso 3, respectivamente. La Tabla 4.2 muestra un resumen de los resultados.

Tabla 4.2 Observaciones y error por Caso

Fuente: Elaboración propia

	NO SE DETIENEN		**SI SE DETIENEN**	
	Porcentaje de casos	***FNR***	***Porcentaje de casos***	***FPR***
Caso 1: 0 log	89,11%	8,33%	53,85%	17,58%
Caso 2: 1 log	10,89%	81,82%	42,01%	4,23%
Caso 3: 2 o más logs	0%	0%	4,14%	0%
Total	***Error tipo II***	***16,34%***	***Error tipo I***	***11,24%***

Considerando los tres casos, el Error Tipo I se encuentra en un valor menor que el Error Tipo II. Sin embargo, se encuentran a 5,1 puntos porcentuales de diferencia. Esto quiere decir que el modelo tiende a decir de forma equilibrada que los buses se detuvieron o no se detuvieron.

Para la no detención, la mayoría de las veces se encuentra en el Caso 1, donde el modelo falla en predecir un 8,33 porciento. Las observaciones restantes corresponden al Caso 2, donde el modelo falla en predecir más del 80 porciento de las veces. Esto refleja que las restricciones asociadas al Caso 2 podrían ser menos estrictas. De esta forma se podría reducir el FNR en este caso, pero aumentaría el FPR. Sin embargo, esto depende del uso que se le entrega al modelo, si se desea multar los casos en que no se detuvo el bus, es recomendable disminuir el FPR en función de un aumento de FNR.

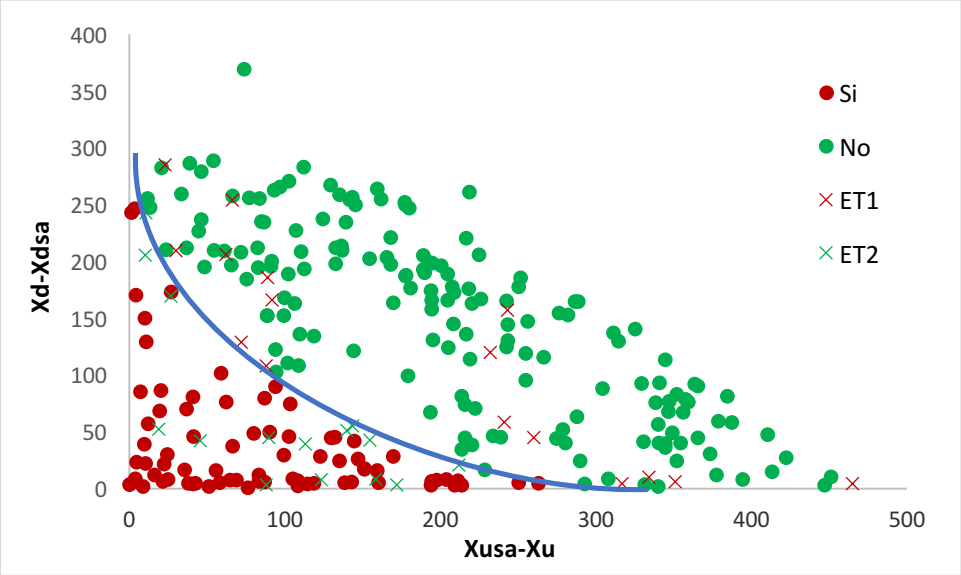

Figura 4.1 Clasificación Modelo Bull y Herrera –Caso 1 al graficar las ditancias de los GPS antes y después del paradero

Fuente: Elaboración propia

La Figura 4.1 muestra las distancias de los GPS antes y después del paradero para las observaciones que se encuentran en el Caso 1. Presentar estas variables es muy importante ya que esta es la única información con la que cuenta el modelo para la predicción.

Los datos rojos representan los buses que se detuvieron y los datos verdes los que no se detuvieron, mientras que los círculos representan los datos bien clasificados y las cruces los mal clasificados. La curva azul separa los datos en dos, donde todos los datos bajo la curva el modelo los clasifica como "Detención", y sobre la curva como "No detención".

Una posibilidad sería hacer uso de la velocidad instantánea, esta información puede traer beneficios a la hora de realizar la clasificación. No obstante, en 30 segundos un bus se puede encontrar con un semáforo o con una detención que no necesariamente se encuentra en el paradero. Debido a esto, aunque estos cambios se implementen, la predicción puede no resultar exitosa.

4.1.2 Cambios realizados al modelo Bull y Herrera

4.1.2.1 Diferenciación entre aceleración y desaceleración

El resultado de este modelo se presenta en la Tabla 4.3, se comparan los indicadores del modelo de Bull y Herrera sin y con los cambios propuestos. Se utilizaron los parámetros $v_{max} = 60,\ h_1 = 5, h_2 = 25$ para ambos modelos, y $a = 1$ en el primer caso, y $a_1 = 0{,}9$ $a_2 = 1{,}1$ en el segundo.

Tabla 4.3 Comparación de indicadores entre modelo original y variación aceleración

Fuente: Elaboración propia

	BULL Y HERRERA	VARIACIÓN 1
Precisión	85,98%	85,71%
FPR	11,24%	12,43%
FNR	16,34%	15,84%

Se observa que no existen cambios sustanciales, la precisión empeora 0,27 puntos porcentuales respecto al modelo original. Esto puede ser causa de que los valores encontrados para a_1 y a_2 no representan lo que ocurre en este caso, ya que fueron obtenidos con la Base de datos 2. Esto demuestra que suponer que la magnitud de la aceleración y desaceleración son iguales, es adecuado.

4.1.2.2 Uso de distancias para distinguir la no detención

El resultado producto de este cambio, se presenta en la Tabla 4.4. En comparación con la propuesta anterior, esta presenta un mayor cambio en los indicadores. A pesar de que se presenta una baja en el valor de FNR, el FPR aumenta significativamente. Esto quiere decir que la restricción propuesta está castigando a muchos buses que sí se detuvieron.

Tabla 4.4 Comparación de indicadores entre modelo original y al utilizar distancias

Fuente: Elaboración propia

	BULL Y HERRERA	VARIACIÓN 2
Precisión	85,98%	79,25%
FPR	11,24%	27,81%
FNR	16,34%	14,85%

La precisión disminuye desde 85,98% hasta 79,25%, debido al aumento significativo de FPR que llega a valores sobre el 27%. La inclusión de la distancia entre dos log GPS consecutivos de la forma propuesta al modelo puede traer beneficios en el FNR. Sin embargo, el aumento significativo del Error Tipo I hace que no sea recomendable.

4.1.2.3 Incorporación de velocidades instantáneas

El resultado de este modelo se presenta en la Tabla 4.5. Al igual que en los casos anteriores se comparan los indicadores del modelo sin (1) y con (2) los cambios propuestos.

Tabla 4.5 Comparación de indicadores entre modelo original y variación con velocidad intantánea

Fuente: elaboración propia

	BULL Y HERRERA	VARIACIÓN 3
Precisión	85,98%	83,56%
FPR	11,24%	12,43%
FNR	16,34%	19,80%

Los cambios propuestos no benefician la predicción del modelo, tanto FNR como FPR aumentan respecto al modelo original. Aunque esta propuesta genera cambios

significativos en la formulación del modelo, la velocidad instantánea no genera los cambios esperados.

4.1.2.4 Relajar restricciones

En la Tabla 4.6 se observa que los cambios implementados no producen los cambios esperados en los indicadores. Esta variación de relajar las restricciones se realiza en función de disminuir FNR. Sin embargo, lo único que se refleja es un aumento de FPR. Esto es coherente, ya que al relajar las restricciones se espera que el modelo prediga de forma más fácil que los buses no se detuvieron.

Tabla 4.6 Comparación de indicadores entre modelo original y variación al relajar restricciones

Fuente: elaboración propia

	BULL Y HERRERA	VARIACIÓN 4
Precisión	85,98%	85,18%
FPR	11,24%	13,02%
FNR	16,34%	16,34%

4.1.3 Análisis de sensibilidad del modelo

4.1.3.1 Parámetros

Esta subsección presenta los resultados del análisis de sensibilidad respecto a los parámetros v_{max} y a. Esta sección se realiza con los datos de la Base de datos 1 y los resultados se analizan respecto a los indicadores FPR, FNR, y precisión.

La Tabla 4.7 muestra cómo varía el valor de FPR al variar los parámetros. Se observa que a medida que la aceleración y la velocidad máxima aumentan, el FPR tiende a disminuir. Esto es coherente con la formulación del modelo. Al tener una velocidad máxima mayor, el modelo le entrega un menor tiempo asociado a recorrer la distancia

entre dos log GPS, teniendo un mayor τ_s para hacer posible la detención. Un alto valor de a, en cambio, sugiere un proceso de detención más corto, exigiendo un menor valor de τ_s para asegurar una posible detención.

Tabla 4.7 Valor del indicador FPR al realizar análisis de sensibilidad

Fuente: Elaboración propia

		VELOCIDAD						
		45 km/h	48 km/h	51 km/h	54 km/h	57 km/h	60 km/h	***Promedio***
ACELERACIÓN	0,8 m/s^2	27,2%	24,8%	24,8%	23,2%	22,4%	22,4%	***24,1%***
	0,9 m/s^2	20,8%	20,0%	18,4%	18,4%	15,2%	12,8%	***17,6%***
	1 m/s^2	15,2%	12,8%	12,0%	11,2%	11,2%	11,2%	***12,3%***
	1,1 m/s^2	11,2%	10,4%	8,0%	7,2%	7,2%	8,0%	***8,7%***
	1,2 m/s^2	9,6%	8,8%	8,0%	7,2%	7,2%	7,2%	***8,0%***
	Promedio	***16,8%***	***15,4%***	***14,2%***	***13,4%***	***12,6%***	***12,3%***	

Tabla 4.8 Valor del indicador FNR al realizar análisis de sensibilidad

Fuente: Elaboración propia

		VELOCIDAD						
		45 km/h	48 km/h	51 km/h	54 km/h	57 km/h	60 km/h	***Promedio***
ACELERACIÓN	0,8 m/s^2	6,5%	6,5%	6,5%	6,5%	6,5%	8,2%	***6,8%***
	0,9 m/s^2	9,0%	9,0%	8,2%	7,4%	8,2%	9,8%	***8,6%***
	1 m/s^2	11,4%	11,4%	11,4%	12,3%	14,7%	16,3%	***12,9%***
	1,1 m/s^2	14,7%	16,3%	16,3%	17,2%	19,6%	21,2%	***17,6%***
	1,2 m/s^2	16,3%	16,3%	20,4%	22,1%	23,7%	25,3%	***20,7%***
	Promedio	***11,6%***	***11,9%***	***12,6%***	***13,1%***	***14,5%***	***16,2%***	

La Tabla 4.8 muestra en cambio cómo varía el valor de FNR al variar los parámetros. Esta tabla muestra el efecto contrario al observado en el análisis respecto a FPR, donde FNR aumenta a medida que la aceleración y la velocidad máxima aumentan. Esto es coherente con lo anterior, ya que a medida que los parámetros aumentan, el

modelo tiende a decir que los buses si se detienen, y a medida que estos disminuyen, la detención se hace menos probable.

El efecto que producen estas variaciones en la precisión del modelo se presenta en la Tabla 4.9.

Tabla 4.9 Valor de la Precisión al realizar análisis de sensibilidad

Fuente: Elaboración propia

		VELOCIDAD						
		45 km/h	48 km/h	51 km/h	54 km/h	57 km/h	60 km/h	***Promedio***
ACELERACIÓN	0,8 m/s^2	84,05%	85,14%	85,14%	85,87%	86,24%	85,35%	***85,3%***
	0,9 m/s^2	85,63%	86,00%	87,17%	87,61%	88,63%	88,83%	***87,3%***
	1 m/s^2	86,85%	87,94%	88,31%	88,23%	86,89%	85,98%	***87,4%***
	1,1 m/s^2	86,89%	86,37%	87,46%	87,38%	86,04%	84,79%	***86,5%***
	1,2 m/s^2	86,73%	87,09%	85,23%	84,71%	83,82%	82,93%	***85,1%***
	Promedio	***86,0%***	***86,5%***	***86,7%***	***86,8%***	***86,3%***	***85,6%***	

En esta tabla se aprecia el efecto causado en FPR y FNR al variar los parámetros. Estos se contraponen, haciendo que el cambio de la precisión al variar los parámetros no sea determinante. Para observar este efecto en mayor detalle se presentan por separado los efectos que producen cada uno de los parámetros, graficando el valor promedio de los indicadores FPR y FNR, y su desviación estándar.

Velocidad máxima

En el caso de FPR, la desviación estándar bordea el valor de 0,055 para los distintos valores de v_{max}, lo que se muestra en la Figura 4.2. Las líneas negras representan el rango de valores entre el valor mínimo y el máximo para el indicador, mientras que las barras o rectángulos representan la desviación estándar para cada valor de v_{max}. La misma información se presenta en la Figura 4.3 para FNR.

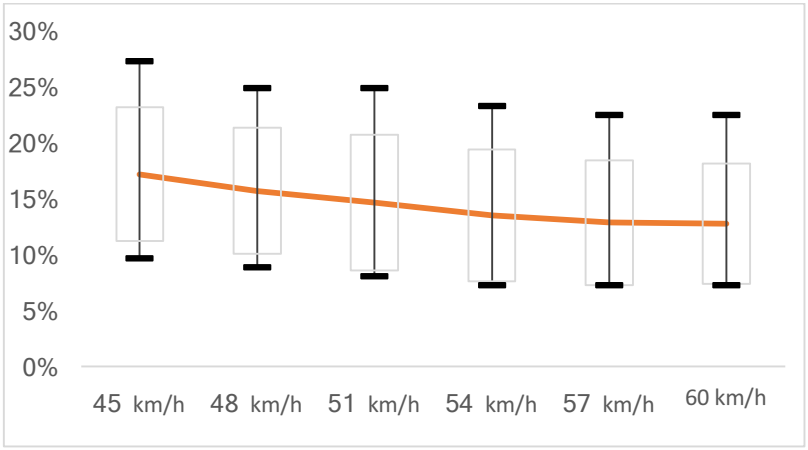

Figura 4.2 FPR y su desviación estándar – Vmax

Fuente: Elaboración propia

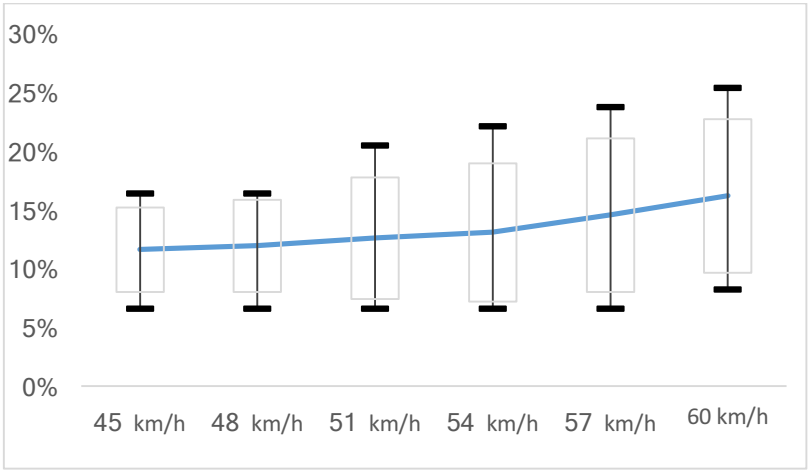

Figura 4.3FNR y su desviación estándar - Vmax

Fuente: Elaboración propia

La principal diferencia que presentan estas figuras, es que FPR presenta desviación uniforme a lo largo de v_{max}, mientras que FNR va aumentando su desviación al aumentar v_{max}. Esto es un factor importante al calibrar el modelo, ya que, aunque este parámetro no presenta mejoras en términos de precisión, si disminuye el Error Tipo I al aumentar v_{max}. Sin embargo, al aumentar v_{max} el Error Tipo II se vuelve más variable.

En conclusión, si el modelador estima que es más importante no incurrir en el Error Tipo I, se sugiere que la velocidad máxima sea 60 km/h de modo ser más exigente con el modelo. No obstante, el modelo se considera robusto respecto a v_{max}, por lo que no influye de forma categórica el valor del parámetro, mientras se encuentre dentro de los valores establecidos.

Aceleración

Las Figuras 4.4 y 4.5 presentan la desviación estándar para FPR y FNR.

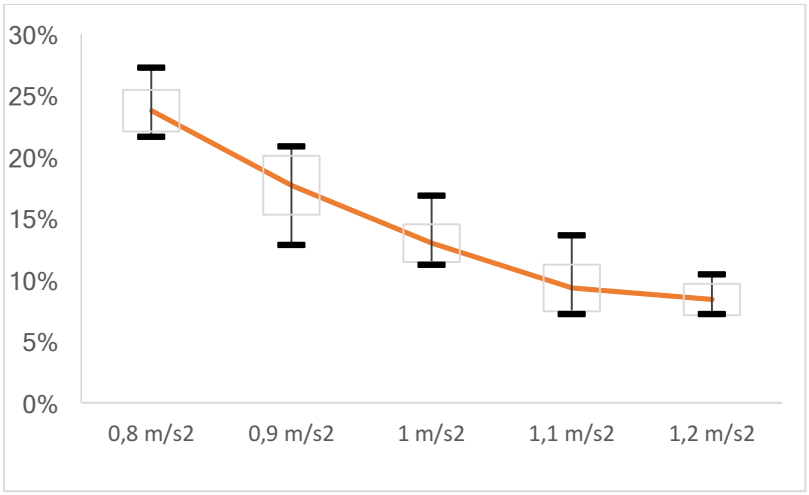

Figura 4.4 FPR y su desviación estándar – Caso aceleración

Fuente: Elaboración propia

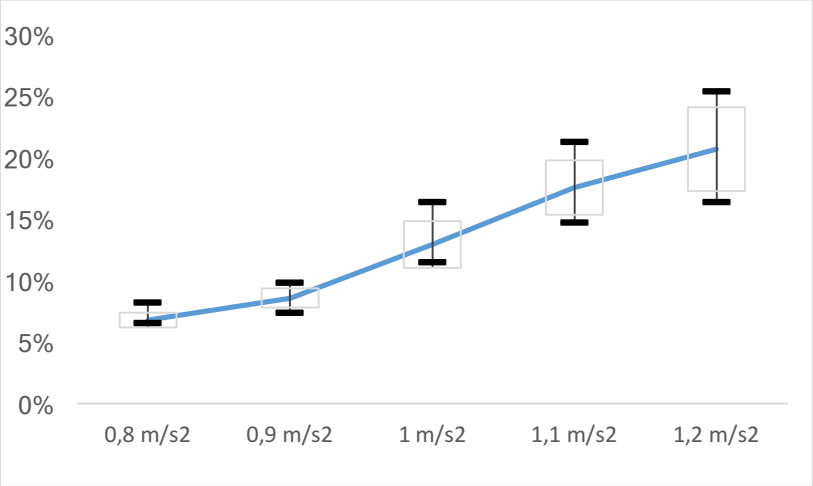

Figura 4.5 FNR y su desviación estándar - Caso aceleración
Fuente: Elaboración propia

La principal diferencia que presentan estas figuras con las presentadas en el caso anterior es que, a pesar de que las curvas de los indicadores son más pronunciadas, la desviación estándar es menor, con valores promedio de 0,015 y 0,013 para FPR y FNR respectivamente. Por otro lado, no se observa una clara relación entre los valores de los indicadores con el valor de la desviación estándar.

A partir de lo observado, el modelo se considera robusto respecto al valor de la aceleración, ya que no se presentan cambios relevantes en la precisión. El valor de a a recomendar puede depender del indicador que se desea priorizar, donde una aceleración 1,2 m/s^2 presenta el menor Error Tipo I. Sin embargo, una aceleración de 1 m/s^2 podría mantener los valores de FNR y FPR más equilibrados.

4.1.3.2 Datos sobre el paradero

En esta subsección se presenta el análisis que se realiza tanto para el largo del paradero como la ubicación del paradero. Se utiliza el valor de los parámetros $v_{max} = 60\ km/h$ y $a = 1\ m/s^2$.

Largo del paradero

La Figura 4.6 grafica los resultados para el análisis realizado con el largo del paradero, donde se observan los indicadores FPR, FNR y precisión.

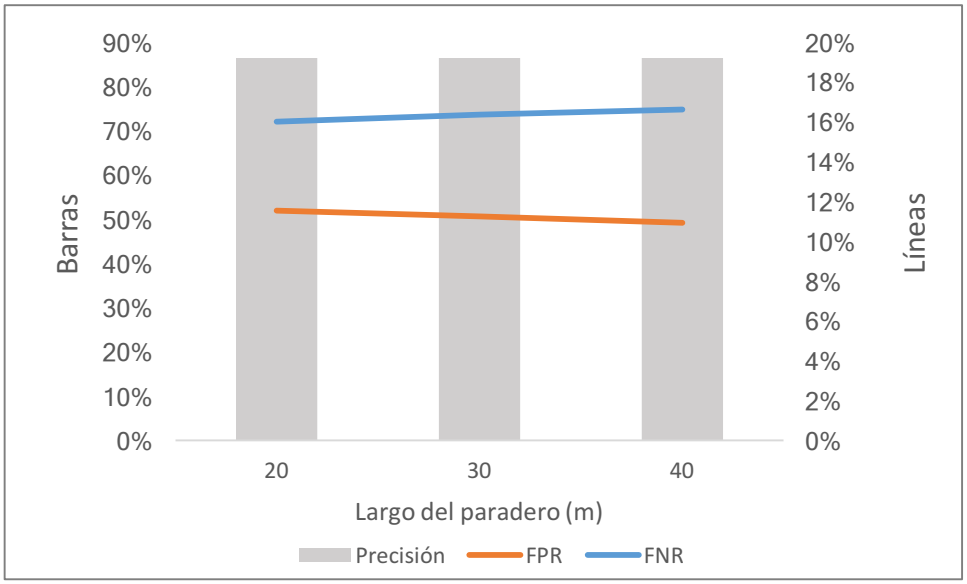

Figura 4.6 Valores de indicadores al variar largo de paradero

Fuente: Elaboración propia

Para los tres indicadores graficados, no se observan variaciones significantes en la Figura 4.6. Esto refleja que la variación respecto a este dato no produce cambios significativos al utilizar valores sensatos de largo de paradero. Esto es coherente con la formulación del modelo y los datos utilizados, ya que, al tener una periodicidad de 30 segundos la mayoría de las observaciones se encuentran en la categoría *Caso 1: No hay log GPS dentro del área de estacionamiento*. Este caso es menos restricto que los demás, por lo que el largo del paradero tiene menos importancia.

Tabla 4.10 Indicadores para distintos largos de paradero

Fuente: Elaboración propia

Largo del paradero	20 m	30 m	40 m
FPR	11,53%	11,24%	10,91%
FNR	15,98%	16,34%	16,60%
Precisión	85,9 %	86,0%	86,0%

La Tabla 4.10, muestra que el valor de FPR tiende a disminuir levemente a medida que el paradero aumenta su largo. Se puede concluir que el valor de este parámetro no produce cambios significativos en la predicción del modelo, por lo que mantenerlo en valores razonables no genera cambios en el comportamiento del modelo.

Ubicación del paradero

La ubicación de paradero se evalúa en tres escenarios posibles. La Figura 4.7 presenta el valor de los indicadores precisión, FPR, y FNR para cada uno de ellos.

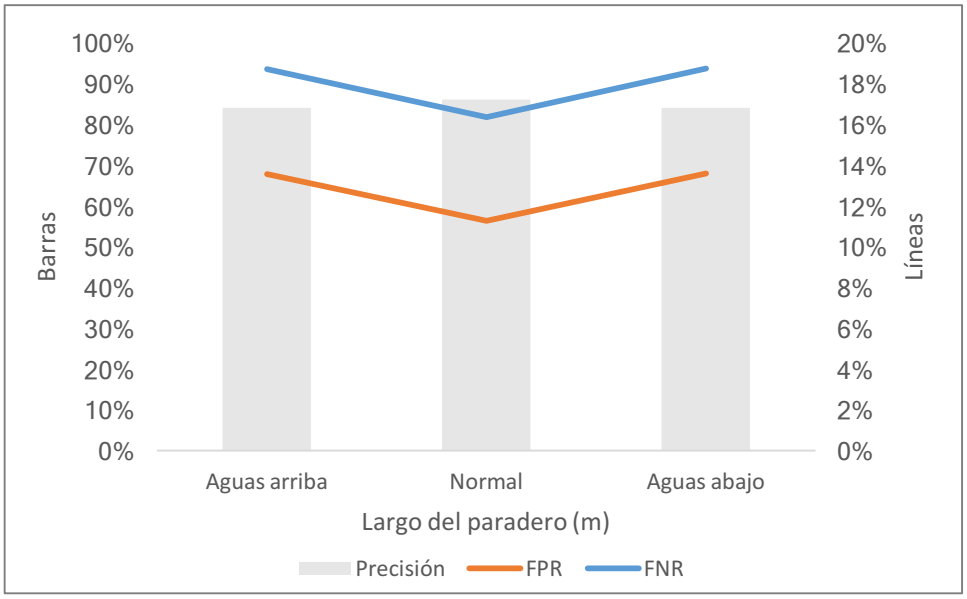

Figura 4.7 Valores de indicadores al variar la ubicación del paradero

Fuente: Elaboración propia

Como es de esperar, los mejores resultados se presentan al ubicar al paradero en su punto de georreferenciación. La precisión es de 85,98% en este caso, cerca de dos puntos porcentuales más respecto a las otras ubicaciones. Esto refleja que sí existe un efecto en la ubicación del paradero, pero no se considera substancial.

Los escenarios propuestos son casos extremos. Un error de 40 metros en la ubicación del paradero es exagerado, pero puede ocurrir, y exigirle a cualquier modelo predecir si los buses se detienen o no en paraderos que no están ubicados de forma correcta dificulta el proceso. No obstante, para datos GPS cada 30 segundos, el modelo no es exigente acerca de la exactitud de la ubicación del paradero.

4.2 Modelos Basados en aprendizaje supervisado

En este subcapítulo se presenta cómo se construyen los modelos basados en aprendizaje supervisado. Como se hace alusión en la sección 3.3, para la construcción de estos modelos se utilizan dos métodos. A continuación, se presentan los parámetros y las variables predictivas que utiliza cada modelo, junto a los resultados en base a los datos destinados a la clasificación. Al igual que en la sección anterior, tanto la calibración como la clasificación se realizan con la Base de datos 1, utilizando los datos reservados para cada etapa.

4.2.1 Máquinas de vectores de soporte (SVM)

Parámetros

Las máquinas vectores de soporte tienen dos parámetros principales. Estos son el valor de C, y la función Kernel a utilizar que se definen en el capítulo 2. Para evitar un sesgo, se utiliza una búsqueda exhaustiva utilizando valores de C desde 1 a 10, y variando la función Kernel entre la función lineal, polinomial, y gaussiana. Se utilizan estas tres funciones porque son las más comúnmente utilizadas (Ding & Liao, 2016).

Variables predictivas

En el proceso de calibración se realiza una búsqueda exhaustiva variando las variables predictivas involucradas (ver tabla 3.7). Una vez realizada la búsqueda exhaustiva respecto a estas variables y los demás parámetros, se seleccionan los modelos que presenten un FPR menor al presentado por el modelo Bull y Herrera, y de estos se elige el con mayor precisión. La Tabla 4.11 presenta las variables predictivas que utiliza el modelo elegido.

Tabla 4.11 Una cruz marca las variables predictivas que se utilizan en el modelo SVM

Fuente: Elaboración propia

VARIABLES	
Nºdatos	
Tipo	x
v_inicial	x
v_final	x
disTotal	x
v_promedio	
v_paradero	x
Xusa-Xu	x
Xd-Xdsa	x
Precision	93,16%
FPR	*4,83%*

Además de estas variables el valor de los costos asociados al Error y Tipo I y Tipo II, son 3 y 1 respectivamente. Se utiliza una función Kernel lineal, y un costo C (asociado ubicar un dato dentro de los vectores de soporte) igual a 2. Se observa que el valor de FPR se encuentra por debajo del mismo indicador respecto al modelo de Bull y Herrera, sin embargo, estos son datos de entrenamiento y este valor se debe corroborar en el proceso de calibración.

Durante la calibración, se observa que los modelos que presentan los mejores resultados, cuentan todos con las variables **v_inicial** o **v_paradero**. Esto demuestra la

importancia de estas variables. Para observar qué tan predictivas resultan ser estas variables respecto a la detención del bus, la Figura 4.8 las grafica.

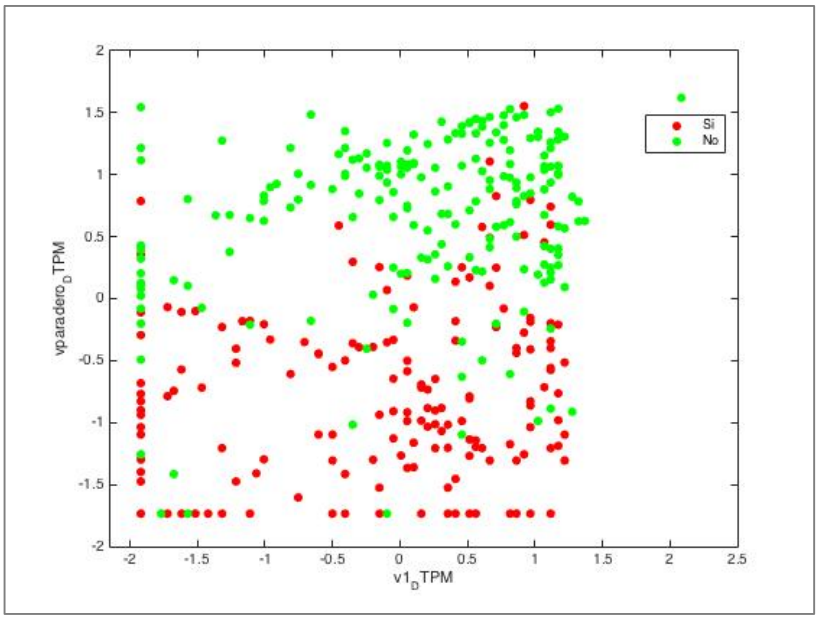

Figura 4.8 Base de datos 1 por clase al graficar las variables v_inicial y v_paradero
Fuente: Elaboración propia

Se observa que estas variables sí son capaces de esbozar una separación entre las dos categorías. Sin embargo, ésta no es clara, lo cual explica la inclusión de las otras tres variables.

Resultados

La Tabla 4.12 presenta un resumen de los resultados, que permite validar el modelo con datos que no fueron utilizados en la calibración.

Tabla 4.12 Indicadores modelos SVM – Clasificación

Fuente: Elaboración propia

INDICADOR	VALOR
Precisión	89,19%
FPR	11,11%
FNR	10,53%

La Tabla 4.12 muestra que el modelo fue capaz de determinar con una precisión del 89,19% las detenciones y no detenciones. Esta es una precisión menor a la que presenta en la etapa de calibración. Para observar qué datos predice bien el modelo y qué datos clasifica de forma errónea, se grafican los resultados de la clasificación en la Figura 4.9.

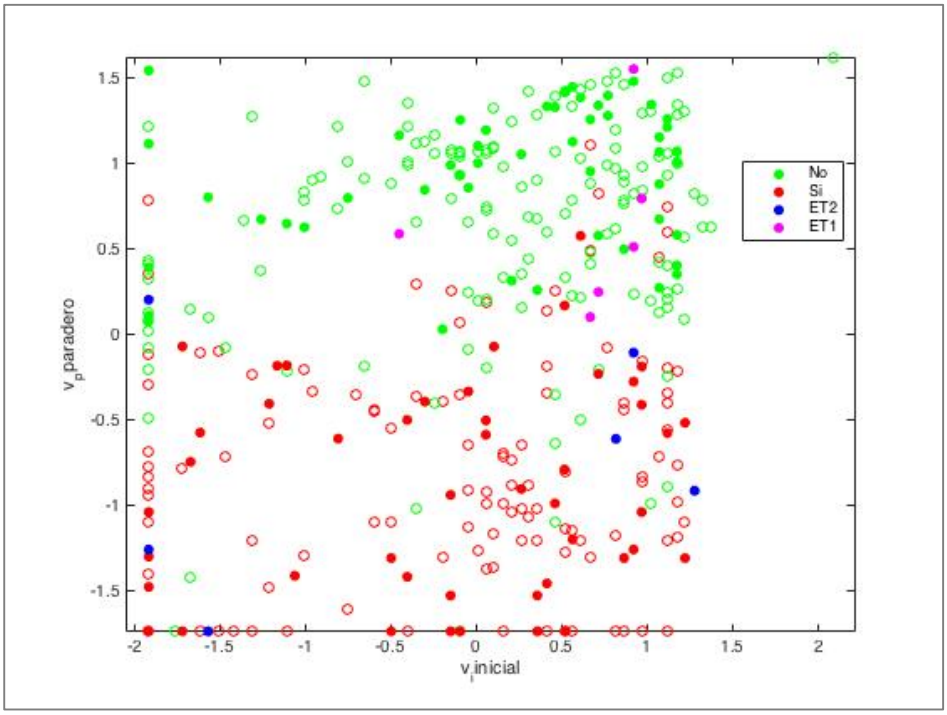

Figura 4.9 Datos de la clasificación del modelo SVM sobre los datos de entrenamiento

Fuente: Elaboración propia

La Figura 4.9 grafica las observaciones en base a las variables que parecieran ser más predictivas, que son **v_inicial** y **v_paradero**. La figura muestra con círculos vacíos los datos con los que se entrena el modelo, mientras que con círculos rellenos los datos que se reservan para la clasificación. Dentro de estos últimos, los de color verde y rojo han sido bien clasificados, mientras que los de color rosado y azul representan clasificaciones mal realizadas. Los datos rosados representan observaciones de buses que sí se detuvieron, y el modelo los clasifica como no detenciones, mientras que los datos azules representan el error contrario.

Estas clasificaciones mal realizadas se encuentran en el área del gráfico donde predomina la otra clasificación, y por esto el modelo no es capaz de predecir de forma correcta su categoría.

4.2.2 Análisis de Discriminante Lineal (LDA)

Calibración

El Análisis de discriminante lineal genera nuevas variables de tal forma de maximizar la separación entre categorías, que en este caso son "Detención" y "No detención". Como máximo se pueden generar $n-1$ variables nuevas, siendo n el número de categorías. Como en este caso $n=2$, se crea una sola variable que caracterice las dos categorías. Además del número de variables, este método no posee parámetros, por lo que su calibración consiste solamente de la selección de variables predictivas a utilizar y de los costos asociados al Error Tipo I y Tipo II.

Al igual que el modelo SVM, se realiza una búsqueda exhaustiva variando los parámetros. El valor de los indicadores en la etapa de calibración se calcula utilizando la validación cruzada, dividiendo los datos de calibración en cinco partes. La Tabla 4.13 muestra las variables predictivas que utiliza el modelo a utilizar.

Tabla 4.13 Variables predictivas que utiliza el modelo LDA

Fuente: Elaboración propia

VARIABLES	
Nºdatos	
Tipo	x
v_inicial	x
v_final	x
disTotal	x
v_promedio	x
v_paradero	x
Xusa-Xu	x
Xd-Xdsa	x
Precision	93,16%
FPR	6,96%

La Tabla 4.13 muestra un comportamiento similar a los modelos SVM, donde el modelo utiliza casi todas las variables predictivas. Además de estas variables el valor de los costos asociados al Error y Tipo I y Tipo II, son 5 y 1 respectivamente.

Resultados

La Tabla 4.14 presenta un resumen de los resultados obtenidos a partir del proceso de clasificación.

Tabla 4.14 Indicadores modelos SVM – Clasificación

Fuente: Elaboración propia

INDICADOR	**VALOR**
Precisión	90,09%
FPR	7,41%
FNR	12,28%

Se evidencia un comportamiento similar al del modelo formulado con SVM, donde la precisión disminuye respecto a los datos de entrenamiento.

Para observar con mayor detalle qué datos predice bien el modelo y qué datos clasifica de forma errónea, se grafican los resultados de la clasificación en la Figura 4.10.

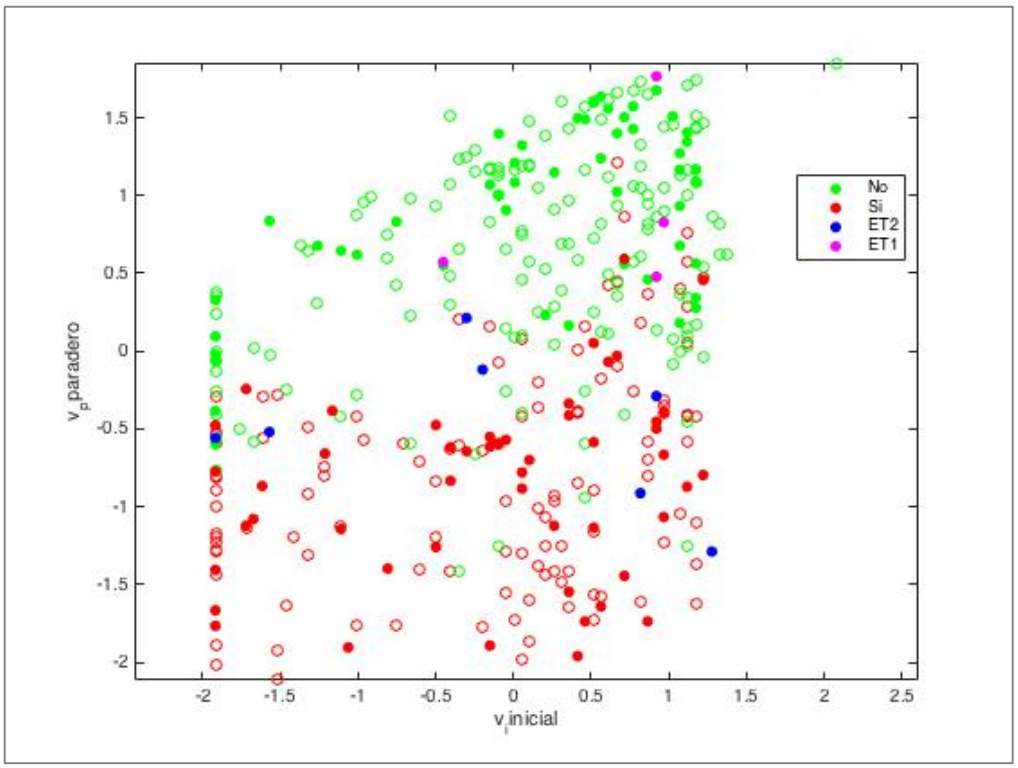

Figura 4.10 Datos de la clasificación del modelo LDA sobre los datos de entrenamiento

Fuente: Elaboración propia

A partir de la Figura 4.10, se puede observar un comportamiento similar al modelo SVM, donde los datos mal clasificados en este modelo también fueron mal clasificados en SVM. Sin embargo, este modelo incurre en un Error Tipo I menos.

4.3 Comparación de modelos

En esta sección se comparan el modelo de Bull y Herrera, ya que ninguna de las propuestas presenta mejoras respecto a la precisión y FPR, el modelo SVM y el LDA. La clasificación y comparación en esta etapa se realiza con el 30% de los datos de la *Base de datos 1*, por esta razón los resultados del modelo Bull y Herrera, varía respecto a la sección 4.1.

En la Figura 4.11 se muestran los indicadores obtenidos a partir de la clasificación.

Figura 4.11 Indicadores de los distintos modelos de predicción- Base de datos 1

Fuente: Elaboración propia

Los tres modelos a comparar presentan una precisión similar, en promedio tienen una precisión de 89,46%, con una desviación estándar de 0,0042. Esto demuestra que los modelos tienen una alta precisión, y sus valores son generalmente parecidos. Sin embargo, FPR y FNR difieren. El modelo que presenta el menor FPR es el modelo LDA, tanto el modelo de Bull y Herrera como el SVM presentan un FPR de 11,11%.

Si se comparan los modelos basados en aprendizaje supervisado, estos presentan resultados similares en cuanto a precisión, sin embargo, se observa un *trade-off* entre el valor de FPR y FNR. Recapitulando, la Figura 4.11 muestra que el modelo de LDA presenta los mejores resultados respecto a la precisión y FPR.

Hay que tener en cuenta que los modelos basados en aprendizaje superviado están calibrados para la situación de la que se obtienen la Base de datos 1. Por lo que implementarlo en una red más grande puede traer distintos resultados. Además el aumentar el costo asociado al Erro Tipo I puede traer beneficios en la disminución de FPR que no se ve reflejado en este modelo en particular.

4.4 Análisis periodicidad datos GPS

Para ver el cambio que produce en los indicadores la variación de τ, se utiliza la Base de datos 2. Esta sección se divide en dos: la primera presenta los resultados del modelo basado en cinemática, mientras que la segunda presenta los resultados de los modelos basados en aprendizaje supervisado.

4.4.1 Modelo basado en cinemática

Debido a lo observado en la sección 4.1, el modelo generado por Bull y Herrera es el que presenta los mejores resultados. Es por esto que en esta sección se presenta un análisis de cómo se comporta el modelo al variar su periodicidad.

Debido a lo estipulado en la sección 4.1, el valor de los parámetros es: $v_{max} = 60, a = 1, h_1 = 5, y\ h_2 = 25$. La Figura 4.12 grafica la variación de la precisión, FPR y FNR, frente a este caso.

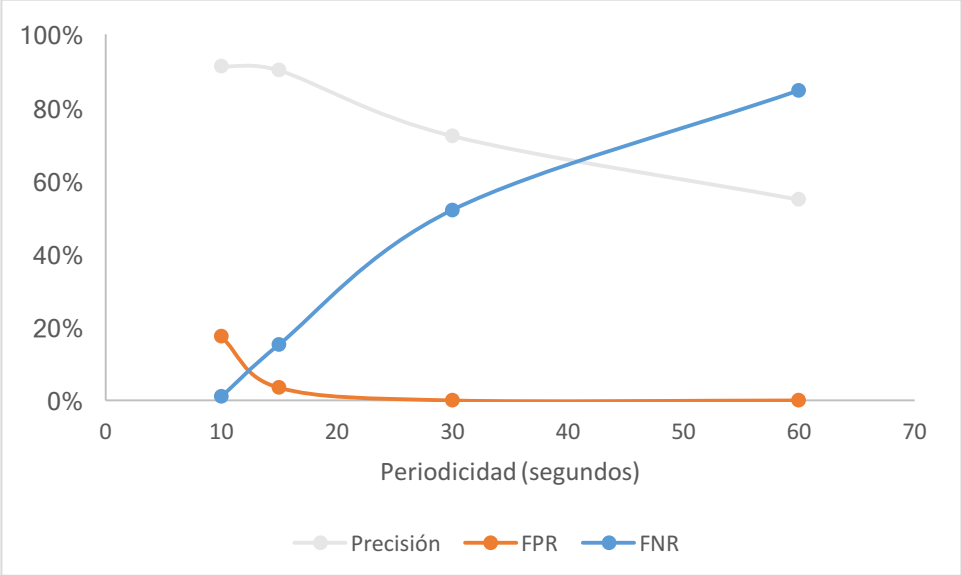

Figura 4.12 Valores de indicadores al variar espaciamiento – Modelo B&H

Fuente: Elaboración propia

A diferencia lo que se observa en la sección 4.1.1, se observa un cambio importante en el valor de los indicadores. Si se comparan los resultados de la validación del modelo con estos cuando $\tau = 30$, estos presentan valores significativamente distintos.

Estos resultados se deben a que la base de datos con la que se realiza la calibración y validación corresponde a buses que transitan por un corredor. Esto quiere decir que los buses transitan a mayor velocidad, ya que se enfrentan a menos detenciones y congestión. Como el modelo utiliza los metros recorridos entre log GPS, al recorrer más metros en el mismo tiempo, las observaciones tienden a parecerse a no detenciones.

Por otro lado, la Base de datos 2 corresponde a buses que transitan por vías exclusivas y vías compartidas durante distintos horarios del día. La velocidad de operación de estos buses es menor que la de la Base de datos 1, por esto el FPR tiende a ser cero, y el FNR aumenta.

La Figura 4.12 también muestra cómo disminuye significativamente el FNR cuando τ toma valores más pequeños. Sin embargo, cuando $\tau = 10$, el Error Tipo I

aumenta significativamente. A pesar de que FNR se encuentra en el valor más bajo observado, 3,7%, el valor de FPR se encuentra en su punto más alto, 18,6%.

Esto es razonable debido a cómo está construido el modelo, éste exige que en cada trio de datos se pueda realizar el proceso de detención completo. A medida que τ se hace más pequeño, se hace más difícil que un proceso de aceleración completo sea posible en τ.

Para $v_{max} = 60\ km/h$ y $a = 1\ m/s^2$, $\tau_a = 8{,}33\ s$. Cuando $\tau = 10$, el modelo espera que el bus haya avanzado lo equivalente a 1,67 segundos a velocidad máxima, esto es no más de 27,78 metros, para asegurar que el bus se pudo haber detenido. Las exigencias del modelo con τ más grandes aseguran el no acusar a buses que se detuvieron de no haberlo hecho. Sin embargo, para valores de τ más pequeños, estas restricciones aseguran lo contrario. La Tabla 4.15 refleja lo estipulado, donde el modelo tiende a fallar en los Casos 2 y 3, cuando el modelo exige al menos un proceso completo de aceleración o detención.

Tabla 4.15 Observaciones y error por casos para τ =10

Fuente: Elaboración propia

	NO SE DETIENEN		**SI SE DETIENEN**	
	Porcentaje de casos	***FNR***	***Porcentaje de casos***	***FPR***
Caso 1	62,7%	0,4%	0%	0%
Caso 2	36,2%	12,5%	11,6%	38,3%
Caso 3	1,1%	82,6%	88,4%	14,7%
Total		***5,7%***		***17,4%***

Si comparamos la precisión de $\tau = 10$ y $\tau = 15$, estas son 89,33% y 90,08% respectivamente. Es decir, un cambio en la periodicidad de 5 segundos no genera mayores cambios en este indicador. Sin embargo, los valores de FPR y FNR se ven revertidos. Frente a este caso sería recomendable utilizar $\tau = 15$ a la hora de utilizar el

modelo de Bull y Herrera. No solo su precisión se encuentra en el valor más alto, sino que su FPR tiene un valor menor al 1%.

En resumen, el valor de τ que presenta mejores indicadores es 15 segundos. Este valor de τ reduce significativamente el Error Tipo 1 respecto a $\tau = 10$ y reduce el Error Tipo 2 respecto a $\tau = 20$.

4.4.2 Modelos basados en aprendizaje supervisado

4.4.2.1 Máquinas de vectores de soporte (SVM)

Por las razones mencionadas en el capítulo anterior, en esta etapa se realizan cuatro modelos utilizando el método SVM. Para cada uno de los modelos se utilizan distintas variables predictivas por la forma en la que se realiza la calibración. El resultado de este proceso se presenta en la Tabla 4.16, donde se muestra las variables utilizadas para cada uno de los modelos (las columnas representan el valor de τ).

Tabla 4.16 Variables predictivas utilizadas en los modelos SVM

Fuente: Elaboración propia

VARIABLES	10	15	30	60
N		X		X
Tipo	X	X	X	X
v_inicial	X	X	X	X
v_final	X	X	X	
disTotal		X	X	X
v_promedio				X
v_paradero	X	X	X	X
Xusa-Xu				X
Xd-Xdsa				

La Tabla 4.16 no muestra una clara relación entre las variables utilizadas y el valor de τ. Por otro lado, sí muestra una consistencia a la hora de utilizar la variable **v_inicial**

y **v_paradero**, lo que es coherente con lo visto en la sección 4.2. Para observar qué tan predictivas resultan ser las variables **v_inicial** y **v_paradero** respecto a la detención del bus, la Figura 4.13 grafica las variables para $\tau = 10$ mientras que la Figura 4.14 grafica las mismas variables para $\tau = 60$.

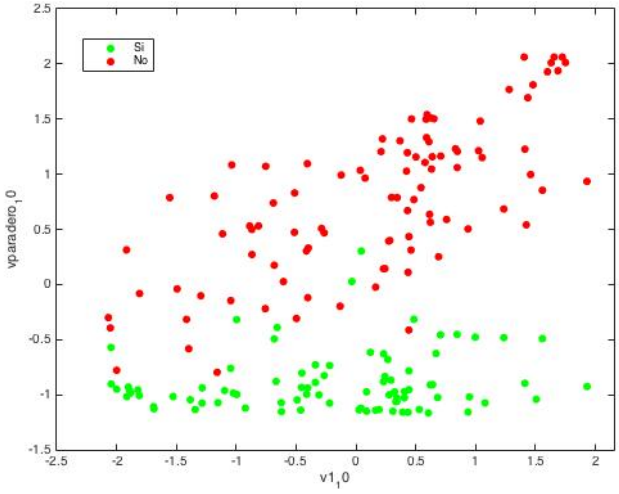

Figura 4.13 Variables v_inicial y v_paradero para τ=10

Fuente: Elaboración propia

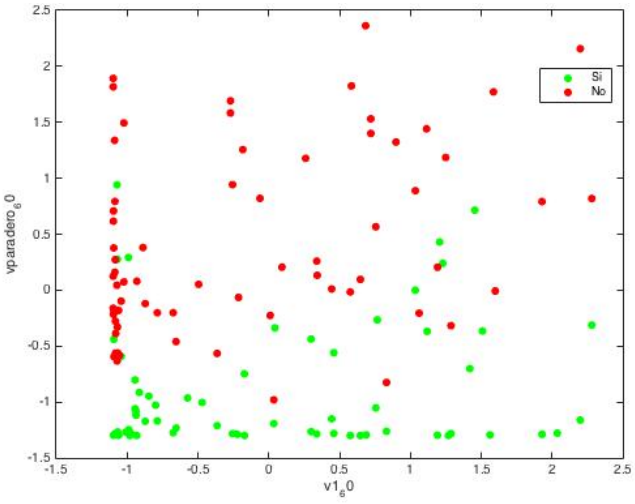

Figura 4.14 Variables v_inicial y v_paradero para τ=60

Fuente: Elaboración propia

En el caso $\tau = 10$ estas categorías son mucho más separables que cuando $\tau = 60$. Debido a esto, el modelo SVM con $\tau = 60$ requiere de un mayor número de variables para hacer la clasificación.

La Figura 4.15 presenta los resultados, que permite validar el modelo con datos que no fueron utilizados en la calibración.

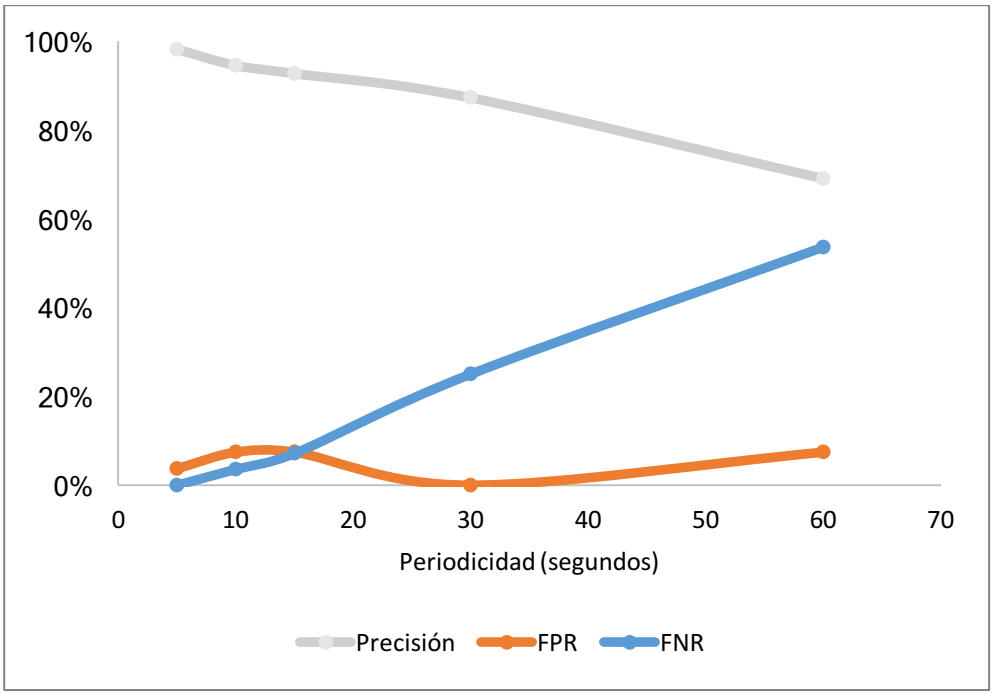

Figura 4.15 Valores de indicadores al variar espaciamiento – Modelo SVM

Fuente: Elaboración propia

La Figura 4.15 evidencia que, al igual que los modelos basados en cinemática, la precisión disminuye a medida que aumenta el valor de τ. Sin embargo, a diferencia de los modelos anteriores, no solo FNR aumenta, sino que también lo hace FPR en cierta medida. Esto ocurre porque el modelo busca hacer una distinción entre las variables de los buses que se detuvieron y de los que no, y a medida que τ aumenta las variables se hacen más parecidas entre las dos categorías, haciendo que las muestras no sean igual de separables que para el caso de $\tau = 10$.

En el caso de $\tau = 30$, los resultados muestran que el modelo fue capaz de determinar con una precisión del 81,82% las detenciones y no detenciones, valor que se encuentra por debajo del 89,19% que se obtiene en la sección 4.2. Esto ocurre por la misma razón que existe una baja en el modelo de Bull y Herrera, los datos pertenecen a otra naturaleza, donde las observaciones no son tan separables por categoría.

Para observar qué datos predice bien el modelo y qué datos clasifica de forma errónea, se grafican los resultados de la clasificación en la Figura 4.16.

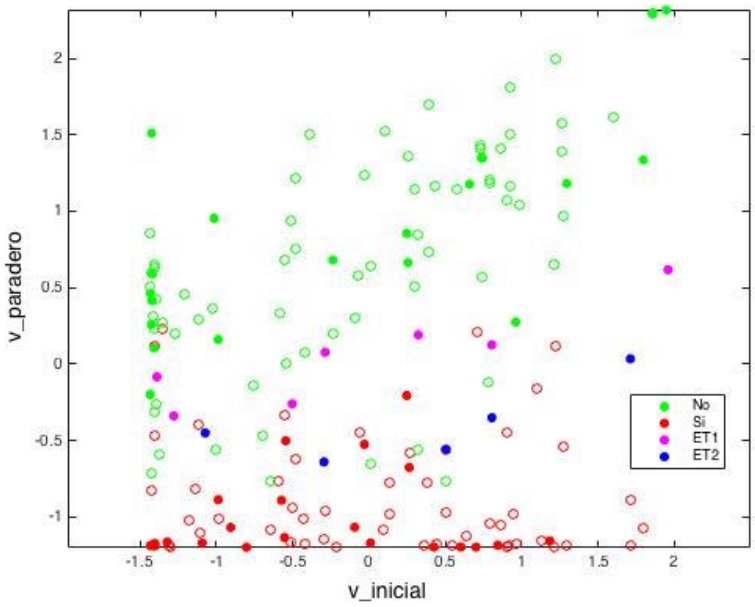

Figura 4.16 Clasificación para modelo SVM τ=30

Fuente: Elaboración propia

La Figura 4.16 muestra con círculos vacíos los datos con los que se entrena el modelo, mientras que con círculos rellenos los datos que se reservan para la clasificación. Dentro de estos últimos, los de color verde y rojo han sido bien clasificados, mientras que los de color rosado y azul representan clasificaciones mal realizadas. Los datos rosados representan observaciones de buses que sí se detuvieron, y el modelo los clasifica como no detenciones, mientras que los datos azules representan el error contrario.

Estas clasificaciones mal realizadas se encuentran en el área del gráfico donde las dos categorías se mezclan, y por esto el modelo no es capaz de predecir de forma correcta su categoría.

4.4.2.2 Análisis de Discriminante Lineal (LDA)

La Tabla 4.17 muestra el resultado del proceso de calibración para los cuatro modelos realizados.

Tabla 4.17 Variables predictivas utilizadas en los modelos de LDA

Fuente: Elaboración propia

VARIABLES	10	15	30	60
N		X		
Tipo	X		X	X
v_inicial	X	X	X	X
v_final		X		
disTotal				
v_promedio				
v_paradero	X	X	X	X
Xusa-Xu				X
Xd-Xdsa		X		

La Tabla 4.17 muestra un comportamiento similar a los modelos SVM, donde las variables **v_inicial** y **v_paradero** son consistentes en su aparición. La variable predictiva **Tipo** también aparece en casi todos los modelos, a excepción del modelo con $\tau = 15$. No obstante, este modelo utiliza la variable **N,** que está correlacionada con la variable **Tipo**.

La Tabla 4.18 presenta un resumen de los resultados obtenidos a partir del proceso de clasificación.

Tabla 4.18 Indicadores modelos LDA– Clasificación

Fuente: Elaboración propia

INDICADOR	10	15	30	60
Precisión	90,91%	87,27%	81,82%	69,09%
FPR	3,70%	3,70%	7,41%	7,41%
FNR	14,29%	21,43%	28,57%	53,57%

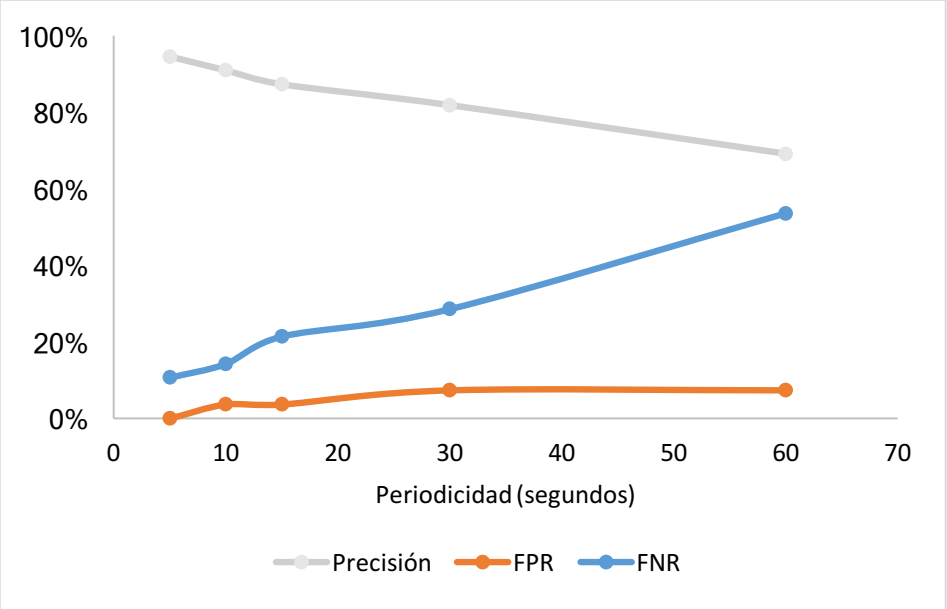

Figura 4.17 Valores de indicadores al variar espaciamiento – Modelo LDA

Fuente: Elaboración propia

Se evidencia el mismo comportamiento de los modelos anteriores, donde la precisión disminuye a medida que aumenta el valor de τ. Al igual que los modelos SVM, FPR y FNR aumentan a una mayor periodicidad.

Para observar con mayor detalle qué observaciones predice bien el modelo y qué observaciones clasifica de forma errónea, se grafican los resultados de la clasificación en la Figura 4.18, para cuando τ = 30.

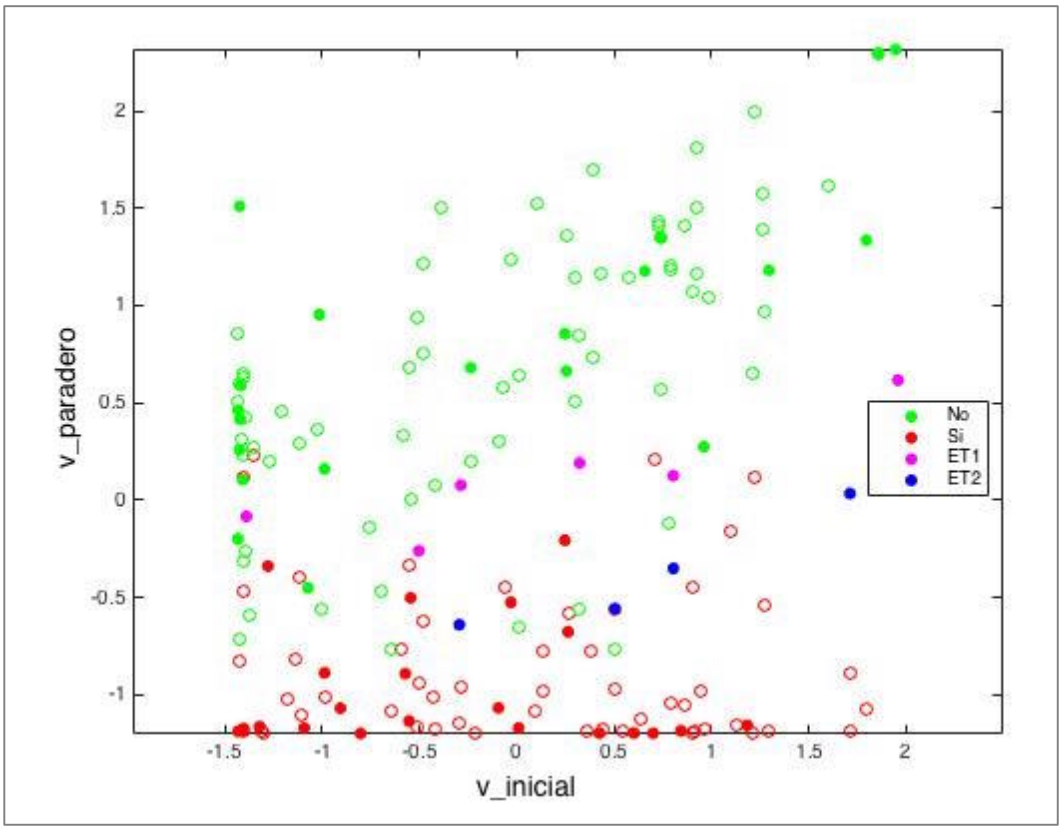

Figura 4.18 Clasificación para modelo LDA τ=30

Fuente: Elaboración propia

A partir de la Figura 4.18, se puede observar un comportamiento similar al modelo SVM, donde los datos mal clasificados en este modelo también fueron mal clasificados en SVM. Sin embargo, este modelo mejora en dos predicciones respecto al SVM, uno correspondiente a cada categoría.

5 CONCLUSIÓN

A partir de este estudio se pudo examinar y desarrollar distintos modelos de predicción, que buscan determinar la no detención de buses en paraderos mediante información GPS. Esto se logra mediante el uso de dos bases de datos distintas, la primera permite comparar los modelos utilizando datos del sistema de buses del transporte público de la ciudad de Santiago. Mientras que la segunda permite validar los modelos y ver cómo se comportan con distintos espaciamientos de datos.

Este capítulo se divide en dos partes. Primero se presentan las principales conclusiones respecto a los objetivos de este trabajo. Segundo, se propondrán extensiones o posibles líneas de trabajo futuro.

5.1 Respecto a los objetivos

El modelo de Bull y Herrera (2013) presenta resultados satisfactorios en ambas bases de datos utilizadas. Esto sugiere que su utilización en otras instancias podría entregar indicadores igual de satisfactorios. Respecto a sus parámetros, el modelo parece no ser sensible al valor de éstos cuando se encuentran en valores razonables. La velocidad máxima, la aceleración, el largo de paradero y su ubicación no presentan cambios significativos en los indicadores. El único parámetro que presenta mayores cambios en la precisión y demás indicadores es la periodicidad de datos GPS. Sin embargo, en la práctica, este no es un parámetro a calibrar del modelo, sino más bien un dato que se ha decidido mediante negociaciones y contratos entre los operadores de los buses y la entidad supervisora, que en el caso de Santiago es el DTPM.

En relación a las debilidades del modelo, se observan fallas cuando hay log GPS en el área de estacionamiento. Los casos en que falla el modelo se tratan de solucionar mediante distintos puntos. Primero se considera el caso en que la magnitud de la aceleración y la desaceleración no tengan los mismos valores. Luego se propone el uso de la distancia entre log GPS como restricción. También se incursiona en la

implementación de la velocidad instantánea. Finalmente se limitan menos las restricciones propuestas en el modelo original. Ninguno de estos cambios genera un cambio sustancial en el rendimiento del modelo. Aunque algunos si disminuyen el FNR (Error Tipo II), lo hacen a costa de un aumento del FPR (Error Tipo I), lo cual es altamente indeseable.

Los modelos basados en aprendizaje supervisado presentan buenos resultados respecto a los indicadores. Tanto el modelo formulado con el método SVM, como el formulado con LDA presentan altas precisiones. Ambos modelos presentas precisiones superiores al modelo de Bull y Herrera al clasificar los datos de la Base de datos 1, sin embargo, la diferencia no es de más de dos puntos porcentuales.

En cuanto a la periodicidad de datos GPS, 15 segundos logra la mayor precisión de los distintos valores utilizados. Además, no solo el modelo de Bull y Herrera presenta buenos resultados con esta periodicidad, sino que los demás modelos también lo hacen. Cabe hacer notar, que, a una menor periodicidad, este modelo deja de funcionar de forma tan precisa, ya que el valor del Error Tipo I aumenta. Lo anterior no ocurre con los modelos basados en aprendizaje supervisado, donde una mayor frecuencia de datos produce un mejor resultado.

Por esto motivos se concluye que el uso de los tres modelos puede ser favorable, su elección va a depender del objetivo del modelador. El modelo de Bull y Herrera ya se encuentra calibrado y listo a implementar, mientras que los modelos basados en aprendizaje supervisado requieren de una preparación de datos previa antes de ser implementados. Además, estos últimos, al ajustar el costo del Error Tipo I pueden llegar a ser más robustos respecto a este.

5.2 Extensiones

Esta investigación tiene distintas extensiones que se pueden realizar, comenzando por el uso de tecnología fuera del espectro GPS. Por ejemplo, en el caso de Santiago, se podría complementar este modelo con la información de los boletos del sistema de

buses, las “Bips”, donde un boleto emitido puede representar una detención en el paradero, si es que no existe un boleto emitido se puede predecir utilizando los GPS. Para esto se debe hacer un pareamiento entre la hora que pasa el bus por el paradero, con la hora que se emite un boleto o se genera una “Bip” en el bus. Otra extensión posible es implementar tecnología no existente en el sistema. Por ejemplo, el uso de cámaras, el enviar información GPS al momento de alcanzar una velocidad cero o al enfrentarse a cambios bruscos de aceleración

De igual forma, sería interesante desarrollar modelos para distintos escenarios, alguno que funcione en horarios y zonas congestionadas, mientras que otros calibrados con otros datos logren representar a los buses en horarios nocturnos o calles descongestionadas. Esto puede sugerir un modelo de Big Data, donde el modelo de predicción no funcione igual para todas las observaciones, sino que sea definido para cada paradero. Es más, su detalle puede ser aún más, mejorando la predicción haciendo diferenciación por horario y día de la semana. Sin embargo, el costo de esto es alto, ya que hoy en día no se cuenta con la información para generar este tipo de modelo.

Por último, para los modelos basados en aprendizaje supervisado se pueden implementar técnicas nuevas que están siendo desarrolladas, que un ahondamiento más profundo en su investigación podría traer beneficios para esta clasificación. Lo mismo puede ocurrir con otros sistemas de aprendizaje supervisado que no fueron incluidos en esta investigación.

BIBLIOGRAFÍA

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, julio). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

Bull, O., & Herrera, J.C. (2015). Using GPS data to identify buses skipping formal stops. *Thredbo 14*. Santiago.

Carmona-Suárez, E. J. (2014). Tutorial sobre Máquinas de Vectores Soporte (SVM). *Dept. of Artificial Intelligence, ETS of Computing Engineering, Universidad Nacional de Educación a Distancia (UNED), Spain*, 1-25.

Cawley, G. C. (2006, julio). Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on* (pp. 1661-1668). IEEE.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Ding, L., & Liao, S. (2016). An Approximate Approach to Automatic Kernel Selection.

Directorio de Transporte Público Metropolitano. (2015). *Gráficos, tablas y anexos Informe de Gestión 2015*. Recuperado el 31 de marzo de 2017, de DTPM - Informe de gestión: http://www.dtpm.cl/index.php/gestion2

Directorio de Transporte Público Metropolitano. (2016). *Gestión de Requerimientos*. Recuperado el 17 de enero de 2017, de DTPM: http://www.dtpm.cl/index.php/gestion-requerimientos

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Izenman, A. J. (2013). Linear discriminant analysis. *Modern Multivariate Statistical Techniques* (pp. 237-280). Springer New York

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Transantiago. (5 de febrero de 2016). *Autoridades de Transportes informan sobre baja en número de conductores de Transantiago multados por no detención en paradas*. Recuperado el 12 de octubre de 2016, de Transantiago: http://www.transantiago.cl/noticias/autoridades-de-transportes-informan-sobre-baja-en-numero-de-conductores-de-transantiago-multados-por-no-detencion-en-paradas;jsessionid=gMvGW0YIBXTSvJGiJKxjrvDX

Vapnik, V., & Chervonenkis, A. (1964). A note on one class of perceptrons. *Automation and remote control*, *25*(1), 103.

A N E X O S

ANEXO 2.1 – CASOS 2 Y 3 – MODELO BULL Y HERRERA 2015

Caso 2 – Un log *GPS en el área de estacionamiento*

Este caso sigue la misma lógica del Caso 1, donde se analiza en un principio si la detención completa se puede realizar entre Para analizar este caso utilizaremos la Figura 5.1. Sea X_1 la ubicación del dato GPS dentro del área de estacionamiento. Algo parecido ocurre con los tiempos y distancias entre datos *GPS*, que se clasifican como τ_{ij} donde i y j, representan el log de origen y destino.

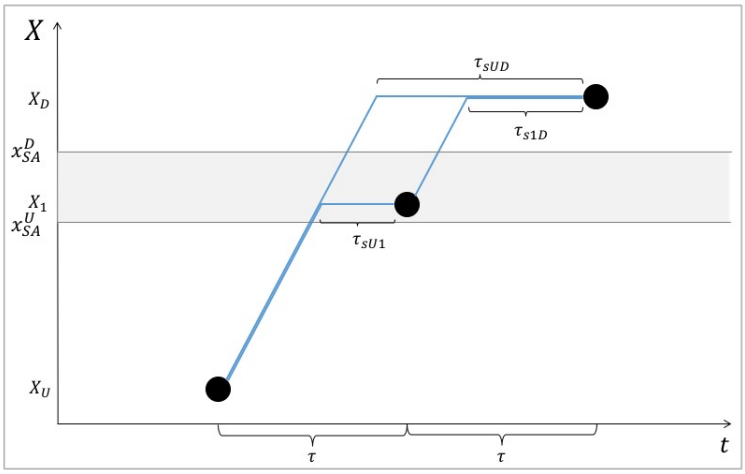

Figura 5.1 Ejemplo - Caso 2
Fuente: Elaboración propia

El primer acercamiento es utilizar $\tau_{sUD} \geq 2\tau_a$, olvidarse del log dentro del área de estacionamiento y solo utilizar los que se encuentran fuera de ella. En este punto surgen dos casos: si la parada se puede realizar de forma completa entre estos dos log o si no se puede. A continuación, se explica que proceso se realiza en cada caso.

- $Si\ \tau_{sUD} \geq 2\tau_a$

Esta restricción establece que el bus se pudo haber detenido entre X_U y X_D. Sin embargo, la información adicional que nos entrega el dato en X_1 nos permite predecir si esta detención ocurre dentro o fuera del área de estacionamiento.

La Figura 5.2 muestra un ejemplo del caso en que ocurre una detención fuera del área de estacionamiento. La línea punteada refleja la trayectoria real del vehículo, mientras el círculo rojo representa el lugar donde ocurre la detención.

La demora τ_{sUD} es lo suficientemente grande para que ocurra una detención entre X_D y X_U. Sin embargo, la ubicación del log GPS dentro de SA, aporta información que permite asegurar que la detención ocurre fuera del área de estacionamiento.

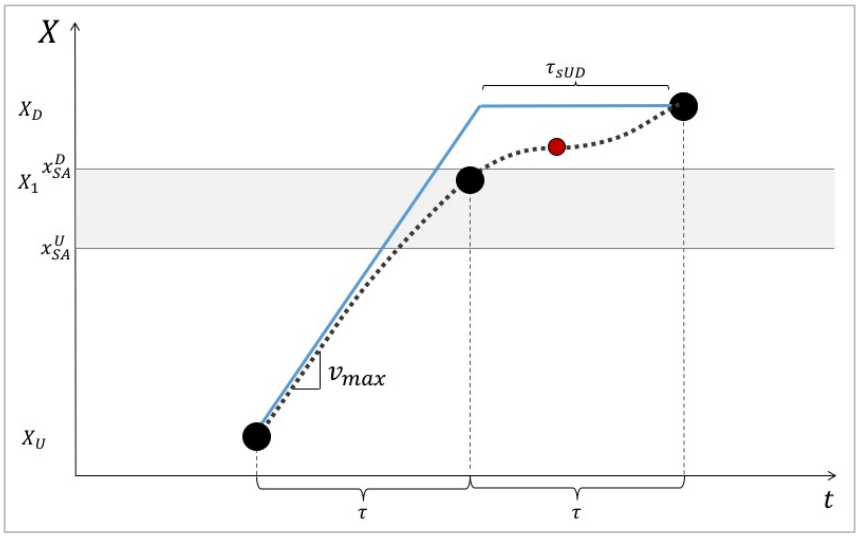

Figura 5.2 Ejemplo 2 - Caso 2
Fuente: Elaboración propia

Este tipo de casos, son percibidos cuando τ_{s1D} o τ_{sU1} son lo suficientemente pequeños. En el caso de la Figura 5.2, el valor de τ_{sU1} sería muy pequeño, ya que casi todo el trayecto se recorre a una velocidad v_{max} o cercana a v_{max} .

Como el modelo es exigente, los autores se sitúan en el caso extremo donde comienza a existir la posibilidad que la detención ocurra dentro del área de

estacionamiento. Estos casos son cuando la detención ocurre en los bordes del área de estacionamiento, o sea, X_{SA}^{U} y X_{SA}^{D}.

En el caso de que la detención ocurra en X_{SA}^{U}, la demora del tramo τ_{U1} debe contener un proceso completo de desaceleración y un proceso de aceleración hasta la ubicación de X_1. Este valor lo expresan los autores como $(2\tau_a)\gamma_U$, siendo γ_U la fracción máxima del proceso de detención contenida en el tramo τ_{U1}. Su valor está dado por la ecuación siguiente.

$$(2\tau_a)\gamma_U = \sqrt{\frac{2(X_1 - X_{SA}^{U})}{a}} - \frac{(X_1 - X_{SA}^{U})}{v_{max}} + \tau_a \tag{2.9}$$

En este caso, la demora mínima que debe tener el tramo τ_{1D} para que la detención ocurra dentro del área de estacionamiento es la diferencia entre $2\tau_a$ y la ecuación (2.9). La restricción asociada a este caso se formula como:

$$\tau_{s1D} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} \geq 2\tau_a \tag{2.10}$$

Esto es análogo a lo que sucede cuando la detención ocurre en X_{SA}^{D}, solo que es el tramo τ_{1D} el que contiene la mayor fracción de la detención. Juntando estos dos casos, se asume que un bus no realiza una parada dentro del área de estacionamiento si:

$$min\{(2\tau_a)\gamma_D;\ \tau_{s1D}\} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a \tag{2.11}$$

- ***Si $\tau_{sUD} < 2\tau_a$***

En este caso solo una fracción del proceso de detención ocurre en τ_{UD}. Los autores deciden relajar esta restricción, haciendo permisible que no todo el proceso de detención ocurra en este tramo.

Se realiza algo parecido a lo realizado en el punto 2.1.2 donde se asume que solo una fracción del proceso de aceleración ocurre en τ_{UD}. Al igual que en el punto 2.1.2, esta fracción consiste en la distancia entre los datos GPS fuera del área de estacionamiento y los bordes de ésta.

Sin embargo, los autores exigen que esta fracción del proceso de detención ocurra en τ_{sU1} o τ_{s1D}, y no en la suma de estos. Por ejemplo, si buscamos el valor mínimo que debe tener τ_{sU1} para que ocurra un proceso de detención en este tramo, éste

debe albergar una fracción del proceso de detención y todo el proceso de detención. Esto es análogo a si el proceso ocurre en τ_{s1D}.

Por lo tanto, si se cumplen las siguientes dos restricciones de forma simultánea, se asume que el bus no se logra detener en el área de estacionamiento.

$$\tau_{sU1} < \sqrt{\frac{2(X_{SA}^{U} - X_U)}{a}} - \frac{(X_{SA}^{U} - X_U)}{v_{max}} + \tau_a \tag{2.12}$$

$$\tau_{s1D} < \sqrt{\frac{2(X_D - X_{SA}^{D})}{a}} - \frac{(X_D - X_{SA}^{D})}{v_{max}} + \tau_a \tag{2.13}$$

Caso 3 – Más de un* log *GPS en el área de estacionamiento

Asumamos que existen n datos GPS dentro de SA, donde la ubicación de cada uno de ellos se denota por X_i, donde i puede tomar el valor desde 1 hasta n. Se sigue el formato del modelo planteado en la sección 2.1.2, donde el modelo busca si el bus se pudo haber detenido entre tres datos. El primer caso a analizar es X_U, X_1 y X_2; si el modelo concluye que el bus se pudo haber detenido entre estos datos, el bus se pudo haber detenido en SA. Si no es posible concluir esto, el modelo analiza los datos X_1, X_2 y X_3; y así, de forma iterativa hasta llegar a X_{n-1}, X_n y X_D. Si en este último trio de datos, el modelo concluye que el bus no se pudo detener, el modelo concluye que el bus no se detuvo.

Los autores diferencian los casos donde el trio de log GPS a analizar contienen un log fuera del área de estacionamiento, al caso en el que el trio de log se encuentre dentro del área de estacionamiento.

- **Dos log GPS dentro del área de estacionamiento y uno fuera**

En el caso de que fuera X_U el log GPS incluido en el trio a evaluar, los autores toman lo realizado en el punto 2.1.3; juntando la ecuación (2.12) y una variación de la ecuación (2.10).

$$\tau_{sU1} < \sqrt{\frac{2(X_{SA}^{U} - X_U)}{a}} - \frac{(X_{SA}^{U} - X_U)}{v_{max}} + \tau_a \ \wedge \ \tau_{s12} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a \tag{2.14}$$

En el caso en que fuera X_D parte de este trio, la restricción quedaría de la forma:

$$\tau_{snD} < \sqrt{\frac{2(X_n - X_{SA}^D)}{a}} - \frac{(X_n - X_{SA}^D)}{v_{max}} + \tau_a \ \wedge \ \tau_{s(n-1)n} + min\{(2\tau_a)\gamma_D;\ \tau_{snD}\} < 2\tau_a \quad (2.15)$$

Si el modelo llega a esta última restricción, y ésta se cumple, es la única forma de determinar que el bus no se detuvo en el paradero.

- **Tres log GPS dentro del área de estacionamiento**

Llamaremos a los log GPS X_{i-1}, X_i y X_{i+1}. La restricción que asume que el bus no se detuvo en este trio de datos, queda de la siguiente forma:

$$\tau_{s(i-1)i} + \tau_{si(i+1)} < 2\tau_a \quad (2.16)$$

ANEXO - PROGRAMACIÓN DEL MODELO

Este modelo, al igual que la preparación de los datos, se realiza en C# mediante Visual Studio. Se pensó realizar el programa de forma que fuera fácil y útil para la DTPM utilizarlo, por lo que el programa está pensado para recibir reclamos sobre no detención de buses y corroborarlos mediante los datos GPS de Transantiago.

Como dato de entrada, el programa recibe la lista de reclamos, datos con los que también cuenta la DTPM. Esta lista se presenta como un archivo de texto, donde cada línea contiene: Código de servicio, dirección, patente, hora, fecha, y paradero.

En el caso de la validación, este archivo de entrada tiene 184 líneas, cada una representa una de las observaciones realizadas para la *Base de datos 1*, con fecha, hora, y paradero correspondiente. Las primeras 86 representan las detenciones y las últimas 98 representan las no detenciones.

El programa selecciona la primera línea del archivo y a partir de la patente, fecha y hora, selecciona los datos GPS abarcando un rango de 20 minutos. Esto quiere decir que si el archivo indica que la observación ocurre a las 8:10, el programa selecciona los datos GPS desde las 8:00 hasta las 8:20 para el bus correspondiente. Esto se realiza porque, en general, la hora del reclamo no concuerda de forma exacta con la hora del GPS del bus. Además, a partir del código de servicio y paradero, extraídos del archivo de entrada, el programa obtiene la distancia del paradero de la base de datos **Paradas**. A este valor se le llama **distParadero.**

Luego, el programa le entrega valores a cada uno de los parámetros. Primero, selecciona la periodicidad entre datos GPS, τ. Una vez realizado esto, divide la base de datos en *sets*, donde en cada *set* los datos están separados cada τ segundos, por lo que ningún dato se repite en los *sets*. Luego, le entrega un valor al largo del paradero, asignando un valor a h_1 y h_2, siguiendo por asignar un valor a α, para finalizar otorgándole un valor a v_{max}.

A partir del valor de distParadero y de los valores de h_1 y h_2, el programa selecciona X_U y X_D, que son los datos justo antes y justo después de la ubicación del

paradero. Si existen, también selecciona los datos ubicados dentro del área de estacionamiento.

Una vez que tiene estos datos, se ingresan al modelo de Bull y Herrera. Una clase de C#, llamada **Comparar**, realiza las restricciones correspondientes al caso. En el archivo de salida se escribe el valor de los parámetros y el número del *set* de datos que se evalúa. Luego escribe "Si", si el modelo indica que bus se pudo haber detenido, o "No" si indica lo contrario.

Luego de realizado lo anterior, se itera el valor de los parámetros v_{max}, a, y largo de paradero, de tal forma que todos los casos sean analizados. Una vez evaluado todos los casos, se procede al siguiente *set* de datos. Cuando se evalúan los τ *sets*, se procede a la siguiente separación de datos, y se realiza el proceso descrito nuevamente. Una vez terminado esto, se procede a la siguiente línea del archivo de entrada y se realizan todas las iteraciones nuevamente.

La Figura 5.3, muestra cómo funciona la clase **Comparar**, haciendo alusión a las fórmulas de la Tabla 5.1. Para mayor detalle de cómo funciona la programación de es este modelo, se puede acceder al Anexo computacional bajo la carpeta **Programa Modelo Bull y Herrera**.

Tabla 5.1 Resumen de restricciones Modelo Bull y Herrera

Fuente: Elaboración propia a partir de Bull y Herrera (2015)

CASO	RESTRICCIONES
CASO 1	$\Delta = \min\{X_D - X_{SA}^D;\ X_{SA}^U - X_U;\ L_a\}$ $\tau_s < \sqrt{2\Delta/a} - \Delta/v_{max} + \tau_a$ (1)
CASO 2	$\boldsymbol{Si\ \tau_{sUD} \geq 2\tau_a}$ $min\{(2\tau_a)\gamma_D;\ \tau_{s1D}\} + min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a$ (2) $\boldsymbol{Si\ \tau_{sUD} < 2\tau_a}$ $\tau_{sU1} < \sqrt{\frac{2(X_{SA}^U - X_U)}{a}} - \frac{(X_{SA}^U - X_U)}{v_{max}} + \tau_a$ (3) $\tau_{s1D} < \sqrt{\frac{2(X_D - X_{SA}^D)}{a}} - \frac{(X_D - X_{SA}^D)}{v_{max}} + \tau_a$ (4)
CASO 3	**Dos log GPS dentro del área de estacionamiento y uno fuera** $\tau_{sU1} < \sqrt{\frac{2(X_{SA}^U - X_U)}{a}} - \frac{(X_{SA}^U - X_U)}{v_{max}} + \tau_a\ \wedge\ \tau_{s12} + \min\{(2\tau_a)\gamma_U;\ \tau_{sU1}\} < 2\tau_a$ (5) $\tau_{snD} < \sqrt{\frac{2(X_n - X_{SA}^D)}{a}} - \frac{(X_n - X_{SA}^D)}{v_{max}} + \tau_a\ \wedge\ \tau_{s(n-1)n} + min\{(2\tau_a)\gamma_D;\ \tau_{snD}\} < 2\tau_a$ (6) **Tres log GPS dentro del área de estacionamiento** $\tau_{s(i-1)i} + \tau_{si(i+1)} < 2\tau_a$ (7)

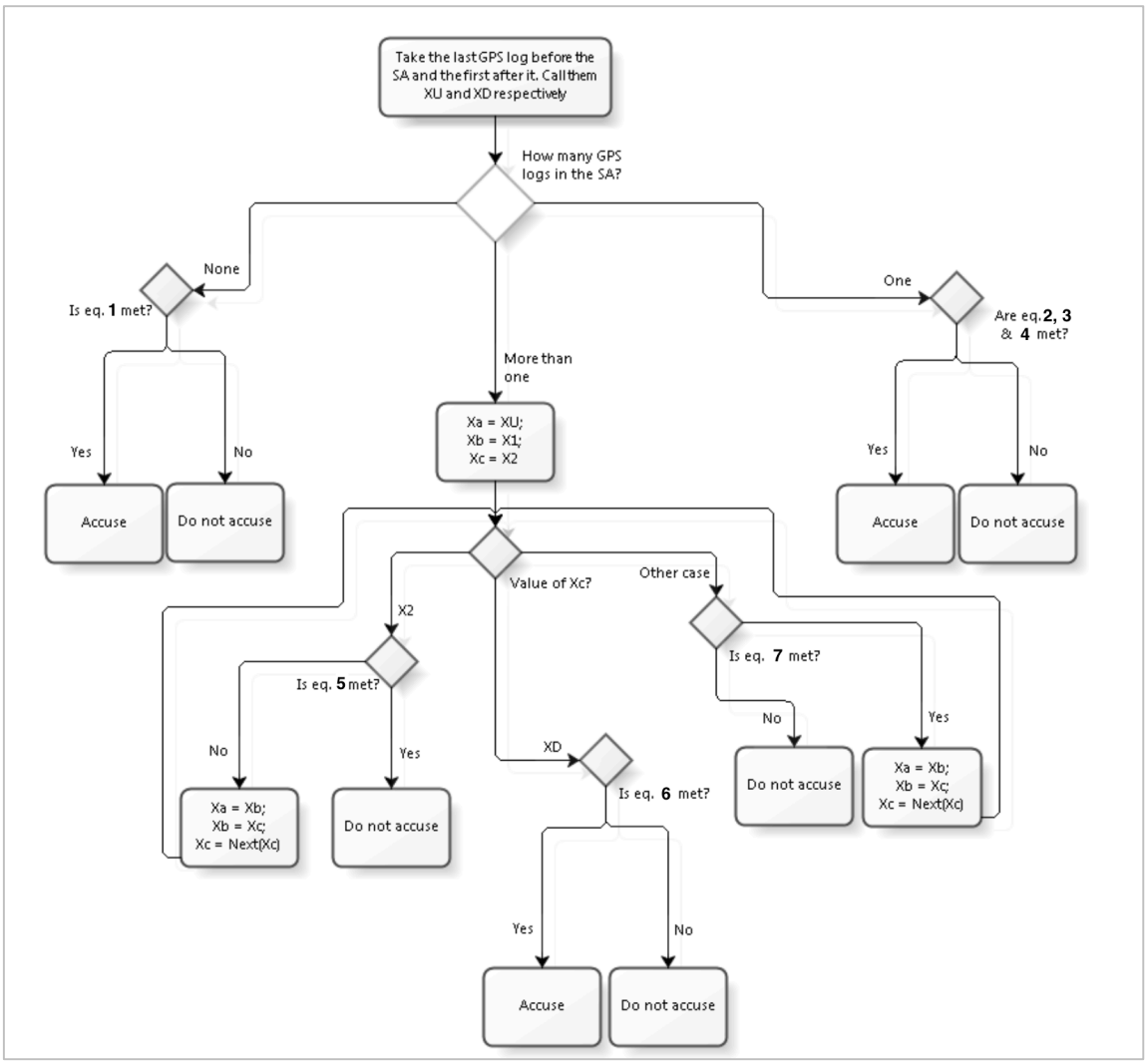

Figura 5.3 Funcionamiento Algoritmo Bull y Herrera

Fuente: Bull y Herrera 2015 – Con modificaciones

ANEXO 3.1.3 - PREPARACIÓN DE RUTAS Y PARADEROS

Los modelos a evaluar, requieren de un set de datos. En esta investigación, las variables a utilizar por los modelos son todo aquello que entregue la información GPS de los buses, ya que es la información accesible por la DTPM.

Para poder hacer el pareamiento de datos, entre los datos GPS y la variable de respuesta, se debe primero hacer un paso previo. Este consiste en construir las rutas de cada servicio, junto con los paraderos. Es importante realizar este paso, ya que genera una base a la que todos los GPS de los buses del mismo servicio están referenciados. Además, logra situar a los paraderos dentro de esta ruta, por lo que hace más fácil la comparación entre los GPS.

La Dirección de Transporte Público Metropolitano (DTPM) posee archivos geográficos, tanto, con la ubicación y forma de las rutas de los servicios de Transantiago, como la ubicación de sus paraderos. Estos fueron descargados de la web (DTPM, 2016) y transformados a archivos legibles para el programa gpAdmin III, que trabaja con bases de datos en SQL. Se decide trabajar con este programa, ya que, con un gran set de datos no disminuye su rendimiento. Además, es posible conectar este programa de forma local con otro tipo de softwares de programación, como Visual Studio, que se utiliza más adelante.

Estos archivos geográficos, transformados a una base de datos llamada **Rutas**, poseen: Latitud, longitud, código de servicio y dirección. Esta última variable, se refiere si es ida o retorno de ese servicio, ya que el punto de origen de la ruta y la distancia desde este, dependen de ello. Para cada par código-dirección va a representar una ruta, y cada punto de la ruta tendrá una distancia asociada.

Para realizar esto, se programa en C# a través de Visual Studio 2010, donde se crea una nueva columna para esta base de datos, llamada **distancia**. Esta variable refleja la distancia que se ha recorrido desde el punto de origen.

Como la base de datos está ordenada de tal forma que dos datos consecutivos, son geográficamente consecutivos. Para cada set de datos, separados por código-dirección, se elige el primer dato de la base de datos y se sitúa como el punto de origen. Esto

significa que su distancia es cero. Luego, el programa selecciona el segundo dato y obtiene la distancia euclidiana respecto al primero, y le asigna a la variable distancia este valor. Asimismo, se selecciona el tercer dato y se calcula lo recorrido desde el segundo. El valor de la variable distancia, resulta ser la suma entre el valor de la distancia del segundo dato, y lo recorrido entre estos dos. Avanzando, de forma iterativa hasta el último dato de este par código-dirección, formando el recorrido que realiza un bus en esa ruta. Para indagar en cómo se realiza este proceso, se puede revisar el programa que se encuentra en el Anexo Computacional, bajo el nombre MapMatching.

Luego de realizar esto, se sitúan los paraderos dentro de la ruta, para que así, posean una distancia desde el punto de origen. La base de datos de paraderos, llamada **Paradas**, tiene las siguientes variables: Código de servicio, sentido, código de paradero, latitud y longitud. Se debe hacer notar, que un mismo código de paradero, aparecerá el número de veces, que servicios se detengan en él, ya que la distancia asociada a su posición en la ruta, depende del servicio que se esté observando.

Estas ubicaciones geográficas podrían ser una limitación, ya que podrían existir instancias donde la ubicación real del paradero, no coincida con este dato. Debido a esto, se realiza un análisis acerca del efecto que tiene la ubicación geográfica del paradero, que se explica más adelante. No obstante, en todas las metodologías a utilizar, se entiende el paradero como una ubicación que posee un largo determinado, y que la georreferenciación es solo uno de los puntos en los que se encuentra el paradero.

Cabe decir, que la ubicación geográfica de los paraderos, que posee la DTPM, no coincide necesariamente con la ubicación geográfica de la ruta. Se debe entregar un valor de distancia al paradero, que refleje su posición en la ruta. Para esto, se debe encontrar el punto más cercano de la ruta, y asignarle al paradero la distancia asociada a este punto.

El punto más cercano, puede ser tanto un punto contenido en la base de datos, como un punto ficticio que se encuentre entre dos datos, como muestra la Figura 7.4. Los puntos GPS que generan la ruta de un servicio son representados en azul, mientras que los puntos que referencian la ubicación de los paraderos son anaranjados, mientras que la línea azul, representa la ruta de un servicio. En el caso del primer paradero, P1, el

punto más cercano de la ruta es un punto GPS. Por lo que asociarle una distancia a P1 resulta trivial, ya que es la misma distancia asociada al punto azul que demarca la ruta. En cambio, el segundo paradero, P2, resulta que su asignación a la ruta resulta en un punto ficticio, que no existe entre los datos que demarcan la ruta. Para poder asignarle una distancia al paradero, se le debe asignar una distancia al punto ficticio primero.

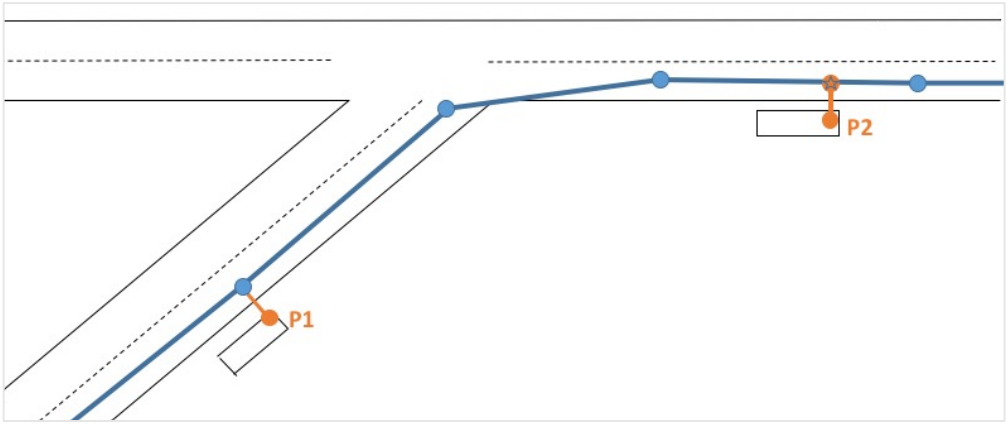

Figura 5.4 Referenciación de paraderos a la ruta

Fuente: Elaboración propia

Este proceso no resulta complicado, ya que, como el punto ficticio se encuentra en la línea recta entre dos puntos reales. Se calcula la distancia desde el punto ficticio al punto real que se encuentra antes de él, y el valor de la variable **distancia** resulta ser el valor de distancia del punto real, más la distancia calculada. Este proceso se puede ver en mayor detalle en el Anexo Computacional bajo MapMatching_Paradas.