UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE INGENIERÍA DE PRODUCCIÓN Y ORGANIZACIÓN EMPRESARIAL

MODELO GENÉRICO PROBABILÍSTICO PARA LA ESTIMACIÓN DE TIEMPOS DE SERVICIO EN UNA RUTA DE REPARTO.

Por:
Kevin Danglau Mejía Maldonado

INFORME DE PASANTÍA
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Ingeniero de Producción

Sartenejas, agosto 2018

UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE INGENIERÍA DE PRODUCCIÓN Y ORGANIZACIÓN EMPRESARIAL

MODELO GENÉRICO PROBABILÍSTICO PARA LA ESTIMACIÓN DE TIEMPOS DE SERVICIO EN UNA RUTA DE REPARTO.

Por:
Kevin Danglau Mejía Maldonado

Realizado con la asesoría de:
Tutor Académico: Gerardo Febres
Tutor Industrial: Aurelio De Pádua

INFORME DE PASANTÍA
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Ingeniero de Producción

Sartenejas, agosto 2018

MODELO GENÉRICO PROBABILÍSTICO PARA LA ESTIMACIÓN DE TIEMPOS DE SERVICIO EN UNA RUTA DE REPARTO.

Elaborado por: Kevin Danglau Mejía Maldonado

RESUMEN

Foxtrot Systems es una empresa norteamericana de tecnología aplicada a la logística de entrega urbana de bienes y servicios. La principal ventaja de Foxtrot es su algoritmo de sugerencia de secuencia de visita en tiempo real a través de un aplicativo móvil directamente manipulado por el conductor, el cual recibe el mejor orden de visita a ser realizada durante determinado momento para los clientes que tiene pendiente durante el día. Existe una limitación al no poder estimar correctamente el tiempo que tardará cada repartidor entregando en cada cliente en ruta, lo cual afecta la toma de decisión por parte del repartidor.

El presente proyecto estuvo enfocado en el diseño de un modelo estadístico apoyado en inteligencia artificial con el reconocimiento de patrones que permitan la estimación del tiempo de servicio para cada cliente en cada ruta a través del estudio de bases de datos de clientes de Foxtrot. Para realizar la inferencia se evaluaron los principales estimadores estadísticos para la base del cálculo.

Luego se establecieron las variables principales que afectan directa e indirectamente el tiempo de servicio, conversaciones con el personal de operaciones de nuestros principales clientes con visitas en operaciones de reparto en Colombia y Argentina. Se realizó un análisis de la arquitectura de base de datos actual encontrando algunas limitaciones en la captación de eventos de GPS asociados a un cliente y las variables disponibles para realizar pruebas con datos reales.

A partir de este análisis se definió, se probaron y compararon cuatro modelos diferentes.

1era versión: valor por defecto por centro de distribución basado en valor esperado de una distribución Lognorm. 2da versión del cálculo: Árbol de Decisiones considerando cantidad de producto. 3ra versión del cálculo: Árbol de Decisiones considerando cantidad de producto y conductor a realizar la entrega. 4ta versión del cálculo: Media Simple por cliente.

Los resultados de pruebas en 3 países diferentes entregando productos similares se obtuvo un error medio de aproximadamente 3 segundos por cliente y 5 minutos de desviación estándar, representando una mejora de 98% con respecto a la media de error de tiempos de servicio antes utilizados y de 59% con respecto a la desviación estándar. Este error depende del grado de sofisticación disponible debido a la madurez de los datos.

Por el contrario, para la distribución de panes en Pittsburgh los tiempos de servicio enviados por el cliente presentaron mejores resultados que las estimaciones, dando paso a un proceso de evaluación de datos de entrada para escoger si usar o no la predicción de Foxtrot.

Finalmente se definió una serie de lógicas que se adaptan a la base de datos actual para cada tipo de cliente y en cada fase de aprendizaje, dando las directrices para la construcción de un nuevo algoritmo.

Palabras clave: Tiempo de Servicio, Inteligencia Artificial, Partición Recursiva, Distribución Urbana, Planificación de Rutas.

AGRADECIMIENTOS

ÍNDICE GENERAL

RESUMEN iii
ÍNDICE DE FIGURAS ix
INTRODUCCIÓN 1
Antecedentes del Problema 1
Planteamiento del Problema 2
Justificación e Importancia del Proyecto 3
Objetivos Específicos 3
CAPÍTULO I ..... **Error! Bookmark not defined.**
MARCO EMPRESARIAL 4
1.1. Mapa Estratégico 4
1.2. Productos 4
1.2.1. Foxtrot SDK (software development kit) 4
1.2.2. App Mobile Android y iOS 5
1.2.3. Paneles de Monitoreamiento 5
1.2.3.1. Real Time Dashboard: 5
1.2.3.2. Inspector de Rutas: 6
1.2.4. Algoritmos de aprendizaje 7
1.3. Estructura Organizativa 7
CAPÍTULO II 9
MARCO TEÓRICO 9
2.1. Tiempo de Servicio (al cliente) 9
2.2. Logística Urbana 9
2.3. Modelo Probabilístico 9
2.4. Modelo Determinístico 10
2.5. Distribución de Weibull 10
2.6. Distribución Exponencial 11
2.7. Distribución Logarítmica normal 11
2.8. Ajuste a Distribución de probabilidad: 12
2.9. Método de Máxima Verosimilitud (MLE) 12
2.10. Bondad de Ajuste: 13

2.11. Tipos de pruebas utilizados en el presente informe: 13
2.11.1. Kolmogorov-Smirnov:..... 13
2.11.2. KSL TEST (Kolmogorov-Smirnov-Lilliefors):..... 14
2.11.3. Krammer Von mises test 14
2.12. Árbol de Decisiones 15
2.13. Partición Recursiva 15
2.13.1. Aprendizaje del Modelo de Partición Recursiva 16
2.14. Prueba Chi Cuadrado 16
2.14.1. Test para independencia estadística:..... 17
2.15. JMP (SAAS)..... 18
2.15.1. Modelo de Partición para crear árboles de decisión en JMP 18
2.15.2. Respuestas y Factores..... 19
2.15.3. Criterio de Partición JMP 19
2.16. Algoritmo 19
2.17. Computación en la Nube 19
2.18. Amazon Web Services (AWS)..... 20
2.19. SQL (Structured Query Language) 21
2.20. Global Positioning System (GPS) 21
2.21. Apuntamiento de Entrega en el Aplicativo 21
2.22. Tiempo de Parada Autorizada 22
CAPÍTULO III 23
MARCO METODOLÓGICO 23
3.1. Definición de Proceso de Entrega Ideal 24
3.2. Variables que Afectan, Limitaciones y Complejidad..... 24
3.2.1. Aleatoriedad de Variables 26
3.3. Base de datos actual y variables disponibles para análisis:..... 26
3.3.1. Limitaciones encontradas 27
3.4. Obtención y Limpieza de Datos (filtros para eliminar valores atípicos)..... 29
3.4.1. Tratamiento de valores atípicos de duración 30
3.5. Estimadores Estadísticos Base 31
3.6. Modelos y clientes para Analizar..... 32
3.7. Metodología de Evaluación y control 33

CAPÍTULO IV 35
RESULTADOS Y ANÁLISIS 35
4.1. Valores máximos considerados en los Datos: 35
4.2. Ajuste continuo para cada tipo de modelo por Centro de Distribución. 35
4.2.1. Distribución Lognorm como la que mejor se ajusta al proceso de distribución: 40
4.3. Árbol de Decisiones considerando cantidad de producto por Parada. 43
4.4. Árbol de Decisiones considerando cantidad de producto por Parada y conductor de camión. 47
4.5. Promedio de tiempo de parada histórico para cada cliente. 51
4.6. Evaluación y Comparación 52
4.6.1. Caso Pittsburgh 58
4.7. Ventaja de Errores Distribuidos normalmente: 60
4.8. Promedio Histórico: Error por Efecto látigo o Estacionalidad Comercial 61
4.9. Tabla comparativa Final 62
4.10. Flujograma de Decisiones para Implementación de Tiempo de Servicio Genérico: .. 63
CONCLUSIONES 65
RECOMENDACIONES 67
REFERENCIAS BIBLIOGRÁFICAS 68

ÍNDICE DE TABLAS

Tabla 1: Ejemplo de tipos de cliente y su tiempo de servicio. 42
Tabla 2: Valor Esperado de Tiempo de Servicio y probabilidad acumulada. 42
Tabla 3: Ejemplo para entender los beneficios de un error distribuido normalmente..... 61
Tabla 4: Comparación de resultados por Centro de Distribución. 63

ÍNDICE DE FIGURAS

Figura 1: Modelo de tiempo de servicio usado en algunas operaciones. 2
Figura 1.1: Aplicación Mobile Foxtrot.....**Error! Bookmark not defined.**
Figura 1.2: Dashboard en tiempo real .....**Error! Bookmark not defined.**
Figura 1.3: Ejemplo Inspector de Rutas .....**Error! Bookmark not defined.**
Figura 1.4: Estructura Organizativa.....**Error! Bookmark not defined.**
Figura 2.1: Ejemplo de Distribución normal empírica Kolmogorov-Smirnov ...**Error! Bookmark not defined.**
Figura 2.2: Ejemplo árbol de decisión en JMP.....**Error! Bookmark not defined.**
Figura 2.3: Ejemplo de Visualización de opciones de entrega en el aplicativo ..**Error! Bookmark not defined.**
Figura 3.1: Descripción General de Metodología .....**Error! Bookmark not defined.**
Figura 3.2: Proceso Genérico de Servicio al Cliente..... 24
Figura 3.3: Visita a operación de descarga en Argentina..... 25
Figura 3.4: Estatus de Entrega y Evento relacionado..... 27
Figura 3.5: Ejemplo de multiples entregas en la misma parada a través del Route Inspector 28
Figura 3.6: Lugar de parada para múltiples entregas en la ciudad de Rosario visto desde Google Street View. 29
Figura 3.7: Histograma y Boxplot de Tiempos de Servicio para Rosario, Guarulhos y Tarija..... 32
Figura 4.1: Distribución normal para el Tiempo de Servicio en Pittsburgh..... 36
Figura 4.2: Ajuste Normal doble para el Tiempo de Servicio en Guarulhos 37
Figura 4.3: Ajuste normal doble para el Tiempo de Servicio en Pittsburgh. 38
Figura 4.4: Ajuste a distribución Logarítmica Normal (Lognorm) para el Tiempo de Servicio en Guarulhos..... 38
Figura 4.5: Ajuste a distribución Logarítmica Normal (Lognorm) Rosario. **Error! Bookmark not defined.**
Figura 4.6: Ajuste a distribución Logarítmica Normal (Lognorm) Tarija. .. **Error! Bookmark not defined.**
Figura 4.7: Ajuste a distribución Logarítmica Normal (Lognorm) Pittsburgh..... 40
Figura 4.8: Distribución y tipos de clientes, arriba Guarulhos y abajo Pittsburgh..... 41
Figura 4.9: Árbol de Decisión por cantidad de paquetes para Guarulhos. 43

Figura 4.10: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes para Guarulhos. 44
Figura 4.11: Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Garulhos. 44
Figura 4.12: Árbol de Decisión por cantidad de paquetes para Rosario. 45
Figura 4.13: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes para Rosario. 45
Figura 4.14: Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Rosario. . 46
Figura 4.15: Árbol de Decisión por cantidad de paquetes para Tarija. 46
Figura 4.16: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes para Tarija. 47
Figura 4.17: Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Tarija. 47
Figura 4.18: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos. 48
Figura 4.19: Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos 49
Figura 4.20: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes y conductor para Rosario. 49
Figura 4.21: Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Rosario. 49
Figura 4.22: Árbol de Decisión por cantidad de paquetes y conductor para Tarija. 50
Figura 4.23: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes y conductor para Tarija. 51
Figura 4.24: Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Tarija. 51
Figura 4.25: Ejemplo de predicciones obtenidas por ID de cliente. 52
Figura 4.26: Distribución de errores por cliente para la industria de bebidas. 53
Figura 4.27: Distribución de Errores General por Cliente considerando la 4ta Versión de estimación (Media por cliente). 54
Figura 4.28: Distribución de Errores por Cliente para Guarulhos – de Izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3 55

Figura 4.29: Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Guarulhos. 55
Figura 4.30: Distribución de errores por cliente para Rosario – de izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3 56
Figura 4.31: Distribución de errores por cliente considerando la 4ta versión de estimación (Media por cliente) para Rosario 57
Figura 4.32: Distribución de errores por cliente para Tarija – de Izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3 58
Figura 4.33: Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Tarija. 58
Figura 4.34: Distribución de errores por cliente para Pittsburgh – En la Izquierda para los tiempos enviados por el cliente y a la derecha para la Estimación con la Versión 1. 59
Figura 4.35: Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Tarija. 60
Figura 4.36: Error medio vs número de visitas consideradas para el cálculo de la estimación - Guarulhos 62
Figura 4.37: Lógica de Algoritmo de Utilización y Aprendizaje de Tiempos de Servicio. 64
Figura A.1: Distribución normal para el Tiempo de Servicio en Guarulhos. 70
Figura A.2: Distribución normal para el Tiempo de Servicio en Rosario. 70
Figura A.3: Distribución normal para el Tiempo de Servicio en Tarija. 71
Figura A.4: Distribución Weibull para el Tiempo de Servicio en Guarulhos. 71
Figura A.5: Distribución Weibull para el Tiempo de Servicio en Rosario. . **Error! Bookmark not defined.**
Figura A.6: Distribución Weibull para el Tiempo de Servicio en Tarija. .... **Error! Bookmark not defined.**
Figura A.7: Distribución Weibull para el Tiempo de Servicio en Pittsburgh ..... **Error! Bookmark not defined.**
Figura A.8: Distribución Exponencial para el Tiempo de Servicio en Guarulhos. 73
Figura A.9: Distribución Exponencial para el Tiempo de Servicio en Rosario. . **Error! Bookmark not defined.**

Figura A.10: Distribución Exponencial para el Tiempo de Servicio en Tarija. .. **Error! Bookmark not defined.**
Figura A.11: Distribución Exponencial para el Tiempo de Servicio en Pittsburg. ..... **Error! Bookmark not defined.**
Figura A.12: Distribución normal doble para el Tiempo de Servicio en Rosario. ..... **Error! Bookmark not defined.**
Figura A.13: Distribución normal doble para el Tiempo de Servicio en Tarija.. **Error! Bookmark not defined.**
Figura B.1: Local de parada corta autorizada en Rosario..... 76
Figura B.2: Local de parada larga autorizada en Rosario..... 77
Figura B.3: Route Inspector de una ruta con un solo cliente (Walmart-Pittsburgh). 77
Figura B.4: Route Inspector en parada Walmart-Pittsburgh. 78
Figura B.5: Route Inspector en tienda de conveniencia en una estación de Gasolina en Pittsburgh. 78
Figura B.6: Route Inspector en parada en un mayorista en Guarulhos. 79
Figura B.7: Route Inspector en parada en un abasto en Guarulhos..... 79
Figura C.1.1: Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte superior izquierda)..... 81
Figura C.1.2: Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte superior derecha) 81
Figura C.1.3: Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte inferior)..... 81
Figura C.2.1: Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte izquierda)..... **Error! Bookmark not defined.**
Figura C.2.2: Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte central)..... **Error! Bookmark not defined.**
Figura C.2.3: Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte derecha). ..... **Error! Bookmark not defined.**
Figura D.1: Error medio vs número de visitas consideradas para el cálculo de la estimación - Rosario..... 83
Figura D.2: Error medio vs número de visitas consideradas para el cálculo de la estimación - Tarija..... **Error! Bookmark not defined.**

Figura D.3: Error medio vs número de visitas consideradas para el cálculo de la estimación - Pittsburgh. ..... **Error! Bookmark not defined.**

INTRODUCCIÓN

Foxtrot Systems es una empresa de soluciones tecnológicas creada en 2014, con sede en la ciudad de California en los Estados Unidos y sucursales en Ciudad de México y São Paulo, siendo esta última la ciudad donde será realizada la presente investigación. Foxtrot ofrece un sistema integrado de optimización de rutas de reparto de bienes y servicios para la mejora de eficiencia y nivel de servicio con resultados en reducción de km recorrido, reducción de tiempo en ruta y aumento en puntualidad. Actualmente existe una limitación al no poder estimar correctamente el tiempo que tardará cada repartidor entregando en cada cliente en ruta, siendo una variable importante dentro del algoritmo de toma de decisión, representa un problema en la optimización de secuencia de visita en cada ruta de reparto. El presente proyecto consistirá en el diseño de un modelo estadístico apoyado en inteligencia artificial con el reconocimiento de patrones que permitan la estimación del tiempo de servicio para cada cliente en cada ruta a través del estudio de bases de datos de clientes de Foxtrot.

En el Capítulo I de este informe se presentó el marco empresarial de Foxtrot Systems, explicando sus orígenes, productos principales e información relevante para el contexto del proyecto; en el Capítulo II se expusieron los conceptos básicos para la realización de este proyecto; en el Capítulo III se describió la metodología empleada para el diseño de los modelos de estimación de Tiempo de Servicio y por último, en el Capítulo IV se presentaron los resultados obtenidos y análisis correspondientes al modelo genérico.

Antecedentes del Problema

Uno de los grandes desafíos de la planificación y ejecución en ruta es la estimación de los tiempos de servicio en el cliente, actualmente los Softwares de planificación de rutas como RoadShow y Roadnet ofrecen un tiempo estimado, pero los métodos de cálculo no son sofisticados y terminan por ser escritos como un parámetro del cliente en su mayoría. Para Foxtrot el tiempo de servicio representa uno de los pilares fundamentales para mejorar la roterización en tiempo real los cuales son: horario de atención, ubicación y Tiempo de Servicio por cliente como características principales del cliente a ser atendido.

Algunos autores han intentado caracterizar procesos de parada de vehículos, como autobuses, considerando otros factores contribuyentes, como por ejemplo, tipo de vehículo, longitud de lugar de parada [1]. Es difícil encontrar algún estudio que busque caracterizar un proceso de entrega genérico, que sirva para varios clientes. Muchos de estos estudios previos consideran variables aleatorias que afectan el tiempo de parada, como muchos son de logística urbana publica, la mayoría son de transporte de pasajeros. El estudio de entrega de productos y servicio por lo general

en el mundo de la logística se ha venido realizando por muestreo y definición de parámetros por tipo de cliente a ser atendido, lo cual es poco dinámico y escalable.

Una idea interesante que es implementada por la mayoría de Software de planificación de ruta (Foxtrot no entra en esa categoría ya que su foco es la ejecución), fue la consideración de un modelo que determina un tiempo fijo y un tiempo variable, siendo el último representado a través de un parámetro lineal bajo el siguiente esquema:

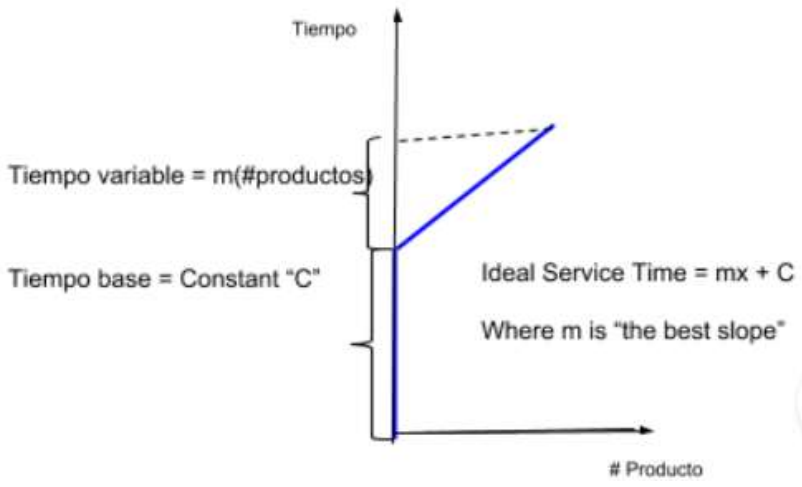

Figura 1 Modelo de tiempo de servicio usado en algunas operaciones. Fuente: Elaboración propia.

El problema cuando se realizó el análisis de esta metodología fue que el parámetro que representa la pendiente era difícil de estimar y no era replicable para diferentes tipos de productos, por lo tanto, no era posible incluir como parte del algoritmo automático, solo por medio de una entrada del parámetro manual, lo cual no representa el foco principal de una empresa de datos para Machine Learning como Foxtrot, donde estos valores deben generares automáticamente.

La idea de celulares, telemetría y datos GPS de camión de entrega ha disponibilizado en varias soluciones tecnológicas la capacidad de percibir tiempos de parada, sin embargo, los mismos no se han relacionado con la planificación dinámica en tiempo real de una ruta, considerando las condiciones del momento actual, limitando las soluciones del mercado local a dar como modelo más sofisticado, un tiempo de parada medio según lo visto históricamente.

Planteamiento del Problema

Existe una limitación al no poder estimar correctamente el tiempo que tardará cada repartidor entregando en cada cliente en ruta, lo cual no permite optimizar correctamente la secuencia de visita en cada ruta.

Justificación e Importancia del Proyecto

- **Precisión En los Tiempos de llegada a cada cliente:** La secuencia sugerida tomará en cuenta la previsión de llegada en cada cliente, sumando no solo un tiempo de conducción con alta precisión debido a información consumida de tránsito y factores en ruta, sino también a un tiempo de parada realista estimado para cada cliente.

- **Mayor precisión en el tiempo de ruta/jornada planificado:** los tiempos de parada por servicio a un cliente siempre fueron variables incontrolables o, a veces, inestimables, a pesar de que estos eventos representan la duración más relevante de la jornada de los conductores en comparación al tiempo de conducción durante la ruta.

- **Especificación del horario de atención:** Al comprender el tiempo de servicio adecuado de un cliente, la definición de su horario de atención y también la sugerencia de secuencia de visita de ese cliente dentro de la ruta en tiempo real será mucho más precisa, a fin de garantizar la puntualidad teniendo una previsión de cuanto se tardará en cada cliente anterior para llegar en la hora correcta.

- **Equilibrio en la carga de trabajo para los conductores:** se debe considerar el tiempo de servicio para equilibrar la carga de trabajo de la manera más eficiente, lo que podría conducir a la reducción de los gastos de sobretiempo y horas extra.

- **Identificación de tendencias, retrasos y comportamientos atípicos por clientes:** teniendo estos datos en el sistema, se pueden usar como inteligencia para caracterizar clientes y negociar nivel de servicio personalizado.

Por estas 5 razones principales es de suma importancia la correcta estimación del tiempo de servicio, representan implícitamente la reducción de las entregas fallidas y costos de logística.

Objetivo General

Diseñar y probar un modelo que permita la estimación del tiempo de servicio para cada ruta y cliente.

Objetivos Específicos

1. Adquirir conocimientos de la empresa
2. Recolectar información requerida
3. Analizar la información recolectada
4. Diseñar propuestas de modelos de inferencia de Tiempo de Servicio
5. Evaluar cada uno de los modelos.

CAPÍTULO I

MARCO EMPRESARIAL

Foxtrot Systems es una empresa de software como servicio (SAS) con sede en San Francisco nacida en el MIT y fundada en 2014, que se centra en potenciar el futuro de la cadena de suministro. A pesar de los avances en la tecnología de planificación de rutas y la rápida proliferación de la telemática de vehículos comerciales, el conductor de "servicio y entrega" actualmente permanece solo en el campo para lidiar con un entorno altamente disruptivo [2]. Solo se necesita un accidente, un obstáculo no planificado o una parada de último minuto para perder la planificación de una ruta completamente. Al aprovechar los avances en la computación en la nube, el aprendizaje automático y la optimización dinámica, Foxtrot se centra en potenciar las flotas de distribución de última milla para que superen sus objetivos y encuentren nuevas formas de ofrecer servicios diferenciados.

1.1. Mapa Estratégico

La empresa busca posicionarse como uno de los principales proveedores de tecnología de logística urbana en el mundo, optimizando rutas de reparto de cualquier tipo, tanto en servicios como productos, utilizando información para beneficiar a todos sus clientes.

Al ser un startup, la visión y misión aún no se encuentra exactamente definida, sin embargo se percibe una misión común de innovación para jugar un papel importante en las nuevas ciudades inteligentes, siendo una empresa de Big Data en busca de transformar el mundo de la logística urbana.

1.2. Productos

Foxtrot Systems mantiene sus contratos por usuario de sus servicios por ruta, sin diferenciar el valor por cada característica, este valor final termina siendo por camión. A continuación de describen los productos principales, sin embargos estos no se ofrecen por separado sino como un todo, además que son complementarios:

1.2.1. Foxtrot SDK (software development kit)

Se trata de una serie de métodos, herramientas y construcción de llamados a los servidores Foxtrot para la integración de todos los algoritmos de aprendizaje y optimización a través de una aplicación móvil existente o la propia Aplicación Foxtrot. Representa el producto principal de la empresa, ya que a través de él, se realizan las sugerencias de secuencias de visita en tiempo real para:

- Roterización en Tiempo Real, ayudando a los conductores a evitar el tráfico, maximizar el rendimiento y mantenerse dentro de los horarios de atención establecidos para cada cliente.
- Previsión de Tránsito y tiempo estimado de llegada actualizado.

1.2.2. App Mobile Android y iOS

Básicamente es la aplicación en el teléfono celular, disponible para Android y iOS, la cual permite realizar las entregas con estatus final, interactuar con usuarios web desde el centro de distribución y soportar todas las integraciones e información generada y consumida a través del SDK. Cabe destacar que está directamente integrada con la aplicación Waze para la navegación entre visitas.

Figura 1.1 Aplicación Mobile Foxtrot. Fuente: Foxtrot Systems 2018.

1.2.3. Paneles de Monitoreo

Foxtrot ofrece una serie de productos web ideal para la supervisión y seguimiento de la operación, siendo algunos para monitoreo de distribución y otros para seguimiento de indicadores.

1.2.3.1. Real Time Dashboard:

Es un resumen de una visualización del progreso de la operación logística en tiempo real.

Es ideal para supervisión en tiempo real desde el Centro de Distribución y / o Central de Ruteo. La idea es presentar la información suficiente para caracterizar el progreso de una ruta durante el día.

Partes del Real Time Dashboard

- **Indicadores Generales Diarios:** son indicadores generales que representan el progreso de todo el grupo de conductores seleccionados.

- **Detalles por ruta:** representa los detalles de cada ruta en tiempo real, contiene el ID de la ruta y el nombre del conductor, así como otros indicadores interesantes sobre el progreso de la ruta.

- **Mapa en Tiempo Real**

Es un mapa que muestra la ubicación del conductor y los clientes que debe visitar para esa ruta, todos diferenciados por colores por ruta.

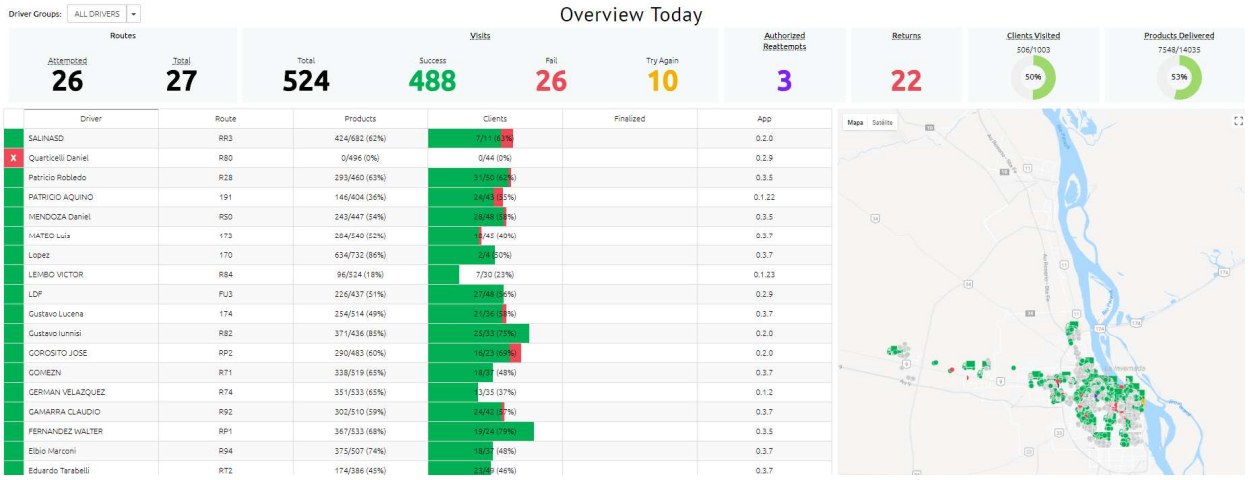

Figura 1.2: Dashboard en tiempo real. Fuente: Foxtrot Systems 2018.

1.2.3.2. Inspector de Rutas

Es una herramienta que permite visualizar todos los eventos que ocurrieron durante la línea de tiempo de la ruta. En un mapa dinámico relacionado con una ruta específica, es posible reconocer en tiempo real la duración y la ubicación de los siguientes tipos de eventos:

- Puntos de GPS: Ruta planificada, camino conducido y paradas (autorizadas y no autorizadas)
- Estatus de Entrega
- Secuencia Real sugerida vs Realizada por el conductor

Figura 1.3 Ejemplo Inspector de Rutas. Fuente: Foxtrot Systems 2018.

Esto da una visión detallada de cada ruta a través de un mapa demostrando todos los eventos y comparando la adherencia del conductor a la secuencia Foxtrot sugerida.

1.2.4. Algoritmos de Aprendizaje

Foxtrot cuenta con algoritmos de aprendizaje de Localizaciones y Horarios de Atención los cuales se desconoce su funcionamiento actualmente, sin embargo, consiguen inferir los horarios de atención y la latitud y longitud de cada cliente en ruta.

El presente proyecto será parte del diseño de un nuevo algoritmo para aprendizaje de tiempos de servicio en cada cliente.

1.3. Estructura Organizativa

La estructura de Foxtrot se define principalmente por la división de tres equipos principales: Ingeniería, Operaciones y Ventas, todos bajo objetivos comunes interconectados. Al ser una startup de apenas 25 funcionarios la estructura no tiene tanta rigidez y la jerarquía no se ve tan marcada. En ventas y operaciones los roles son mixtos, y por lo general muchas tareas de ventas son asignadas a operaciones. Foxtrot cuenta con un CEO y un CTO los dos hermanos fundadores de la empresa y un Director de Operaciones Global como principales líderes. En cuanto a las bases alrededor del mundo, todo el personal de ingeniería de desarrollo se encuentra en San Francisco, en México y Brasil en su mayoría solo hay personal de ventas y operaciones.

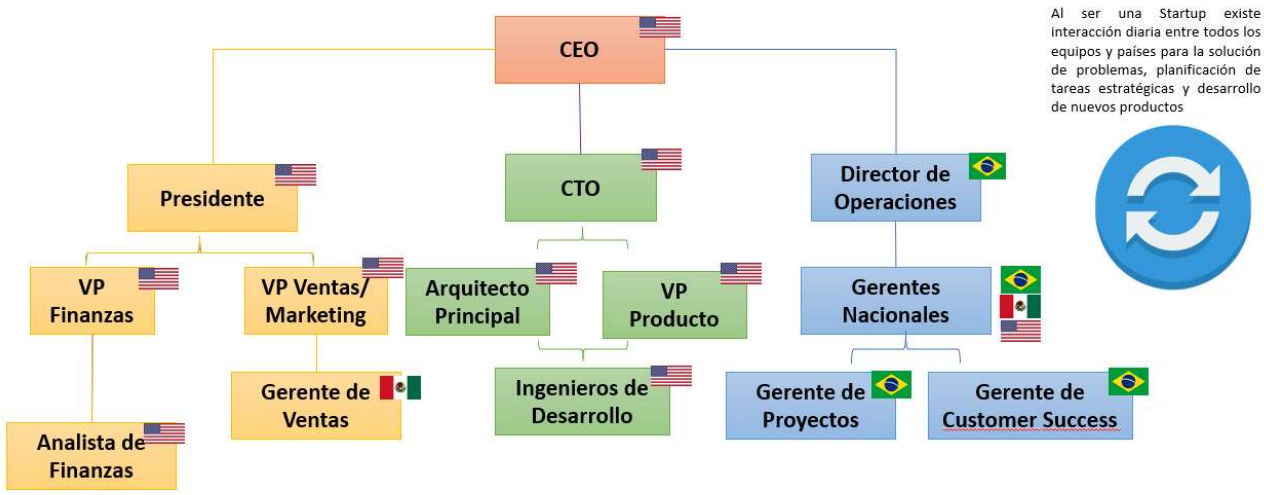

Figura 1.4 Extructura organizativa. Fuente: Foxtrot Systems 2018.

Cabe destacar que este proyecto será realizado desde el área de operaciones (Brasil), sin embargo, al ser un desarrollo de producto tecnológico, las tareas con el área de ingeniería serán de suma importancia para la investigación, pruebas e implementación de esta nueva característica.

CAPÍTULO II

MARCO TEÓRICO

2.1. Tiempo de Servicio (al cliente)

En Logística Urbana de distribución se refiere al tiempo de parada en cada cliente para terminar el servicio, sea para solo entregar productos o para hacer alguna tarea [3].

2.2. Logística Urbana

La Distribución Urbana de Mercancías (DUM), o logística de la última milla es el último eslabón de servicio en la cadena de abastecimiento. El concepto de la DUM incluye todos los movimientos relacionados con actividades comerciales y de suministro y distribución de productos para la industria, su comercialización y el consumo de bienes en las ciudades. La DUM tienen un papel clave en el desarrollo económico de las ciudades, representando una parte fundamental de la actividad comercial y de servicios, pero a su vez, constituye uno de los principales generadores de congestionamiento del tránsito, pudiendo también generar hasta un 25% de las emisiones de gases de efecto invernadero en las áreas metropolitanas, e interfiriendo con el resto del transporte urbano con respecto al uso del espacio público [3].

Desde el punto de vista del sector privado, la DUM representa una gran parte de los costos de las empresas, pudiendo la logística de última milla representar hasta el 28% de los costos logísticos totales de las empresas. La eficiencia en la distribución urbana, con entregas más frecuentes y tiempos más reducidos, es un factor clave para la competitividad local de las ciudades.

2.3. Modelo Probabilístico

Un modelo es estocástico cuando al menos una variable del mismo es tomada como un dato al azar y las relaciones entre variables se toman por medio de funciones probabilísticas. Sirven por lo general para realizar grandes series de muestreos, quitan mucho tiempo en el computador son muy utilizados en investigaciones científicas. Para lograr modelar correctamente un proceso estocástico es necesario comprender numerosos conceptos de probabilidad y estadística [4].

Dentro del conjunto de procesos estocásticos se encuentran, por ejemplo, el tiempo de funcionamiento de una máquina entre avería y avería, su tiempo de reparación y el tiempo que necesita un operador humano para realizar una determinada operación.
Modelo probabilístico es la forma que pueden tomar un conjunto de datos obtenidos de muestreos de datos con comportamiento que se supone aleatorio.

Es un tipo de modelo matemático que usa la probabilidad, y que incluye un conjunto de asunciones sobre la generación de algunos datos muéstrales, de tal manera que asemejen a los datos de una población mayor. Las asunciones o hipótesis de un modelo estadístico describen un conjunto de distribuciones de probabilidad, que son capaces de aproximar de manera adecuada un conjunto de datos. Las distribuciones de probabilidad inherentes de los modelos estadísticos son lo que distinguen a los modelos de otros modelos matemáticos deterministas.

2.4. Modelo Determinístico

Un modelo determinista es un modelo matemático donde las mismas entradas o condiciones iniciales producirán invariablemente las mismas salidas o resultados, no contemplándose la existencia de azar, o incertidumbre en el proceso modelada mediante dicho modelo.

Está estrechamente relacionado con la creación de entornos simulados a través de simuladores para el estudio de situaciones hipotéticas, o para crear sistemas de gestión que permitan disminuir la propagación de errores. Los modelos deterministas sólo pueden ser adecuados para sistemas deterministas no caóticos, para sistemas azarosos (no-determinista) y caóticos (determinista impredecible a largo plazo) los modelos deterministas no pueden predecir adecuadamente la mayor parte de sus características.

La inclusión de mayor complejidad en las relaciones con una cantidad mayor de variables y elementos ajenos al modelo determinista hará posible que éste se aproxime a un modelo probabilístico o de enfoque estocástico. Por ejemplo, la planificación de una línea de producción, en cualquier proceso industrial, es posible realizarla con la implementación de un sistema de gestión de procesos que incluya un modelo determinista en el cual estén cuantificadas las materias primas, la mano de obra, los tiempos de producción y los productos finales asociados a cada proceso [4].

2.5.Distribución de Weibull

En teoría de la probabilidad y estadística, la distribución de Weibull es una distribución de probabilidad continua [5]. Recibe su nombre de Waloddi Weibull, que la describió detalladamente en 1951, aunque fue descubierta inicialmente por Fréchet (1927) y aplicada por primera vez por Rosin y Rammler (1933) para describir la distribución de los tamaños de determinadas partículas. La función de densidad de una variable aleatoria con la distribución de Weibull x es:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} & ; \ x \geq 0 \\ 0 & ; \ x < 0 \end{cases}, \tag{2.1}$$

donde λ>0 es el parámetro de forma y κ>0 es el parámetro de escala de la distribución.

La distribución modela la distribución de fallos (en sistemas) cuando la tasa de fallos es proporcional a una potencia del tiempo:

- Un valor k<1 indica que la tasa de fallos decrece con el tiempo.

- Cuando k=1, la tasa de fallos es constante en el tiempo.
- Un valor k>1 indica que la tasa de fallos crece con el tiempo.

Su función de distribución de probabilidad es:

$$F(x;\lambda,k) = 1 - e^{-(\frac{x}{\lambda})^k}\ , \tag{2.2}$$

2.6. Distribución Exponencial

En estadística la distribución exponencial es una distribución de probabilidad continua con un parámetro λ cuya función de densidad [6] es:

$$f(x) = P(x) = \begin{cases} \lambda e^{-\lambda x} & ;\ x \geq 0 \\ 0 & ;\ x < 0 \end{cases}, \tag{2.3}$$

y su función de distribución acumulada es:

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & ;\ x \geq 0 \\ 0 & ;\ x < 0 \end{cases}, \tag{2.4}$$

El valor esperado y la varianza de una variable aleatoria X con distribución exponencial son:

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2} \tag{2.5}$$

2.7.Distribución Logarítmica Normal

En probabilidades y estadísticas, la distribución normal logarítmica es una distribución de probabilidad de una variable aleatoria cuyo logaritmo está normalmente distribuido. Es decir, si X es una variable aleatoria con una distribución normal, entonces exp(X) tiene una distribución log-normal .

La base de una función logarítmica no es importante, ya que $\log_a X$ está distribuida normalmente si y solo si $\log_b X$ está distribuida normalmente, solo se diferencian en un factor constante. Log-normal también se escribe log normal o lognormal.

Una variable puede ser modelada como log-normal si puede ser considerada como un producto multiplicativo de muchos pequeños factores independientes. Un ejemplo típico es un retorno a largo plazo de una inversión: puede considerarse como un producto de muchos retornos diarios.

La distribución log-normal tiende a la función densidad de probabilidad:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(ln(x)-\mu)^2/2\sigma^2} \tag{2.6}$$

Donde la media y varianza son:

$$E(X) = e^{\mu+\sigma^2/2}\ , \tag{2.7}$$

$$V(X) = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}\ , \tag{2.8}$$

2.8. Ajuste a distribución de probabilidad

El ajuste de distribución de probabilidad o simplemente el ajuste de distribución es el ajuste de una distribución de probabilidad a una serie de datos relativos a la medición repetida de un fenómeno variable.

El objetivo del ajuste de distribución es predecir la probabilidad o pronosticar la frecuencia de ocurrencia de la magnitud del fenómeno en un cierto intervalo.

Hay muchas distribuciones de probabilidad (ver lista de distribuciones de probabilidad) de las cuales algunas pueden ajustarse más estrechamente a la frecuencia observada de los datos que otras, dependiendo de las características del fenómeno y de la distribución. Se supone que la distribución que da un ajuste perfecto conduce a buenas predicciones.

En la adaptación de la distribución, por lo tanto, es necesario seleccionar una distribución que se adapte bien a los datos.

2.9.Método de Máxima Verosimilitud (MLE)

Es un método para estimar los parámetros de un modelo estadístico, dadas las observaciones. MLE intenta encontrar los valores de los parámetros que maximizan la función de probabilidad, dadas las observaciones. La estimación resultante se denomina estimación de máxima verosimilitud, que también se abrevia como MLE.

El método de máxima verosimilitud se utiliza con una amplia gama de análisis estadísticos. Como ejemplo, supongamos que estamos interesados en las alturas de los pingüinos hembras adultas, pero no podemos medir la altura de cada pingüino en una población (debido a limitaciones de costo o tiempo). Suponiendo que las alturas se distribuyen normalmente con alguna media y varianza desconocidas, la media y la varianza se pueden estimar con MLE mientras que solo se conocen las alturas de alguna muestra de la población general. MLE lograría eso tomando la media y la varianza como parámetros y encontrando valores paramétricos particulares que hacen que los resultados observados sean los más probables dado el modelo normal [7].

Supóngase que se tiene una muestra x1, x2, …, xn de n observaciones independientes e idénticamente distribuidas extraídas de una función de distribución desconocida con función de densidad (o función de probabilidad) f_0. Se sabe, sin embargo, que f_0 pertenece a una familia de distribuciones { f(·|θ), θ ∈ Θ }, llamada modelo paramétrico, de manera que f_0 corresponde a θ = θ_0, que es el verdadero valor del parámetro. Se desea encontrar el valor (o estimador) que esté lo más próximo posible al verdadero valor θ_0.

Tanto xi como θ pueden ser vectores. La idea de este método es la de encontrar primero la función de densidad conjunta de todas las observaciones, qué bajo condiciones de independencia, es

$$f\left(x_1, x_2, \dots, x_n \mid \theta\right) = f\left(x_1 \mid \theta\right).f\left(x_2 \mid \theta\right) \dots f\left(x_n \mid \theta\right), \tag{2.9}$$

Observando esta función desde otro punto de vista, se puede suponer que los valores observados x1, x2, …, xn son fijos mientras que θ puede variar libremente. Esta es la función de verosimilitud:

$$L\left(\theta \left| x_1, \dots, x_n \right.\right) = \prod_{i=1}^{n} f\left(x_i \left| \theta \right.\right), \tag{2.10}$$

En la práctica, se suele utilizar el logaritmo de esta función:

$$\hat{l}\left(\theta \left| x_1, \dots, x_n \right.\right) = ln\, L = \sum_{i=1}^{n} f\left(x_i \left| \theta \right.\right), \tag{2.11}$$

El método de la máxima verosimilitud estima θ_0 buscando el valor de θ que maximiza. Este es el llamado estimador de máxima verosimilitud (MLE) de θ_0:

$$\widehat{\theta_{mle}} = argmax\, \hat{l}\left(\theta \left| x_1, \dots, x_n \right.\right). \tag{2.12}$$

2.10. Bondad de ajuste:

La bondad de ajuste de un modelo estadístico describe lo bien que se ajusta un conjunto de observaciones. Las medidas de bondad en general resumen la discrepancia entre los valores observados y los que valores esperados en el modelo de estudio. Tales medidas se pueden emplear en el contraste de hipótesis, ej. el test de normalidad de los residuos, comprobar si dos muestras se obtienen a partir de dos distribuciones idénticas (por ejemplo, test de Kolmogorov-Smirnov), o si las frecuencias siguen una distribución específica (por ejemplo test chi cuadrado) [8].

2.11. Tipos de pruebas utilizados en el presente informe:

2.11.1. Kolmogorov-Smirnov:

La determinación de la Kolmogorov-Smirnov (K-S) se basa en la asignación de funciones de distribución empíricas (ECDF). Dados los puntos Y1, Y2, Y3, ..., YN, la función ECDF está definida como:

$$E_n = n(\text{i}) / \text{N}\,, \tag{2.13}$$

donde n (i) es el número de puntos menos que Yi, y Yi se ordena del menor valor al máximo. Esta es una función escalonada que aumenta por 1 / N en el valor de cada punto ordenado de datos.

El gráfico siguiente es el trazado de la función de distribución empírica con respecto a la Distribución normal acumulativa de valores para 100 números de referencia normal. La prueba K-S se basa en la distancia mínima entre estas dos curvas [9].

La prueba de Kolmogorov-Smirnov se define por las siguientes hipótesis nula y alternativa:

H0: Los datos siguen una distribución específica.
Ha: Los datos no siguen la distribución especificada.

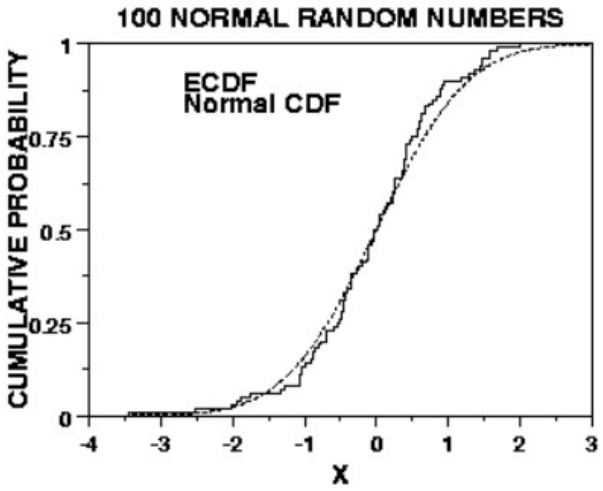

Figura 1.4 Ejemplo de distribución normal empírica Kolmogorov-Smirnov. Fuente: National Institute of Standars an Technology US.

Donde el estadístico de prueba se define como:

$$D = Max_{1<i<N}(F(Yi) - \frac{i-1}{N}, \frac{i}{N} - F(Yi)) \tag{2.14}$$

donde F es la distribución acumulativa teórica de la distribución que se está probando, que debe ser una distribución continua (es decir, no distribuciones discretas como Binomial o Poisson), y debe estar completamente especificada, es decir, los parámetros de ubicación, escala y forma no pueden estimarse a partir de los datos.

2.11.2. KSL TEST (Kolmogorov-Smirnov-Lilliefors):

Es similar con el test Kolmogorov-Smirnov Simple, la única modificación ocurre en la estimación de los parámetros de la función acumulativa teórica a ser evaluada.

Supongamos que queremos evaluar un ajuste de distribución normal, siguiendo la prueba KSL utilizaremos los parámetros μ y σ muestrales en vez de teóricos, partiendo de la media y la desviación estándar de la muestra para construir la función teórica a ser evaluada.

Las hipótesis y estimador no cambian.

2.11.3. Krammer Von Mises test

Este método se ocupa del modelado de una distribución de probabilidad, busca verificar la compatibilidad entre una muestra de datos {x1, …, xn} y una distribución de probabilidad candidata elegida previamente. La prueba de bondad de ajuste de Cramer-von-Mises permite funciona en el caso unidimensional y con una distribución continua [10].

Esta prueba de bondad de ajuste se basa en la distancia entre la función de distribución acumulativa de la muestra {x1, …, xn} y el de la distribución candidata, denotado F. Esta distancia ya no es la desviación máxima como en la prueba de Kolmogorov-Smirnov, ahora es la distancia al cuadrado y Integrado en todo el dominio de variación de la distribución:

$$D = \int_{-\infty}^{\infty} [F(x) - \widehat{F_N}(x)]^2 \, dF \,, \tag{2.15}$$

Siendo F(x) la función candidata y F_N la función distribución de la muestra.

Considerando una muestra {x1, …, xn} la distancia es estimada por:

$$\widehat{D_N} = \frac{1}{12N} + \sum_{i=1}^{N} \left[\frac{2i-1}{2N} - F(x_i)\right]^2 \,, \tag{2.16}$$

La distribución de probabilidad de la distancia es asintóticamente conocida (es decir, ya que el tamaño de la muestra tiende a infinito). Si N es lo suficientemente grande, significa que para una probabilidad alfa y un tipo de distribución candidato, se puede calcular el valor crítico / umbral d_α de manera que podemos comparar:

$$\begin{aligned} \widehat{D_N} &> d_\alpha \,, Rechaza\ el\ ajuste \\ \widehat{D_N} &\leq d_\alpha \,, Acepta\ el\ ajuste \,, \end{aligned} \tag{2.17}$$

Para rechazar o aceptar la distribución candidata con un riesgo de error α.

2.12. Árbol de Decisiones

Un árbol de decisión es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

2.13. Partición Recursiva

Los modelos de partición recursiva (RP) son un método flexible para especificar la distribución condicional de una variable dado un vector de valores predictores X. Tales modelos utilizan una estructura de árbol para recursivamente divida el espacio del predictor en subconjuntos donde la distribución de Y es sucesivamente más homogéneo. Los nodos terminales del árbol corresponden a las distintas regiones de la partición, y la partición se determina dividiendo las reglas asociadas con cada uno de los nodos internos. Por moviéndose desde el nodo raíz hasta el nodo terminal del árbol, cada observación es luego asignado a un nodo terminal único donde se determina la distribución condicional de Y [11].

Los dos tipos de respuesta más comunes son continuos y categóricos, con tareas correspondientes a menudo conocidas como regresión y clasificación.

Dado un conjunto de datos, una estrategia común para encontrar un buen árbol es usar un algoritmo para hacer crecer el árbol y luego “podarlo” de nuevo para evitar el sobreajuste. Esto genera una secuencia de árboles, cada uno de los cuales es una extensión del árbol anterior, luego se selecciona un solo árbol cortando el árbol más grande de acuerdo con un criterio de selección de modelo tal como pruebas de poda, validación cruzada o hipótesis de complejidad de costos de si dos nodos contiguos deben ser colapsado en un solo nodo.

2.13.1. Aprendizaje del Modelo de Partición Recursiva

Para aprender o estimar un modelo de RP, se asume una muestra de entrenamiento que consiste en coordenadas, es decir entradas y respuestas (xi, yi), i =1,. . . , n deben estar disponibles. Tanto el árbol "T" como los parámetros de cada nodo terminal (Candidatos) deben ser estimados utilizando datos de entrenamiento.

Para una T fija, un supuesto común es que los valores de respuesta son Independientes e idénticamente distribuidos. dentro de cada terminal nodo. Los datos en cada nodo terminal se pueden considerar como una muestra separada y convencional. las técnicas de estimación (por ejemplo, máxima verosimilitud) producen estimaciones de parámetros de nodos familiares tales como la media de la muestra para una respuesta normal continua y proporciones de la muestra para una categoría respuesta multinomial.

Para considerar la estimación de T, primero debe especificarse una función objetivo, que proporciona un mecanismo para evaluar la calidad de un árbol T. La probabilidad de registro de los datos de entrenamiento es uno de esos criterios. Para una respuesta normal, en este modelo, el criterio correspondiente sería la minimización de una suma residual de cuadrados.

Con una función objetivo que cuantifica la calidad de un árbol, el problema de estimación se convierte en una búsqueda sobre todos los árboles posibles para optimizar el objetivo. La búsqueda sobre el conjunto de árboles es, por lo tanto, una búsqueda combinatoria sobre un espacio discreto finito pero muy grande.

El algoritmo de búsqueda más común consiste en que todas las observaciones de entrenamiento inicialmente se agrupan en un solo nodo. El algoritmo considera dividir en dos nodos hijos, examinando todas las divisiones posibles en todas las variables posibles.

La regla de división es aquella que da el mejor valor de la función objetivo (por ejemplo, la suma residual de cuadrados más pequeña cuando se suma sobre la que seleccionan dos nodos hijos). El procedimiento se repite en cada nodo hijo recursivamente hasta que un gran árbol se construye.

Se pueden emplear varias estrategias para decidir qué tan grande crecerá un árbol. En el algoritmo CHAID de Kass (1980) [12], se utilizaron pruebas de hipótesis para decidir cuándo dejar de subdividir, dando como resultado un árbol.

CHAID también se puede extender para que se aplique al caso en el que tenemos una variable de respuesta continua, por ejemplo, ventas registradas. Sin embargo, en este caso se utilizan pruebas de Fisher en lugar de pruebas de Chi-cuadrado. Las variables predictoras continuas también pueden incorporarse mediante la determinación de valores de corte para crear grupos ordinales de variables, basadas, por ejemplo, en determinados percentiles de la variable.

2.14. Prueba Chi Cuadrado

La prueba de Chi cuadrado de Pearson es una prueba estadística que se aplica a conjuntos de datos categóricos para evaluar la probabilidad de que cualquier diferencia observada entre los conjuntos surgiera por casualidad [13].

2.14.1. Test para independencia estadística:

En este caso, una "observación" consiste en los valores de dos resultados y la hipótesis nula es que la aparición de estos resultados es estadísticamente independiente. Cada observación se asigna a una celda de una matriz bidimensional de celdas (llamada tabla de contingencia) de acuerdo con los valores de los dos resultados [14]. Si hay r filas y c columnas en la tabla, la "frecuencia teórica" para una celda, dada la hipótesis de independencia, es:

$$E_{i,j} = N_{pi.p.j} \ , \tag{2.18}$$

donde N es el tamaño total de la muestra (la suma de todas las celdas en la tabla), y

$$p_{i.} = \frac{O_{i.}}{N} = \sum_{j=1}^{c} \frac{O_{i,j}}{N} \ , \tag{2.19}$$

es la fracción de observaciones de tipo i que ignora el atributo de columna (fracción de totales de fila), y

$$p_{.j} = \frac{O_{.j}}{N} = \sum_{i=1}^{r} \frac{O_{i,j}}{N} \ , \tag{2.20}$$

es la fracción de observaciones de tipo j que ignora el atributo de fila (fracción de totales de columna). El término "frecuencias" se refiere a números absolutos en lugar de a valores ya normalizados.

El valor de la estadística de prueba es:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = N \sum_{i,j} p_{i.} p_{.j} \left(\frac{\frac{O_{i,j}}{N} - p_{i.} p_{.j}}{p_{i.} p_{.j}} \right)^2 , \tag{2.21}$$

Tenga en cuenta que cX^2 es 0 si y solo si

$$O_{i,j} = E_{i,j} \ \forall \ i,j \quad , \tag{2.22}$$

es decir, solo si el número esperado y verdadero de observaciones es igual en todas las celdas.

Ajustar el modelo de "independencia" reduce el número de grados de libertad en p = r + c - 1. El número de grados de libertad es igual al número de celdas rc, menos la reducción en los grados de libertad, p, que reduce a (r - 1) (c - 1).

Para la prueba de independencia, también conocida como la prueba de homogeneidad, una probabilidad de xi cuadrado de menor o igual a 0.05 (o la estadística de ji cuadrado en o mayor que el punto crítico de 0.05) es comúnmente interpretada por los trabajadores aplicados como justificación para rechazar la hipótesis nula de que la variable de fila es independiente de la variable de columna. La hipótesis alternativa corresponde a las variables que tienen una asociación o relación donde no se especifica la estructura de esta relación.

2.15. JMP (SAAS)

JMP (pronunciado jump) es una herramienta potente e interactiva de visualización de datos y análisis estadísticos. Con JMP puede aprender más acerca de sus datos realizando análisis e interactuando con los datos mediante tablas de datos, gráficos, diagramas e informes.

JMP permite a los investigadores crear una amplia variedad de análisis estadísticos y modelizaciones. También resulta útil para los analistas de negocio que deseen descubrir rápidamente tendencias y patrones presentes en datos. Con JMP no es necesario ser un experto en estadística para obtener información a partir de datos [15].

Por ejemplo, JMP se puede utilizar para:

- Crear diagramas y gráficos interactivos para explorar datos y descubrir relaciones.
- Descubrir patrones de variación con múltiples variables a la vez.
- Explorar y resumir grandes cantidades de datos.
- Desarrollar modelos estadísticos para predicción de variables.

2.15.1. Modelo de Partición para crear árboles de decisión en JMP

La plataforma JMP de partición recursiva particiona los datos de acuerdo con una relación entre los predictores y los valores de respuesta, creando un árbol de decisión. El algoritmo de partición busca todas las divisiones posibles de predictores para predecir mejor la respuesta. Estas divisiones (o particiones) de los datos se hacen recursivamente para formar un árbol de reglas de decisión. Las divisiones continúan hasta que se alcanza el ajuste deseado. El algoritmo de partición elige divisiones óptimas de un gran número de divisiones posibles [16].

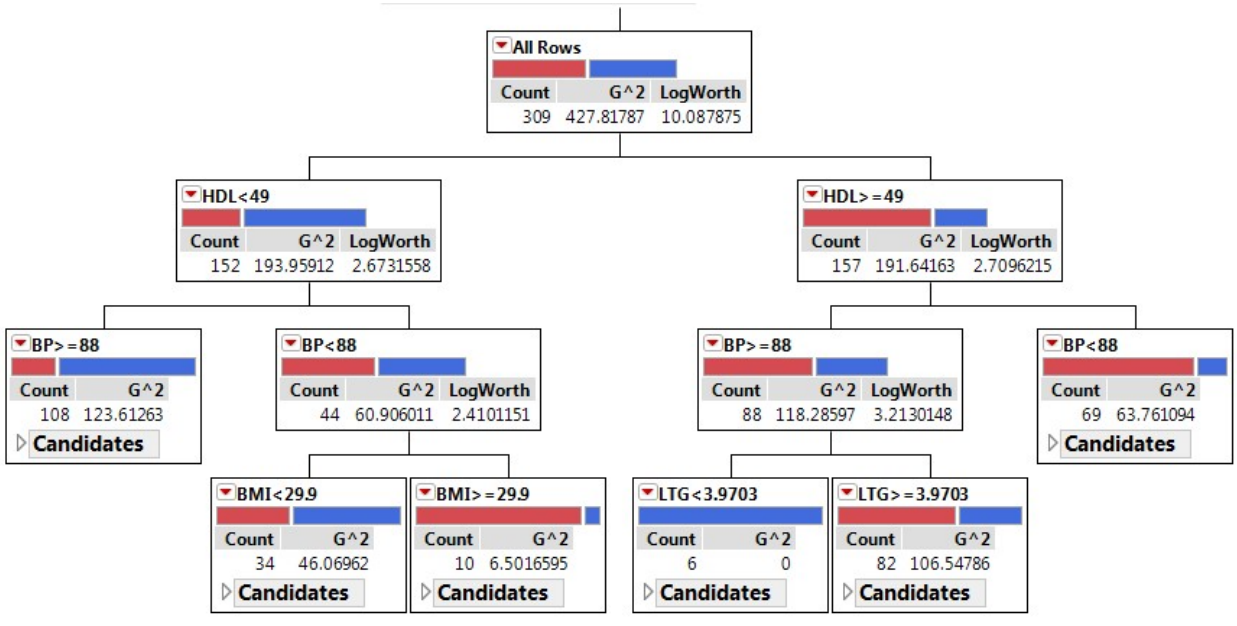

Figura 2.2 Ejemplo de un árbol de decisión en JMP. Fuente: JMP 2018.

2.15.2. Respuestas y Factores

La respuesta (candidatos) puede ser continua o categórica (nominal u ordinal):

- Si la respuesta es categórica, entonces está ajustando las probabilidades estimadas para los niveles de respuesta, minimizando la probabilidad de registro residual chi-cuadrado [2 * entropía].
- Si la respuesta es continua, entonces la plataforma se ajusta a los medios, minimizando la suma de los errores cuadrados.

Los factores (variables independientes) pueden ser continuos o categóricos (nominales u ordinales):

- Si el factor es continuo, entonces la partición se realiza de acuerdo con un valor de "corte" de división para el factor.
- Si el factor es categórico, divide las categorías X en dos grupos de niveles y considera todas las agrupaciones posibles en dos niveles.

2.15.3. Criterio de Partición JMP

La división de nodos se basa en la estadística LogWorth, que se informa en los informes del candidato para los nodos. LogWorth se calcula de la siguiente manera:

$$LogWorth_i = log_{10}(Pvalor_i) \quad , \tag{2.23}$$

Donde el P-Valor ajustado se calcula de una manera compleja que toma en cuenta el número de formas diferentes en que pueden ocurrir las divisiones. Para respuestas continuas, la suma de cuadrados (SS) se informa en los informes de nodos. Este es el cambio en la suma de cuadrados de error debido a una determinada partición.

El SS para un candidato que ha sido elegido es:

$$SS_{test} = SS_{parent} - \left(SS_{right} + SS_{left}\right) \quad , \tag{24}$$

También se informa para las respuestas continuas la estadística de la diferencia. Esta es la diferencia entre los valores predichos para los dos nodos secundarios de un nodo principal.

2.16. Algoritmo

Un algoritmo es una secuencia de pasos lógicos necesarios para llevar a cabo una tarea específica, como la solución de un problema. Los algoritmos son independientes tanto del lenguaje de programación en que se expresan como de la computadora que los ejecuta. En cada problema el algoritmo se puede expresar en un lenguaje diferente de programación y ejecutarse en una computadora distinta; sin embargo, el algoritmo será siempre el mismo.

2.17. Computación en la Nube

Se conoce como Computación en la Nube o Cloud Computing, como el acceso y almacenamiento de los datos en aplicaciones, servidores y plataformas en la nube. Es decir, a través de internet en lugar de hacerlo en el disco duro de cada ordenador.

Cuando hablamos de nube, se hace mención a granjas de servidores físicos de alta potencia y perpetua disponibilidad, con diferentes fines [17].

2.18. Amazon Web Services (AWS)

AWS por sus siglas en inglés, es una filial de amazon.com, que vende herramientas y servicios de cómputo en la nube, ofreciendo una muy amplia gama de opciones especializadas según las necesidades de los usuarios [18].

El mayor diferencial de AWS son servicios que están preparados tanto para autónomos, como pequeñas y medianas empresas o grandes corporaciones, ya que existen posibilidades para escalar las instancias o el almacenamiento según la empresa vaya creciendo. Las categorías en las que Amazon Web Services ofrece herramientas son:

- **Cloud computing:** ofrece todo lo necesario para la creación de instancias y el mantenimiento o el escalado de las mismas. Amazon EC2 está catalogado como el rey indiscutible dentro de los servicios de computación en la nube de Amazon.
- **Bases de datos:** puedes escoger distintos tipos de bases de datos y pueden permanecer en la nube mediante el servicio Amazon RDS, que ofrece diferentes tipos de sistemas gestores de bases de datos (SGBD) a elegir como MySQL, PostgreSQL, Oracle, SQL Server y Amazon Aurora, o Amazon DynamoDB para NoSQL.
- **Creación de redes virtuales:** permite la creación de redes privadas virtuales (VPN) a través de la nube, gracias principalmente al servicio Amazon VPC.
- **Aplicaciones empresariales:** Amazon WorkMail es un servicio de correo empresarial, al que pueden unirse otros servicios como Amazon WorkDocs y Amazon WorkSpaces, que facilitan los procesos de oficina de cualquier empresa.
- **Almacenamiento y gestores de contenido:** ofrece diferentes tipos de almacenamiento, tanto para archivos con acceso regular, poco frecuente o incluso como archivo. Amazon S3 es el servicio principal, aunque complementan la oferta otros como Amazon Glacier o Amazon EBS.
- **Inteligencia de negocios:** ofrece sistemas para análisis de datos empresariales a gran escala y otros servicios para la gestión de flujos de datos, que transforman datos en información valiosa para las empresas.
- **Gestión de aplicaciones móviles:** las herramientas como Amazon Mobile Hub permiten la gestión, desarrollo, pruebas y mantenimiento de aplicaciones móviles a través de la nube.
- **Internet de las cosas (IoT)**: ofrece servicios para establecer conexiones y análisis de todos los dispositivos conectados a internet y los datos recogidos por los mismos, para optimizar recursos mediante machine learning.
- **Herramientas para desarrolladores:** ofrece opciones para almacenar código, implementarlo automáticamente o incluso publicar software mediante un sistema de entrega continua.
- **Seguridad y control de acceso:** permite establecer niveles de autenticaciones en varios pasos para poder proteger el acceso a sus sistemas internos, ya están en la nube o instalados de forma local en sus instalaciones.

2.19. SQL (Structured Query Language)

Es un tipo de lenguaje vinculado con la gestión de bases de datos de carácter relacional que permite la especificación de distintas clases de operaciones entre éstas. Gracias a la utilización del álgebra y de cálculos relacionales, el SQL brinda la posibilidad de realizar consultas con el objetivo de recuperar información de las bases de datos de manera sencilla.

2.20. Global Positioning System (GPS)

Sistema de Posicionamiento Global o GPS es un sistema creado por el Departamento de Defensa de los Estados Unidos y permite, a través de una red de 24 satélites, indicar la posición de un cuerpo en la superficie terrestre con gran precisión. El GPS recurre al método matemático conocido como trilateración para trabajar con la información que aportan los satélites y así determinar la ubicación del objeto, Para conocer una posición, el equipo receptor (conocido también como GPS) localiza al menos tres satélites de la red, recibiendo señales de ellos que señalan la identificación y el horario. Al calcular el tiempo que demoran las señales en llegar desde los satélites hasta el equipo, se mide la distancia existente entre los artefactos. Luego, con estas distancias ya establecidas, es posible determinar la posición relativa del objeto (es decir, sus coordenadas) [19].

2.21. Apuntamiento de Entrega en el Aplicativo

Dentro de Foxtrot existen 2 estatus finales y uno transitorio para seleccionar al momento de visitar un cliente (Figura 2.3).

Figura 2.3 Ejemplo de visualización de opciones de entrega en el aplicativo. Fuente: Foxtrot Systems 2018.

Successful (confirmar entrega, botón verde): Este es un estatus final para un cliente, indicando que los productos fueron entregados correctamente.

Visit Later (visitar más tarde, botón amarillo): Este es un estatus transitorio, donde es obligatorio escoger un motivo indicando por qué se debe pasar más tarde y no ahora (ej: cliente sin dinero), esto colocará en la secuencia nuevamente para ser visitado más tarde.

Failed (entrega fallida, botón rojo): Este es un estatus de entrega final para un cliente, donde es obligatorio escoger un motivo de rechazo, a diferencia de "intentar más tarde", este estatus no deja el cliente como pendiente.

2.22. Tiempo de Parada Autorizada

Durante la ruta Foxtrot es capaz de determinar el tiempo de parada a partir de los cambios de dirección, velocidad y movimiento detectado por el GPS del celular, y exactamente este tiempo de parada es el que define el tiempo de servicio, siempre y cuando este cerca de un cliente.

Cuando la parada se realiza a una distancia de máxima de 150 metros del cliente, se considera como una parada autorizada.

CAPÍTULO III

MARCO METODOLÓGICO

La metodología descrita tiene como objetivo definir: una idealización para el proceso de parada y atendimiento de un cliente, a partir de este definir variables que influyen y están disponibles como entradas dentro de la base de datos. Una vez entendiendo esto, se procederá a elegir un estimador para el tratamiento de datos históricos. Se definirán distintos escenarios donde se tomarán en cuenta nuevos parámetros y variables que logren sofisticar el modelo. Finalmente, la evaluación de cada modelo será realizada a través de una simulación del tiempo estimado según las condiciones de ese momento, para compararlo con el tiempo real que tomo cada evento de parada (según GPS del celular) asociado con cada cliente. La viabilidad de cada modelo será caracterizada por su desviación estándar, el error medio y la cantidad de datos necesaria para el aprendizaje.

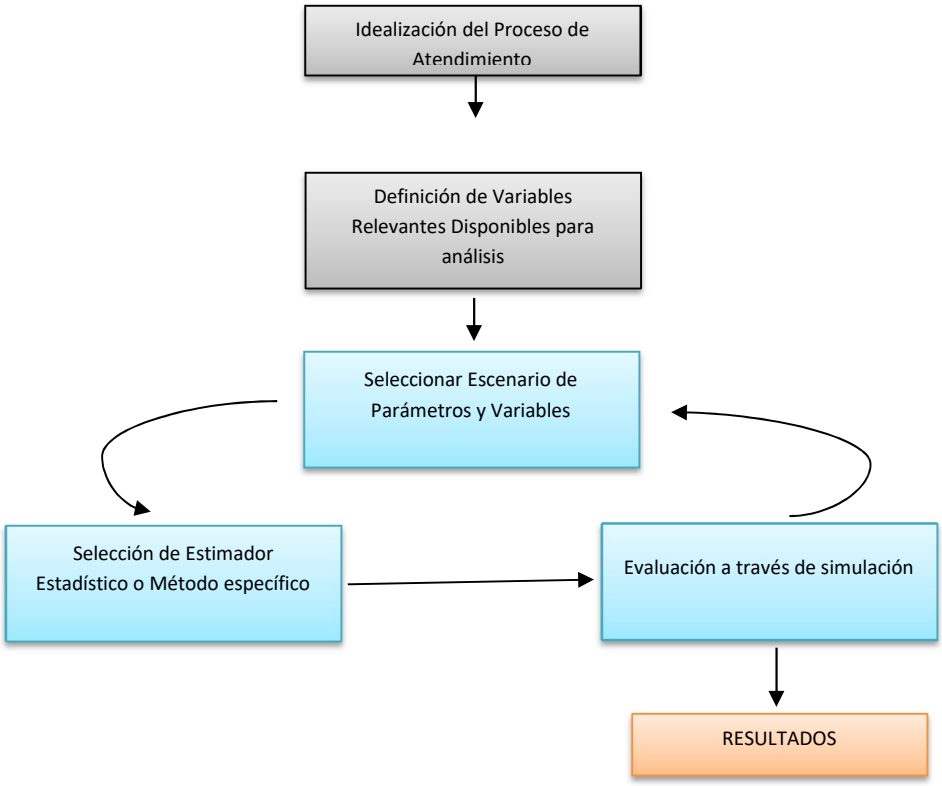

Figura 3.1: Descripción general de la metodología. Fuente: Elaboración propia.

3.1. Definición de Proceso de Entrega Ideal

Foxtrot Systems opera con clientes que realizan entrega de todo tipo de productos, los cuales varían, pero por lo general se reducen a "Cajas/Paquetes" con contenido, pueden ser bebidas, pan, productos de farmacia, etc. Además, con los supervisores de las distintas operaciones de entrega de bienes y servicio que usan actualmente la inteligencia Foxtrot, se realizaron reuniones presenciales y vía telefónica para discutir los detalles del proceso.

Finalmente fue posible encontrar puntos comunes entre las operaciones y definir un proceso genérico que contemple las tareas realizadas durante el tiempo de servicio (Tiempo de para en el cliente). La operación logística que determinará los tiempos de servicio para cada cliente en un caso ideal se puede describir como:

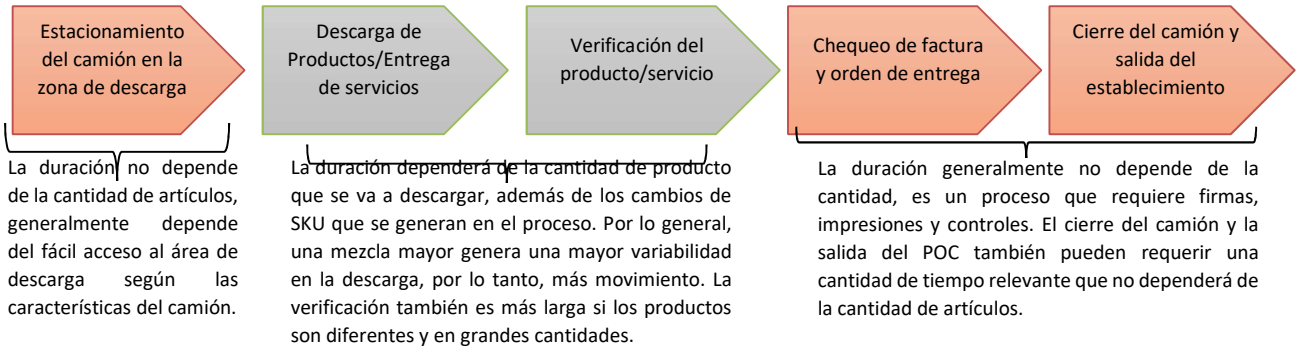

Figura 3.2 Proceso genérico de servicio al cliente. Fuente: Elaboración propia.

Tomando en cuenta el proceso anterior, es posible definir el tiempo de servicio de la siguiente forma:

$$Tiempo\ de\ Servicio = Te + Td + Tv + Tc + Ts, \tag{3.1}$$

siendo Te el tiempo de desacelerar el vehículo y estacionar en la zona de descarga, Td como el tiempo de descarga de productos / entrega de servicios, Tv como tiempo de verificación, Tc siendo el tiempo de chequeo y facturación y Ts representando la ultima parte de cierre del vehículo y salida del local. Con esta ecuación (26) definiremos las variables que afectan a cada uno de los tiempos a estimar.

3.2. Variables que afectan, limitaciones y complejidad

Para entender las variables que podrían surgir, se realizó una lluvia de ideas con diferentes expertos en el proceso de operación por cada cliente, entre los participantes externos se conversó con:

- Supervisor de Distribución de una empresa de entrega de alimentos en México y Brasil.
- Supervisor de Ruta de una empresa de reparto de bebidas en Argentina.
- Gerente de Logística de una empresa de distribución de E-Comerce en Brasil.
- Choferes de una empresa de distribución de E-Comerce en Brasil.
- Choferes de una empresa de distribución de bebidas en Brasil, Colombia y Argentina.

Después de realizar estas entrevistas vía videoconferencia y algunas presenciales, procedimos a visitar la operación y a acompañar el proceso de descarga en vivo.

Figura 3.3: Visita a operación de descarga en Argentina. Fuente: Elaboración propia.

Durante la visita a la operación no se realizó muestreo de tiempos de descarga, debido a que el diseño del modelo debe ser genérico y cada operación, cliente, producto, conductor, tiene tiempo estándar distinto y son situaciones completamente diferentes según lo observado, sin embargo, como mencionamos antes, todas encajan perfectamente en el modelo genérico. Realizar análisis de movimientos se hace inviable para la cantidad de conductores y clientes que maneja Foxtrot, además Foxtrot Systems busca la menor interferencia posible en el trabajo de los conductores y la operación en general buscando un modelo “Plug and Play”, donde todo el cálculo no proviene de parámetros si no de inteligencia generada por datos automáticamente.

Al realizar las consultas externas se compartió un archivo internamente para ser discutido por todos los integrantes de la empresa, donde se debatieron las ideas sobre las variables relevantes en cada proceso con otros integrantes del equipo en varios países, según la experiencia observada en las operaciones y situaciones vividas anteriormente, por ejemplo en los Estados Unidos el tiempo de estacionar no es una variable muy significativa por el difícil acceso al lugar de descarga, si no por las filas que se podrías generar al tener otros proveedores descargando en el mismo local, caso contrario en Colombia donde existen barrios con difícil acceso donde el camión debe maniobrar para situarse cerca de la zona de descarga.

Finalmente, después de esta discusión interna, se llegó a los siguientes factores que afectan cada parte de la Ecuación (3.26) definida para tiempo de servicio.

Tiempo de desacelerar el vehículo y estacionar en la Zona de Descarga
Dificultad de acceso, conductor, tipo de Camión, fila (tiempo de espera), disponibilidad de lugar para estacionar.

Tiempo de Descarga de Productos / Entrega de Servicios
Cantidad de productos, tipo de producto, variedad de producto, agilidad de los tripulantes para descarga, cantidad de personas descargando, distancia del lugar de estacionamiento al lugar de entrega, layout del camión.

Tiempo de Verificación
Variedad de productos**,** tipo de empaquetado del producto**,** confianza del cliente.

Tiempo de Chequeo y Facturación
Tipo de pago, proceso de facturación de la empresa.

Cierre del Vehículo y salida del local
Dificultad de salida, conductor, tipo de camión.

3.2.1. Aleatoriedad de variables

Algunas de las variables antes mencionadas podrían ser aleatorias, sin embargo, muchas de ellas son conocidas previamente, como el conductor, equipo de reparto, tipo de camión, productos a entregar, horario, infraestructura de salida y entrada del cliente, tipo de pago a ser realizado, etc.

La aleatoriedad irá representada en el tiempo de parada con el comportamiento del conductor, estado de ánimo, cansancio, disposición de recepción del conductor a determinada hora, etc. Las filas son particulares de los lugares mayoristas o grandes Centros de Distribución, donde no representan ni el 10% del volumen total de las entregas realizadas con el Sistema Foxtrot, por lo tanto, el estudio del comportamiento de estas filas es poco significativo para objetivos de este trabajo.

Es necesario entender que existen en este estudio variables que son aleatorias y variables que son predeterminadas, por lo tanto, los modelos propuestos deben contemplar ambos tipos de variables.

Al ser de naturaleza estocástica, la evaluación de la desviación estándar para cada estimación de tiempo de parada es de suma importancia.

3.3. Base de datos actual y variables disponibles para análisis:

Foxtrot tiene la capacidad de captar la duración de distintos eventos en ruta a través de señal GPS, pero solo es almacenado para fácil análisis en la base de datos la duración en milisegundos de un evento siempre y cuando este asociado a un evento de visita o un “POC Status” en la APP del celular del conductor, en la figura XXX, los distintos eventos y estatus que se tienen disponibles en Foxtrot.

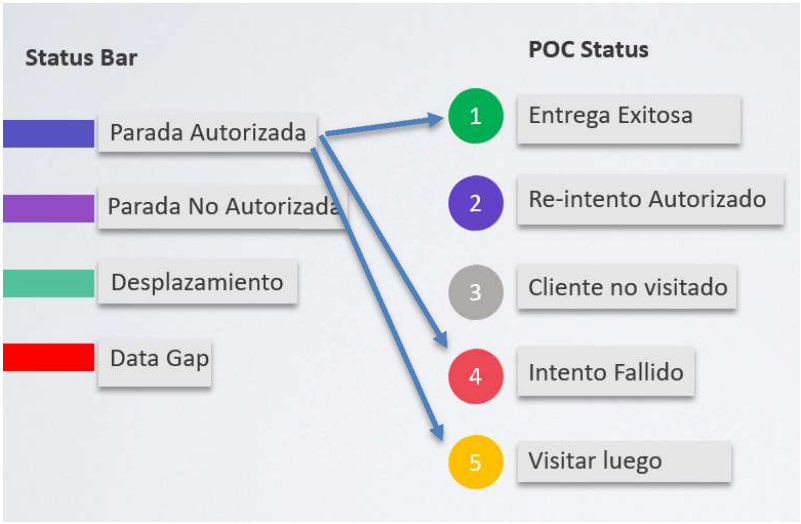

Figura 3.4 Estatus de entrega y evento relacionado. Fuente: Elaboración propia.

Lo que diferencia una parada autorizada de una no autorizada es que la autorizada está a menos de un valor límite de distancia de un cliente, lo cual la hace una parada autorizada. Las paradas no autorizadas no están cerca de un cliente por lo general, son paradas para almorzar o abastecer gasolina. Para efectos de análisis de tiempo de servicio nos enfocaremos en las Paradas Autorizadas para cada cliente, específicamente cuando fue una entrega "Exitosa" o un "Visitar Luego". Se pudo descubrir que el estatus "Visitar Luego" es utilizado en algunos casos donde se entrega el producto y se pasa más tarde solo para cobrar y cambiar es estatus a entrega "Exitosa" por lo que ambos eventos de parada sumados constituyen el Tiempo de Servicio real en el cliente.

Para cada uno de estos eventos el SDK de Foxtrot es capaz de recibir algunas variables de la lista de variables generadas con el proceso de entrega para relacionarlas con el evento de parada:

- Identificación del conductor/tripulantes
- Identificación del cliente
- Productos para entregar en cada cliente
- Tipo de productos
- Hora Exacta de Entrega

La disponibilidad de estas entradas depende de cada cliente y la información que envía al momento de generar una ruta en el ambiente Foxtrot. Para el Centro de Distribución ubicado en Pittsburgh no estará disponible información sobre los productos.

El diseño de los modelos propuestos utilizará estas variables como determinísticas.

3.3.1. Limitaciones encontradas

Para diseñar este tipo de modelo y agregar variables lo primero que se necesita es contar con tiempo de parada histórico almacenado en la base de datos del sistema, el cual tiene algunas limitaciones

que implicarían cambios en la forma en cómo se almacenan los datos, para finalmente definir el "Tiempo de Servicio Realizado" de un cliente (tiempo de parada).

- **En las bases de datos el tiempo se almacena solo por cada vez que el cliente apunta como visita en un cliente:** Ejemplo: el repartidor llega a un cliente, estaciona, descarga la mercancía y después mueve el camión a otra zona del mismo cliente para recibir el dinero, al momento de recibir el dinero, apunta la visita en la aplicación del teléfono celular. En la base de datos solo quedará el tiempo de la segunda parada, que no representa todo el tiempo de servicio. Se hace necesario almacenar todos los tiempos de parada asociados a un cliente dentro de una visita y no solo la parada que va junto con el accionar del botón de entrega del aplicativo.

- **Múltiples entregas en la misma parada:** en algunos casos, sobre todo en zonas con alto tránsito, como el centro de ciudades o centros comerciales, ocurre que en un mismo evento de parada los conductores hacen entrega en distintos clientes, haciendo difícil la división del tiempo de servicio para cada uno. Posible solución: hacer una ponderación con la cantidad de producto a ser entregada en cada cliente para definir en cual cliente tardó más que en otro.

- **Comportamiento del conductor:** Para asociar una visita con una parada, es necesario que el conductor opere bajo [illegible]cir, si el conductor apunta la visita en el aplicativo una vez en movimiento saliendo del cliente y no al estar detenido en el cliente no será posible asociar una parada a una visita.

- **Data Gap**: básicamente se trata de teléfonos/zonas con mala señal de GPS que no permiten captar parte de la ruta, entre ellas algunas paradas incomple[illegible]an perjudicar el cálculo.

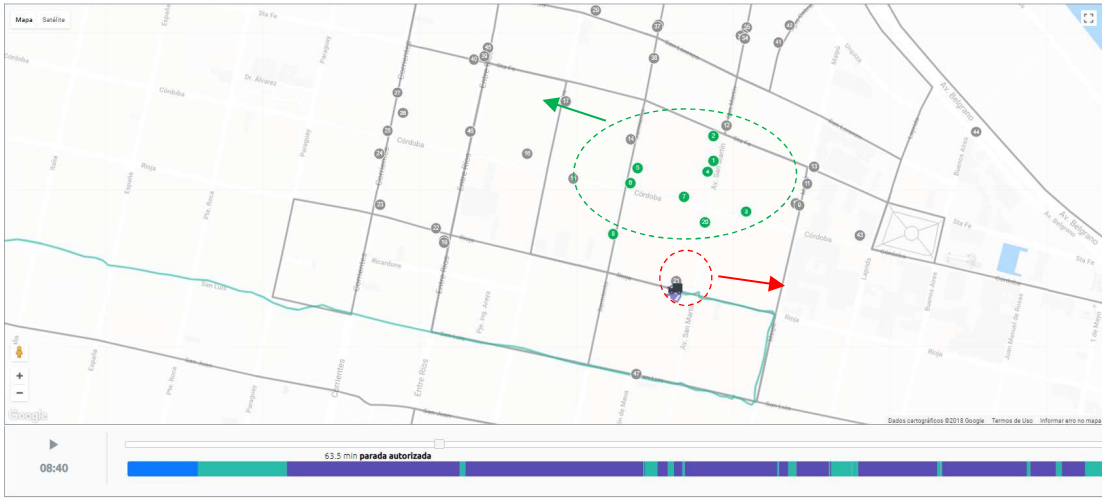

Figura 3.5 Ejemplo de múltiples entregas en la misma parada a través del Route Inspector. Fuente: Foxtrot Systems 2018.

Figura 3.6 Lugar de parada para múltiples entregas en la ciudad de Rosario visto desde Google Street View. Fuente: Foxtrot Systems 2018.

3.4. Obtención y limpieza de datos (filtros para eliminar valores atípicos)

Para la obtención de Datos se escribió una consulta en idioma SQL (Strucutred Query Lenguage) para extraer los datos de la base de datos la cual extrae la información bajo un formato con las siguientes columnas:

Fecha, Nombre del Conductor, Nombre de la Ruta, Tiempo de Servicio Enviado por nuestros Clientes, Duración del Evento de Parada, ID del cliente Asociado a la parada, Nombre del Cliente Asociado a la Parada, Tipo de Evento, Horario de Inicio de parada, Intento de Entrega, Estatus de Visita, Numero de Productos a entregar, Tipo de Producto a Entregar.

La construcción de esta consulta nunca se había realizado antes, lo cual fue un avance interesante en la visualización de los datos.

Una vez descargado el set de datos se deben eliminar los valores atípicos y ayudar a solventar las limitaciones mencionadas anteriormente, para esto se realizarán las siguientes operaciones con la base de datos cruda:

Definición de tiempo de parada (para considerar entregas de múltiplos en la misma parada)

Cuando varias entregas de un conductor tienen el mismo inicio de evento de parada en ruta, se trata claramente de un caso de múltiple entrega por parada, por lo tanto, es posible considerar el tiempo de parada para cada cliente de dos formas distintas:

Promedio normal:
Para calcular un promedio normal se sigue la siguiente lógica con los datos:

If

Concat (Duración del Evento de Parada1, Nombre del Conductor1) =
Concat (Duración del Evento de Parada2, Nombre del Conductor2)

Then

Tiempo de servicio = Duración del Evento de Parada / Count
(matching_route_event_start_time & Driver_ID)

Else

Tiempo de servicio = Duración del Evento de Parada

Resumiendo, la lógica en la siguiente ecuación:

$$Tservicio = \frac{(Tparada\ total)}{\#\ Clientes\ en\ Parada}, \tag{3.2}$$

dando así el mismo valor de duración de parada a cada cliente atendido.

Promedio Ponderado: existe un tratamiento más sofisticado de este tipo de casos, se trata de ponderar la duración que pasa en cada tiempo dependiendo de la cantidad de producto que está entregando en cada cliente de la parada, de esta forma el tiempo por cliente se vería definido de la siguiente forma:

$$Tparadacliente1 = \frac{qt1 * (Tptot)}{qttotal}, \tag{3.3}$$

donde Qt1 es la cantidad de producto entregada para el cliente 1, Tptot representa el tiempo de parada total (puede ser para entregar en varios clientes en la misma parada), y Qttotal es la cantidad de producto total entregado durante la parada, o sea la cantidad de productos a ser entregada para todos los clientes que fueron atendidos durante esa parada

La utilización de promedio ponderado dependerá de la disponibilidad de la variable "cantidad de productos" por parte del cliente, más adelante se podrá observar que no todas las operaciones analizadas en la presente investigación cuentan con ese dato.

3.4.1. Tratamiento de valores atípicos de duración

Para tratar la sensibilidad a los valores atípicos, simplemente se realizarán los cálculos para un subconjunto de los datos, la idea es evitar valores atípicos de datos, como paradas de 8 horas por ejemplo.

Un método para hacer esto es rechazar valores por encima del percentil 99 del Centro de Distribución. Por ejemplo: suponiendo que un Centro de Distribución tiene una distribución de tiempo de servicio por cliente con un percentil 99% de 90 minutos, todas las paradas con valores por encima de 90 minutos serían considerados valores atípicos, de misma forma podríamos definir la misma lógica para un tiempo mínimo de duración de parada.

De forma alternativa, es posible definir un “tiempo de servicio razonable máximo/mínimo", que entraría en la configuración de Centro de Distribución. Probablemente, cada Centro de Distribución (CD) sabría cuál es este valor para su operación. La ventaja de trabajar con percentiles como límite es no tener que agregar un elemento codificado en la configuración de CD; la desventaja es que tendrá que calcularse agregando un paso más de cálculo a la lógica del algoritmo. A fines del presente estudio nos apoyaremos del valor del percentil para definir valores mínimos y máximos, definidos a través del percentil 99 % de cada distribución de frecuencia.

La idea es que esta definición de un “valor atípico de duración” sea un concepto importante para tener en cuenta durante el monitoreo de la operación. Por ejemplo, si se observa que un conductor supera el percentil de tiempo permitido dentro de un cliente, el sistema Foxtrot podría emitir una alerta sobre el caso.

3.5. Estimadores estadísticos base - Media vs Mediana (para base time)

Está claro que es necesario utilizar alguna medida para estimar el tiempo de servicio y entender la distribución de la muestra para los clientes. Esto significará la definición del estimador para el modelo probabilístico que se desea implementar, a partir de este analizar la posibilidad de ajustar distribución. Para eso se evaluará elegir entre media y mediana.

La mediana minimiza la predicción de errores para casos típicos, es decir, la mayoría de los puntos dentro de la distribución estadística que está en la “cola derecha” (Figura 3.7).

Observando la Figura 3.7, es fácil comprender que la Media se sesgará más adelante por la cola de la distribución; esto significa que, en la mayoría de los casos, la media estará sobreestimando el tiempo de servicio, en el orden de varios minutos. Además, las medias son sensibles a los valores atípicos, y las medianas no lo son; esto hace que usar la mediana sea una elección más simple. Sin embargo, resulta que las sobreestimaciones y subestimaciones de tiempo de servicio por cliente a lo largo de una ruta se anulan entre sí al utilizar la media, dando un mejor resultado viendo la ruta como un todo.

La intuición para esto es la idea de que la media sobreestima los casos típicos de modo que cuando, inevitablemente, una entrega lleva mucho más tiempo de lo esperado, la precisión en el tiempo de servicio final por ruta es más acertada. El uso de la mediana subestima la duración total de la ruta en aproximadamente 1 o 2 minutos por cantidad de paradas en la ruta. En una ruta con ~ 30 entregas, esto significa que la ruta en promedio se estima que demora entre media hora y una hora menos de lo que realmente será. En este sentido, usar la media es una mejor opción para la planificación de rutas.

En términos de computación, la media es mucho más rápida, porque se puede usar un promedio continuo. Es decir, cuando aparece un nuevo punto de datos, el promedio se actualiza según ese nuevo punto de datos; no es necesario extraer todos los puntos de datos antiguos y volver a calcular el promedio. Para calcular la mediana, debe consultar todos los datos pasados cada vez que se introduce un nuevo punto de datos.

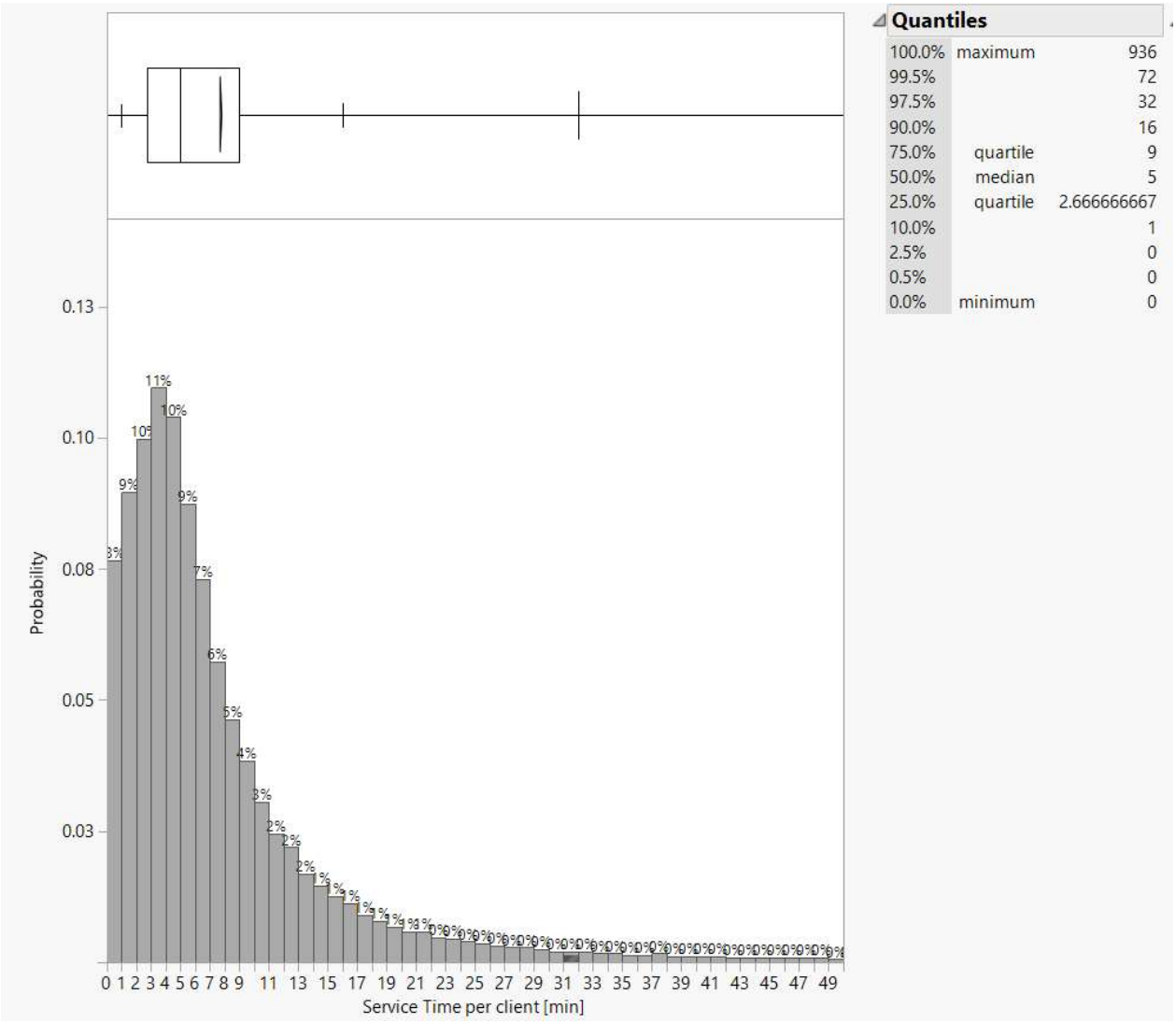

Figura 3.7 Histograma y Boxplot de Tiempos de Servicio para Rosario, Guarulhos y Tarija. Fuente: Elaboración propia.

En el presente estudio se realizarán pruebas de ajuste a cada distribución, donde la "Esperanza" será el estimador principal a evaluar.

3.6. Modelos y clientes para Analizar

Los clientes que serán analizados fueron elegidos debido a su urgencia de usar el modelo, diferencias en la ubicación geográfica (puede ser una diferencia interesante en el comportamiento de cada región) y la disponibilidad de cantidad de datos en cada uno de ellos. Se utilizarán dos meses de datos históricos para definir el aprendizaje de cada método, construyendo distribuciones y árboles de decisión a partir del registro de entregas de junio y julio de 2018. Para la evaluación, se realizará una prueba con un mes de predicción para comparar los modelos con el tiempo de parada real (agosto 2018). Dos meses serán suficientes debido a la frecuencia de visita de los clientes, la cual es máximo semanal. La muestra será diferenciada por Centro de Distribución al igual que sus resultados, los CDs seleccionados son:

- Distribuidor de bebidas en Rosario (Argentina)
- Distribuidor de Bebidas en Guarulhos (Brasil)
- Distribuidor de bebidas en Tarija (Bolivia)
- Distribuidor de Alimentos en Pittsburg (Estados Unidos)

Modelos:

1era versión: el primer modelo consiste en buscar establecer un tiempo por defecto para aquellas operaciones con baja frecuencia de entrega o con pocos datos por cliente, por lo que se intentará definir el modelo que se adapte mejor a cada centro de distribución utilizando el método de Máxima Verosimilitud, evaluando el ajuste con algunas pruebas de bondad de ajuste. Finalmente se considerará la media estimada como predicción de general de tiempo de servicio para cada cliente, los modelos serán evaluados a través de la herramienta estadística de JMP considerando los siguientes tipos de modelos: Normal, Weibull, Exponencial, Normal Doble (dos distribuciones normales), Logarítmica normal.

2da versión: Árbol de Decisiones considerando cantidad de producto. Se construirá un árbol de decisión a partir del método de particiones recursivas, considerando utilizando como variable independiente y continua, la cantidad de producto a ser entregado generando candidatos (valores de tiempo de servicio) partiendo del input de cantidad de productos a ser entregados. La información detallada sobre este método es posible consultarla en el Capítulo II.

3ra versión: Árbol de Decisiones considerando usuario (conductor/equipo de reparto) y cantidad de producto a entregar. Se construirá un árbol de decisión a partir del método de particiones recursivas, considerando utilizando dos variables independientes para definir candidatos, una es continua (cantidad de producto) y la otra nominal (nombre del repartidor), de esta forma se generarán ramas en el árbol de decisión considerando quién realizará la entrega, asi como la cantidad de producto que entregará. Se espera un resultado en el modelo como el siguiente ejemplo: "El tiempo de parada será de 6 minutos cuando el Conductor A o el Conductor B estén entregando menos de 10 paquetes, si fuera el Conductor C tardará 10 minutos".

4ta versión: Promedio por cliente, para este caso se estimará el próximo tiempo de servicio a través de la media simple de los datos válidos (limpios) por cliente.

3.7. Metodología de Evaluación y control

Para realizar la evaluación de cada modelo se utilizará una medición de error de estimación en minutos por cliente definido como:

- Distribución normalizada de errores (minutos) por cliente, siendo

$$Error = Tiempo\ de\ Parada\ Real - Tiempo\ de\ Parada\ Estimado\ , \quad (3.4)$$

Se realizarán estimaciones a partir de dos meses de datos (junio y julio) para predecir del mes siguiente y comparar con los tiempos de parada reales.

Todos los clientes en el presente estudio comparten su propio tiempo de servicio, establecido bajo input manual en su sistema de planificación de rutas, se evaluará que tan buena es su predicción con respecto a la calculada bajo cada uno de los modelos expuestos en este informe.

CAPÍTULO IV

RESULTADOS Y ANÁLISIS

En esta sección se presentarán los resultados y análisis de los modelos a ser comparados para la estimación de tiempos de servicio, presentados en el mismo orden descrito en la metodología.

4.1. Valores máximos considerados en los datos:

Para la limpieza de datos se realizó el análisis de los datos para cada Centro de Distribución separando los dos tipos de industria: Bebidas y Panes. Como fue mencionado en la metodología se halló el percentil 99 % para estos dos tipos de distribución dando como resultado:

75 minutos para la industria de bebidas.

127 minutos para la distribución de pan.

Como se observará más adelante, existe una notable diferencia entre ambos tipos de industria en cuanto tiempos de parada se refiere. Esto se debe a que el proceso en el caso de la distribución de pan conlleva una puesta a punto en las áreas de exhibición y un proceso de conteo de inventario y toma de orden de compra que convierte al repartidor en un vendedor, lo cual sale un poco de la idealización del proceso de entrega, sin embargo, se puede asumir que las variables elegidas para hacer parte del análisis siguen impactando en la operación de igual forma. Por lo tanto, los valores de los percentiles son distintos, para evitar valores atípicos se excluirán todas las muestras con valores por encima de estos percentiles para realizar la construcción de todos los modelos.

4.2. Ajuste continuo para cada tipo de modelo por Centro de Distribución.

Para evaluar cada tipo de modelo en cada centro de distribución se comenzará por la distribución normal.

Cada ajuste evaluado reflejará todos sus parámetros y el resultado de la prueba de bondad de ajuste.

Para Guarulhos el ajuste normal (Figura A.1) se obtuvo una media 9,26 minutos por cliente con una dispersión de casi el mismo valor, un resultado de la prueba KSL que rechaza la hipótesis nula con bastante diferencia. Claramente no sigue una distribución normal, pero da una idea del valor medio del tiempo de parada en Guarulhos Brasil.

Para Tarija en Bolivia (Figura A.3), se observa una media de entrega menor, lo cual es lógico debido a que el volumen de venta por cliente es más pequeño. Los resultados en cuanto a bondad de ajuste son similares a Guarulhos, la distribución normal no es adecuada.

Para Rosario según el ajuste normal (Figura A.2), con una media de 6,47 el tiempo es bastante similar a Tarija, ambas operaciones son controladas bajo directrices parecidas debido a que están bajo la misma supervisión dentro de la empresa (Latinoamérica Sur).

La media del ajuste normal en Pittsburg (Figura 4.1) demuestra que el Tiempo de Servicio por cliente es de casi media hora, afirmando lo mencionado anteriormente sobre la diferencia entre la duración de tiempo de parada entre la distribución de bebidas y la de panes. Es interesante observar que en todos los casos que se muestran hasta ahora, la desviación estándar es grande.

Bajo la prueba KSL se rechaza la hipótesis nula de normalidad al 1%. Ninguno de los Centros de Distribución analizados sigue una distribución normal.

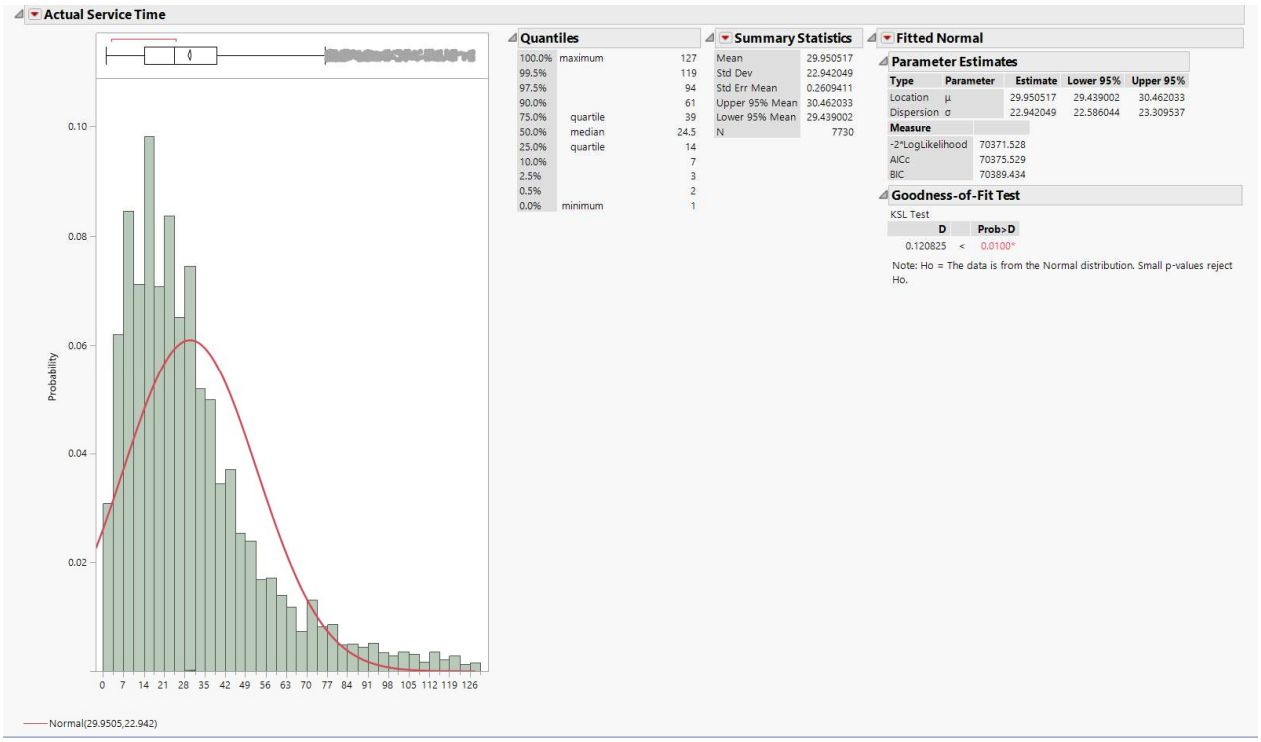

Figura 4.1 Distribución normal para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

La distribución Weibull parece tener mejores resultados de ajuste a simple vista para Guarulhos que la distribución (Figura A.4). Con un valor esperado de 9,92 minutos por cliente, ligeramente mayor que con la distribución normal. Sin embargo, la hipótesis nula se rechaza bajo el test de Cramer Von Misses con 99% de confianza.

El ajuste Exponencial es uno de los menos adecuados que se han propuesta hasta ahora ya que deja de considerar una cantidad significativa de datos fuera del ajuste. Además, se rechaza la hipótesis nula bajo el test de Kolmogorov por una gran diferencia.

Cuando se prueba el ajuste Normal Doble para todos los casos muestra que la distribución cuenta con una acumulación grande de clientes con tiempo de parada corto y luego otros que representan una minoría, con un tiempo de servicio alto, esto ayuda a entender la necesidad de un agrupamiento por tipo de cliente.

El desafío es identificar cuales clientes entran en cada clasificación. En este caso al ser dos curvas normales, no es posible evaluar con una única distribución para caracterizar el valor esperado de la curva, sin embargo, da para reconocer el porcentaje de clientes que entran en cada distribución.

En el caso de Guarulhos (Figura 4.2) se dividen los clientes en dos grupos, uno que tarda 5.72 minutos en media y otro 19.97 minutos en media, siendo 75% la probabilidad de entrar en una distribución con media de 5.72 minutos.

En el caso de Rosario (Figura A.12) se dividen los clientes en dos grupos, uno que tarda 4.27 minutos en media y otro 14.59 minutos en media, siendo 78% la probabilidad de entrar en una distribución con media de 5.72 minutos.

En el caso de Tarija (Figura A.13) se dividen los clientes en dos grupos, uno que tarda 4.05 minutos en media y otro 17.01 minutos en media, siendo 82% la probabilidad de entrar en una distribución con media de 4.05 minutos.

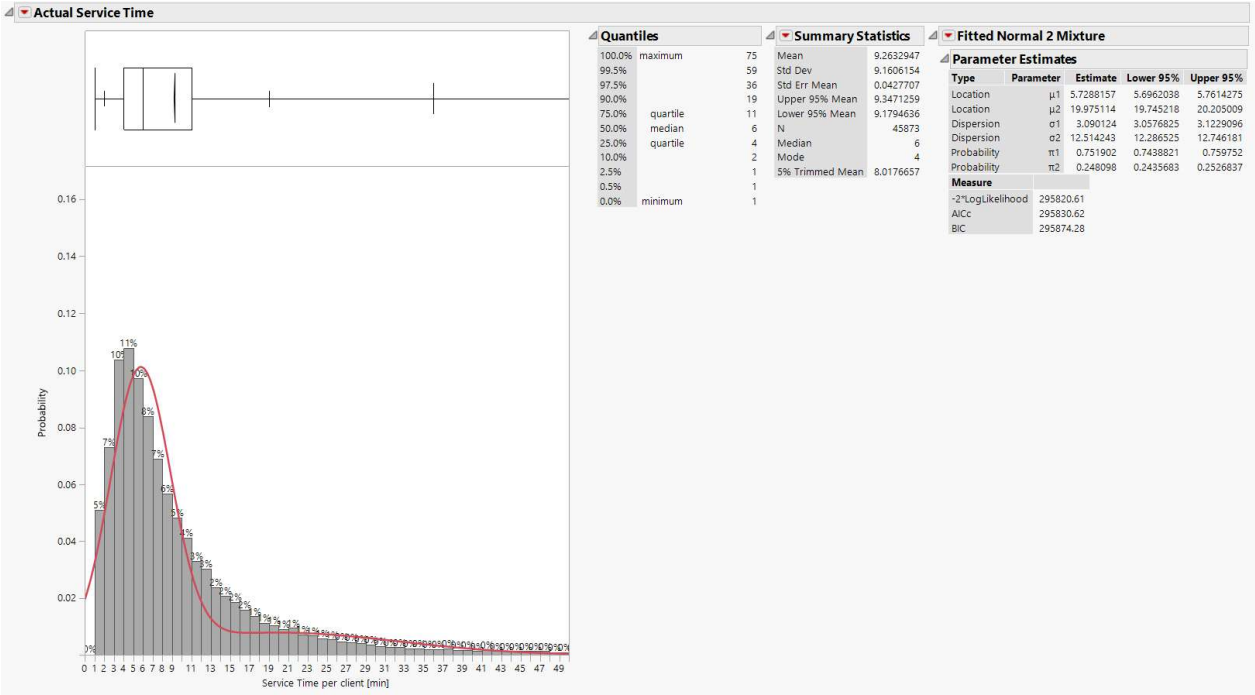

Figura 4.2 Ajuste Normal doble para el Tiempo de Servicio en Guarulhos. Fuente: Elaboración propia.

En el caso de Pittsburgh (Figura 4.3) se dividen los clientes en dos grupos, uno que tarda 20.21 minutos en media y otro 55.28 minutos en media, siendo 72% la probabilidad de entrar en una distribución con media de 20.21 minutos.

Descubrir estas divisiones, consigue dar vista general de la posibilidad de realizar un Árbol de Decisiones que pueda direccionar a una partición de la distribución dependiendo de distintas variables, generando medias diferentes en cada caso.

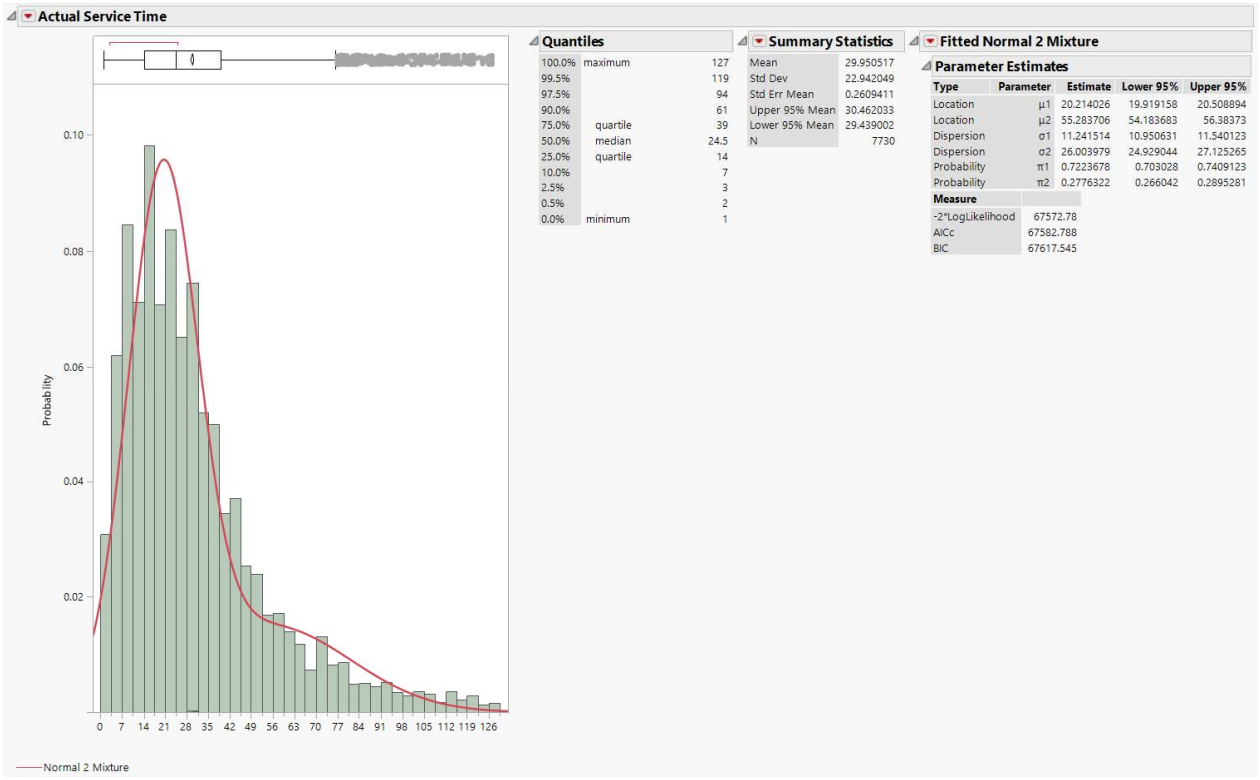

Figura 4.3 Ajuste normal doble para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

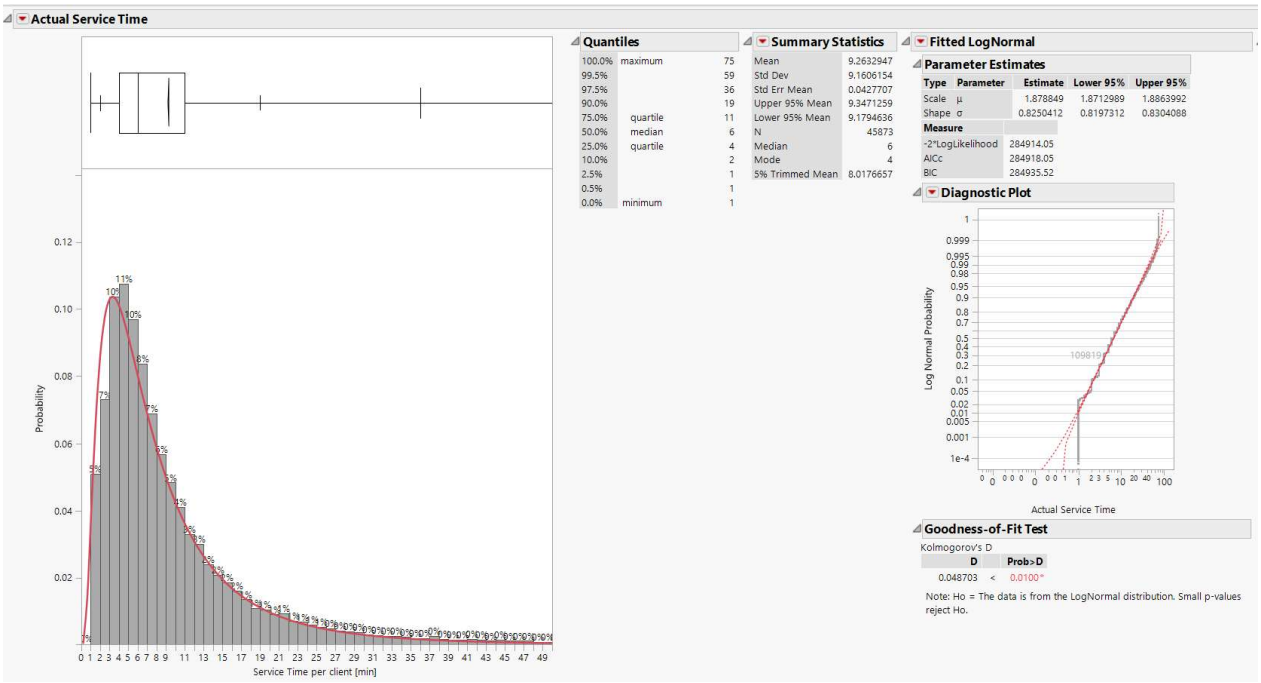

Figura 4.4 Ajuste a distribución Logarítmica Normal (Lognorm) para el Tiempo de Servicio en Guarulhos. Fuente: Elaboración propia.

Finalmente se probó con la distribución Logarítmica Normal (Lognorm), con la cual se obtuvo los resultados más adecuados en comparación a los analizados anteriormente, si utilizamos la ecuación del valor esperado Lognorm (Ecuación 2.7) a partir de esta distribución con parámetros $\mu = 1{,}87$ y $\sigma = 0{,}82$ para Guarulhos (Figura 4.4) se obtiene un tiempo medio de 9,85 minutos por cliente, con un ajuste bajo la prueba de Kolmogorov que si bien rechaza la hipótesis nula a 99% (P valor 0.048) funciona para una confianza del 95% lo cual es un resultado mucho que los anteriores.

Para el ajuste Lognorm en Rosario (Figura 4.5) con parámetros $\mu = 1{,}53$ y $\sigma = 0{,}80$ se obtuvo un tiempo medio esperado de 6,11 minutos, con una prueba de Kolmogorov que si bien rechaza la hipótesis nula a 99% (P valor = 0.036) funciona para una confianza del 95% lo cual es el mejor ajuste obtenido en comparación a las otras distribuciones.

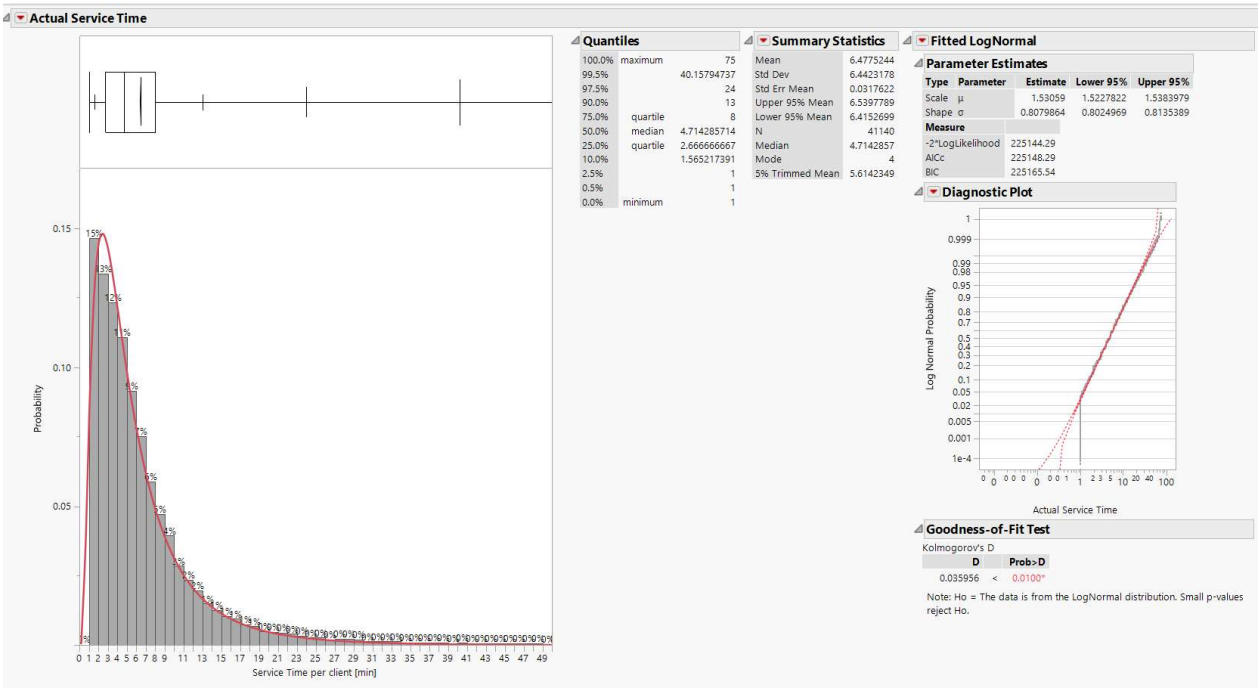

Figura 4.5 Ajuste a distribución Logarítmica Normal (Lognorm) para el Tiempo de Servicio en Rosario. Fuente: Elaboración propia.

Para el ajuste Lognorm en Tarija (Figura 4.6) con parámetros $\mu = 1{,}47$ y $\sigma = 0{,}79$ de un tiempo medio de 6,12 minutos, con un ajuste evaluado bajo la prueba de Kolmogorov que si bien rechaza la hipótesis nula a 99% (P valor 0.076) funciona para una confianza del 90%.

Para el ajuste Lognorm en Pittsburgh (Figura 4.7) con parámetros $\mu = 3{,}09$ y $\sigma = 0{,}84$ de un tiempo medio de 32,54 minutos, con un resultado de la prueba de Kolmogorov que si bien rechaza la hipótesis nula a 99% (P valor 0.036) funciona para una confianza del 95%.

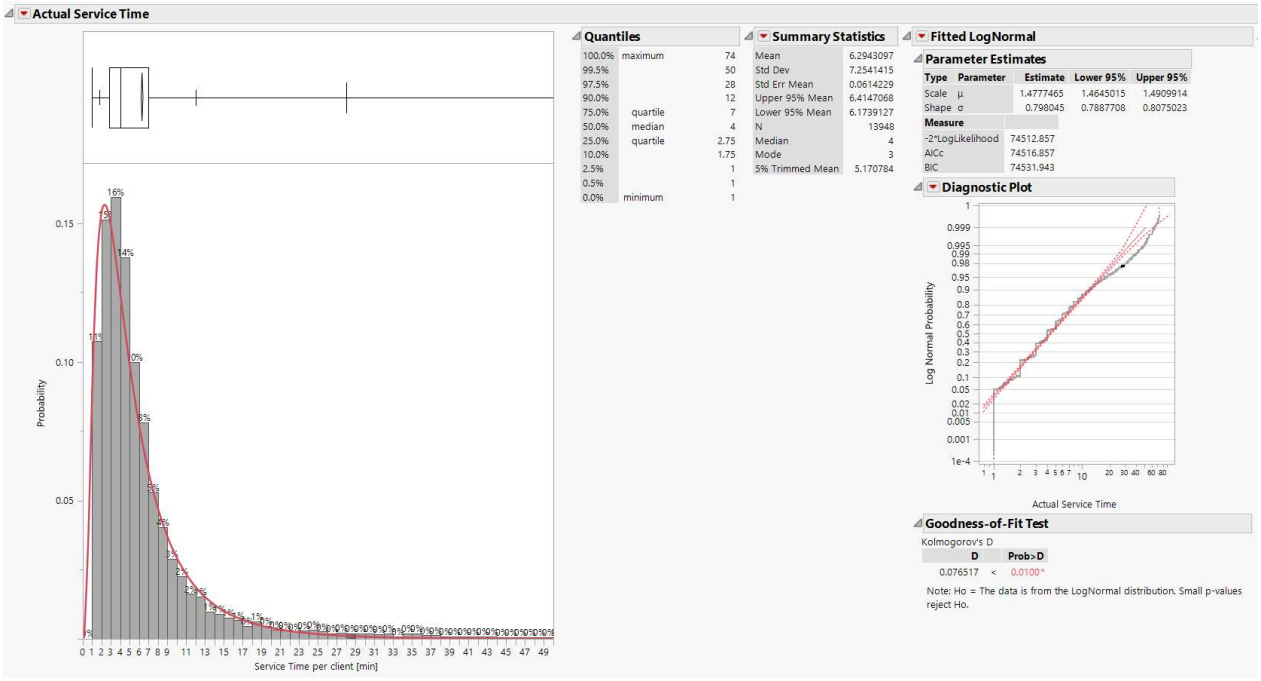

Figura 4.6 Ajuste a distribución Logarítmica Normal (Lognorm) para el Tiempo de Servicio en Tarija. Fuente: Elaboración propia.

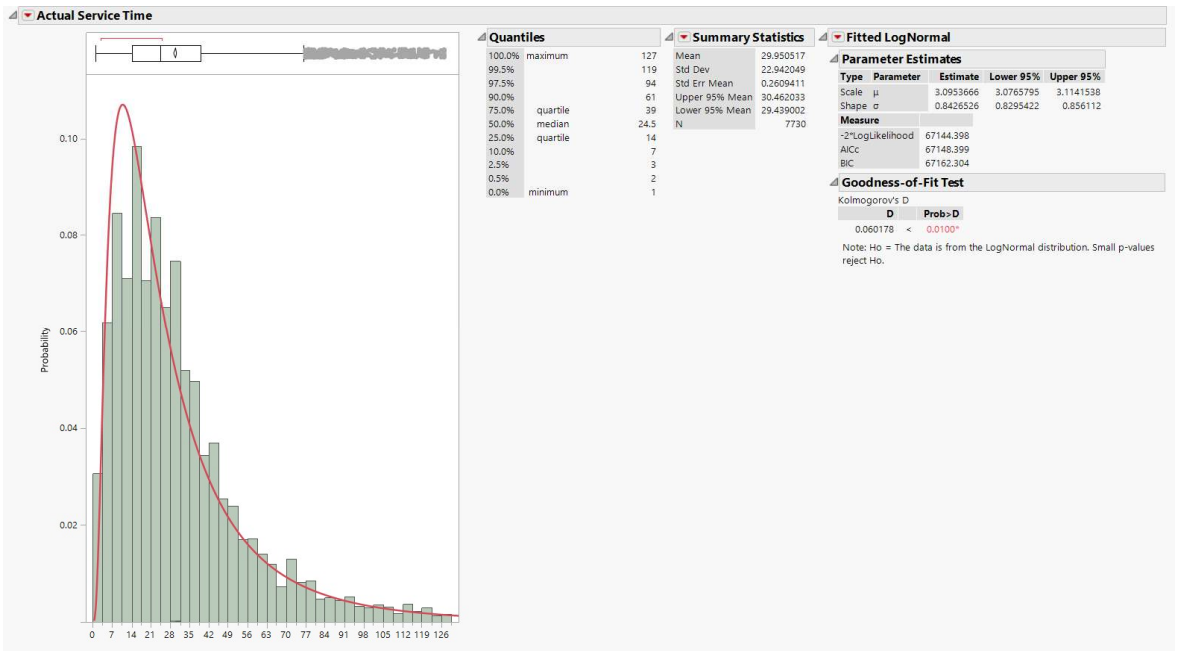

Figura 4.7 Ajuste a distribución Logarítmica Normal (Lognorm) para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

4.2.1. Distribución Lognorm como la que mejor se ajusta al proceso de distribución:

Es sumamente interesante como ambos tipos de operación de descarga siguen la misma distribución con parámetros diferentes. Finalmente la distribución Lognorm se muestra como la más precisa para definir una función de probabilidad continua para el proceso de tiempo de parada

en diferentes clientes debido a que es la función de mejor se comporta ante este tipo de curvas con alta frecuencia en valores bajos de tiempo (minutos de parada) y una cantidad de entre 20% y 15% de paradas que ocurren con un tiempo mucho mayor, como podemos ver claramente en la Figura 4.8 donde se ve las normales dobles.

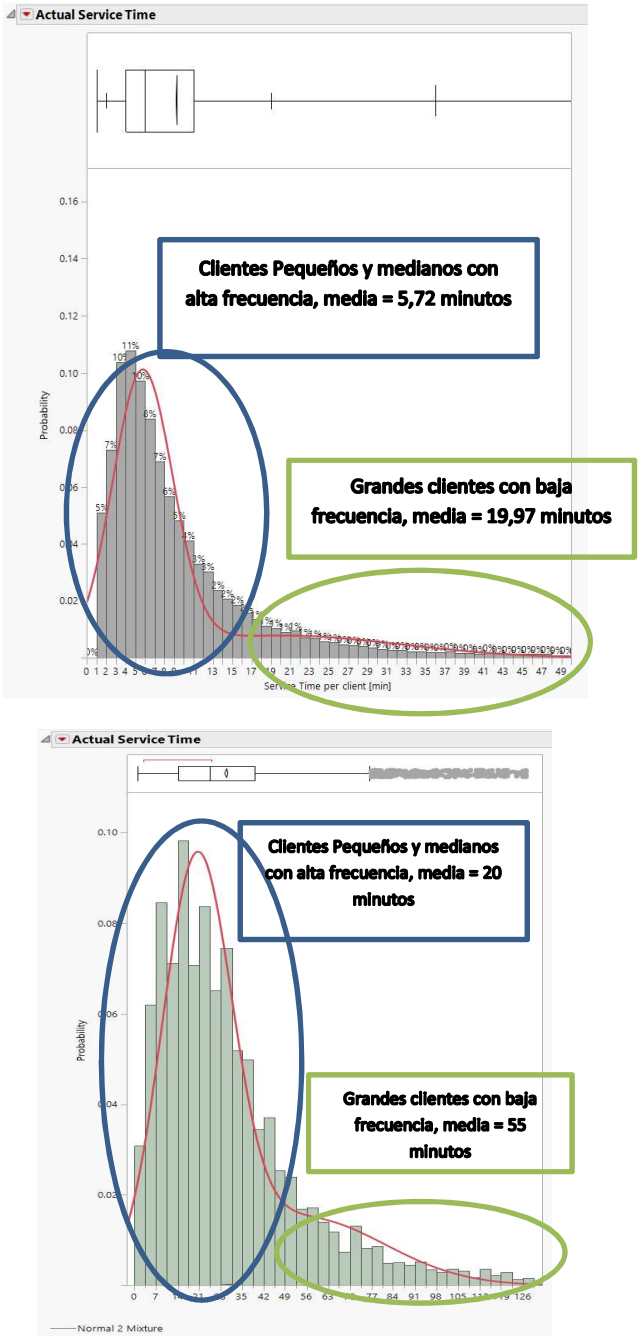

Figura 4.8 Distribución y tipos de clientes, arriba Guarulhos y abajo Pittsburgh. Fuente: Elaboración propia.

Es posible explicar este comportamiento en la distribución del tiempo de servicio si analizamos los tipos de clientes que atienden las industrias de consumo de bebidas y alimentos. Todos los centros de distribución analizados entregan en distintos tipos de clientes, desde bares, restaurantes, supermercados, minimercados y mayoristas, solo que, en la operación de distribución, el tiempo

de servicio se ve influenciado por la frecuencia de pedido por cliente y consecuentemente en el tamaño del pedido a ser entregado en cada caso.

Un cliente que pide con alta frecuencia, dos veces a la semana, por ejemplo, demanda una cantidad menor de productos a ser entregados por visita, y como se discutió en el Capítulo III, la cantidad de productos es una característica fundamental en la definición del tiempo de servicio y los mismos representan la mayoría de las visitas individuales que son cada muestra del presenta estudio. Si se analizan algunos casos específicos es más fácil entender los tipos de cliente que se está haciendo referencia en cada caso y comparar con el agrupamiento antes descrito. En la Tabla 1 se muestran algunos ejemplos.

Tabla 1 Ejemplo de tipos de cliente y su tiempo de servicio.

Centro de Distribución	Tipo de Cliente	Tiempo de Servicio	Figura
Pittsburgh	Grande (Walmart)	118 minutos	xx
Pittsburgh	Pequeño/medio (Puesto de gasolina)	23 minutos	xx
Guarulhos	Grande (Makro)	300 minutos	xx
Guarulhos	Pequeño/medio (abasto)	9 minutos	xx
Rosario	Grande (Supermercado)	69 minutos	xx
Rosario	Pequeño/medio (abasto)	6 minutos	xx

En los ejemplos se puede notar la diferencia entre los tipos de clientes en diversas industrias ya la diferencia en su frecuencia, tal y como se explica en la Figura 4.8 de distribución normal doble de Guarulhos y Pittsburg, donde los productos a entregar son distintos.

Con esta explicación se avanza bajo la premisa que los centros de distribución analizados en el presente informe se comportan de manera similar debida a su cartera de clientes. El mejor ajuste con una función continua para el Tiempo de Servicio por Centro de Distribución es la Logarítmica normal.

Para complementar los resultados del análisis se procederá a reportar el valor estimado por cada centro de Distribución y la probabilidad de sobrepasar algunos límites de tiempo, lo cual es interesante para la emisión de alertas sobre duración de tiempo excedido durante la operación en tiempo real.

Tabla 2 Valor Esperado de Tiempo de Servicio y probabilidad acumulada.

Centro de Distribución	Valor Esperado [min]	Prob (x > 5 min)	Prob (x > 15 min)	Prob (x > 30 min)	Prob (x > 45 min)	Prob (x > 60 min)	Prob (x > 90 min)
Rosario	6.43	46.21%	7.36%	1.06%	0.25%	0.08%	0.01%
Guarulhos	9.86	63.39%	17.94%	4.34%	1.47%	0.61%	0.15%
Tarija	6.12	43.84%	6.52%	0.89%	0.21%	0.06%	0.01%
Pittsburgh	32.54	96.00%	68.16%	36.89%	20.99%	12.68%	5.33%

Las probabilidades son claras. Demuestran la realidad de las entregas y su frecuencia en cada CD, por ejemplo, en Pittsburgh para prácticamente todos los casos, los clientes tardan más de 5 minutos en ser atendidos, mientras que, en Tarija Bolivia, menos de la mitad tardan más de 5 minutos. La idea es dependiendo del porcentaje de confianza podrían caracterizar la operación clasificando comportamientos atípicos y clientes especiales, por lo tanto, describir el comportamiento del Tiempo de Servicio para distintas industrias de bienes de consumo bajo una misma distribución es un logro importante, de esta forma se concluye la estimación para el primer modelo.

4.3. Árbol de Decisiones considerando cantidad de producto por Parada.

Para realizar la estimación a partir de árbol de decisiones, se considerará como variable independiente solo el número de productos a ser entregados, para el caso de Tarija, Rosario y Guarulhos, los productos que se reciben en la base de datos Foxtrot son número de cajas. Pittsburgh no envía cantidad de producto, por lo tanto, saldrán de los análisis que necesiten variables independientes como entradas para generar una predicción, en este caso las dos próximas versiones de estimación de tiempo de servicio a través de árbol de decisión solo considerará Tarija, Rosario y Guarulhos.

Guarulhos

A partir de una base de datos con 36623 visitas en diferentes clientes para la ciudad de Guarulhos durante los meses julio y junio de 2018, se construyó un árbol de decisiones definido por las siguientes condiciones:

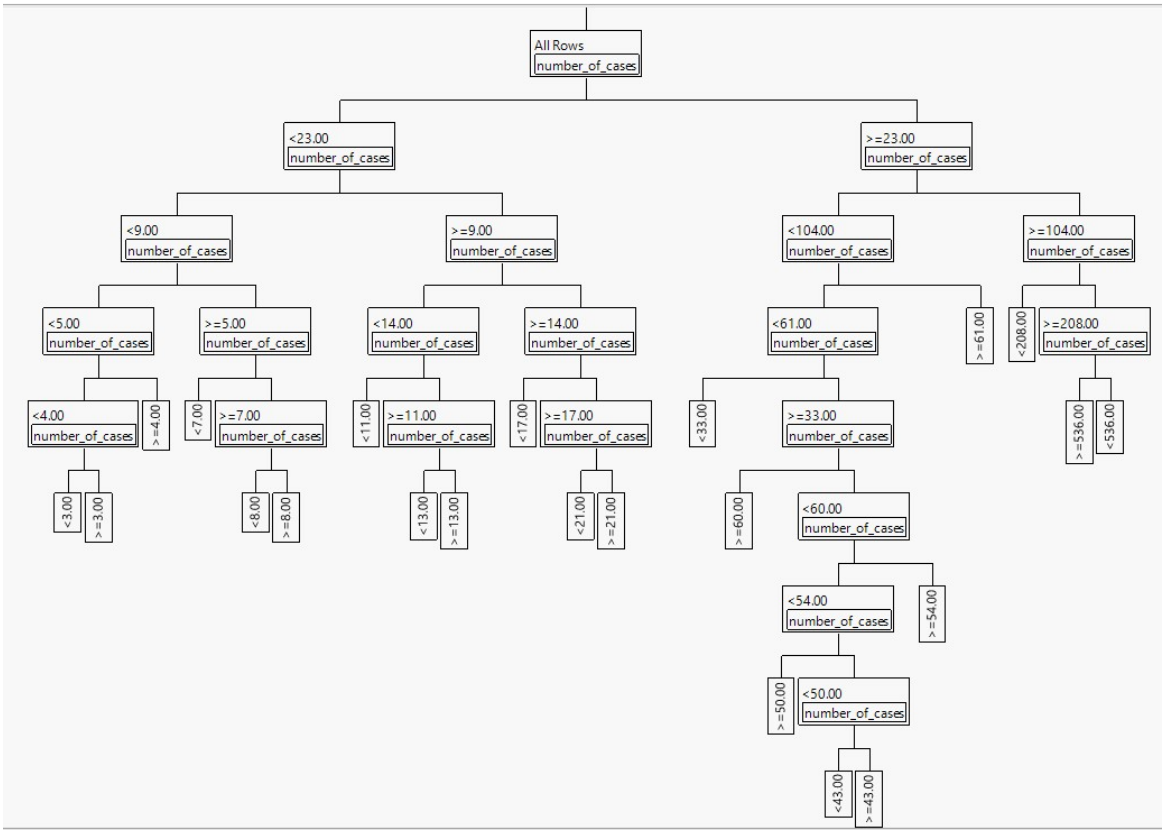

Figura 4.9 Árbol de Decisión por cantidad de paquetes para Guarulhos. Fuente: Elaboración propia.

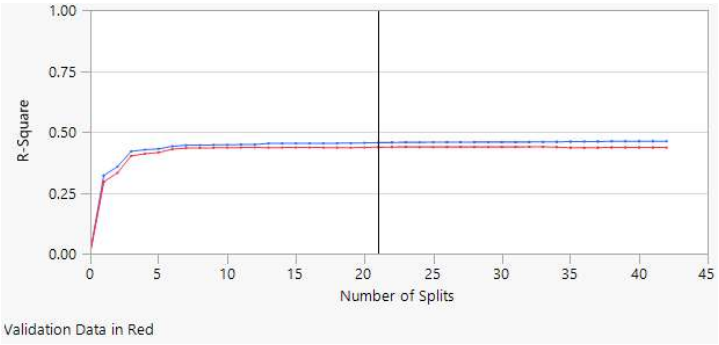

Figura 4.10 R^2 en relación al número de particiones Árbol de Decisión por cantidad de paquetes para Guarulhos.
Fuente: Elaboración propia.

La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,458 con 22 valores finales lo cual se mantiene constante al aumentar la cantidad de particiones (Figura 4.10).

Cada una de esas condiciones arroja un valor diferente, realizando particiones en cada caso, obteniendo al final 22 posibles valores finales o "candidatos" como fue explicado en la sección XX del marco teórico, cada "Mean" representa la cantidad de minutos del valor candidato y cada uno de estos va acompañado de su desviación estándar asociada.

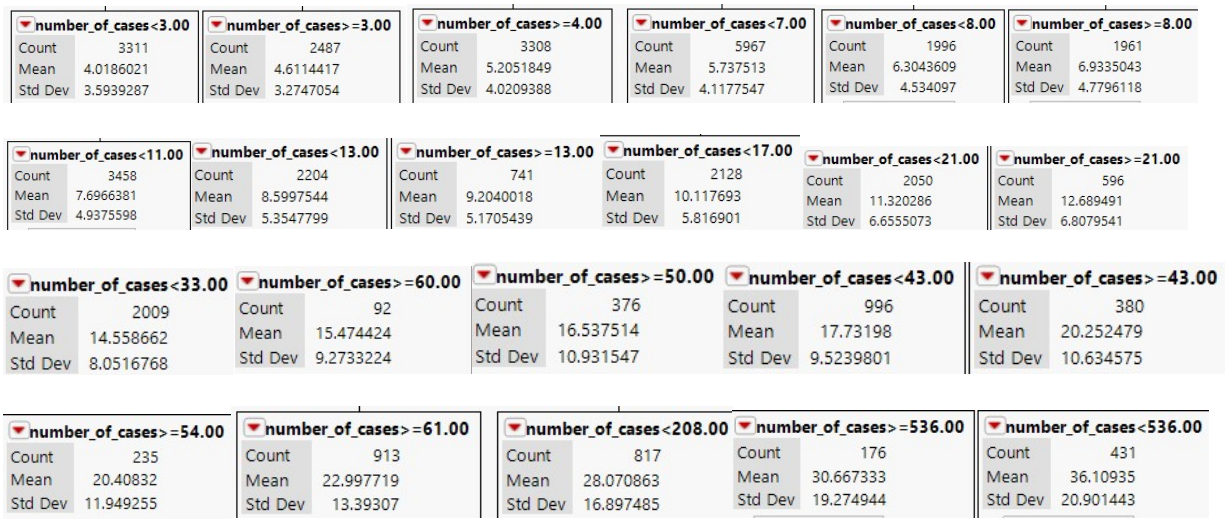

Figura 4.11 Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Garulhos.
Fuente: Elaboración propia.

Cada valor candidato será tomado como predicción dependiendo del resultado del árbol de decisiones generado. El ajuste del árbol se fue haciendo manualmente, recortando o "podando" las ramas que no hacían diferencia hasta que la variación fuera significativa de al menos 30 segundos o la muestra fuera menor a 50 (Count).

Rosario

A partir de una base de datos con 32851 visitas en diferentes clientes para la ciudad de Rosario durante los meses julio y junio de 2018, se construyó un árbol de decisiones definido por las siguientes condiciones:

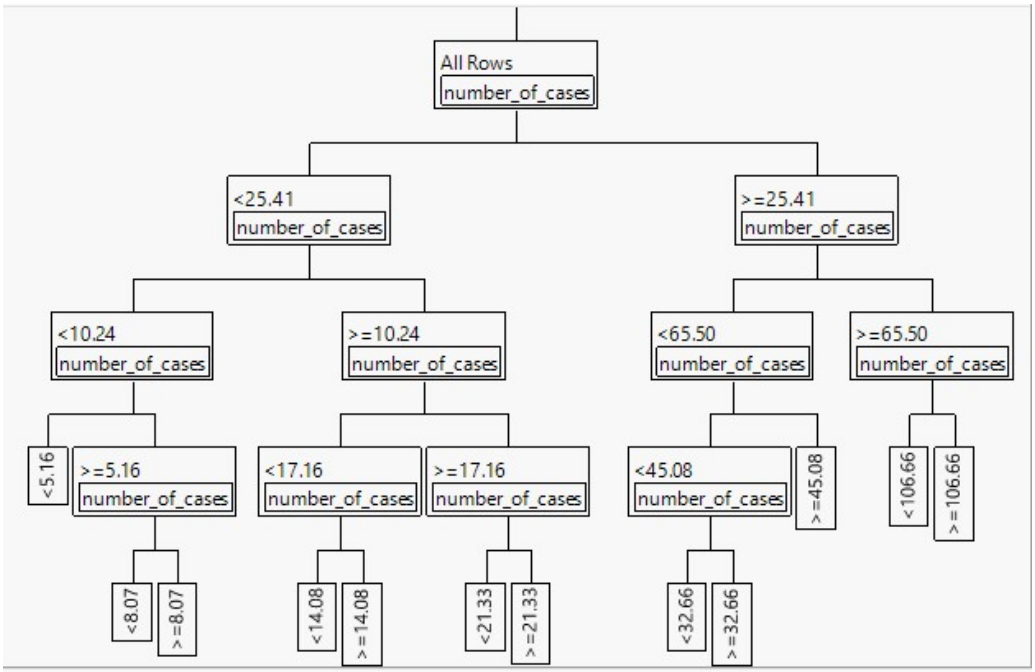

Figura 4.12 Árbol de Decisión por cantidad de paquetes para Rosario. Fuente: Elaboración propia.

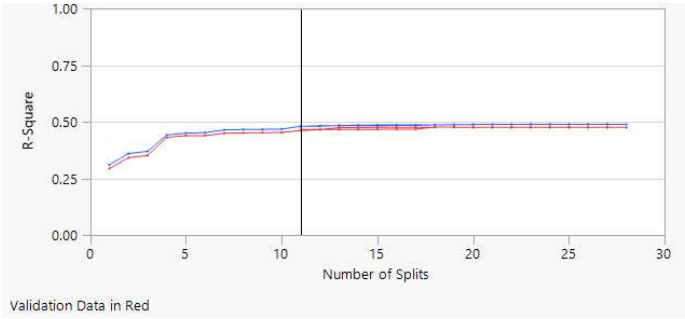

Figura 4.13 R^2 en relación a número de particiones Árbol de Decisión por cantidad de paquetes para Rosario. Fuente: Elaboración propia.

La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,482 con 12 valores finales lo cual se mantiene constante al aumentar la cantidad de particiones (ver figura xx). Obteniendo al final 12 posibles valores finales o “candidatos”:

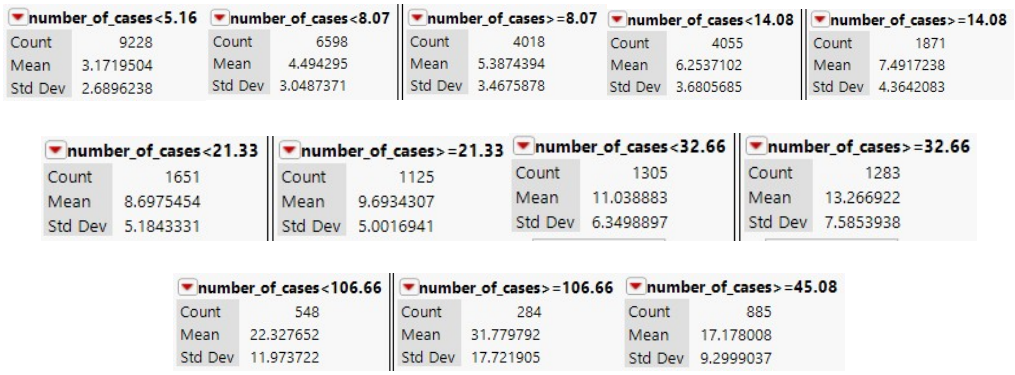

Figura 4.14 Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Rosario. Fuente: Elaboración propia.

Tarija

A partir de una base de datos con 11100 visitas en diferentes clientes para la ciudad de Tarija durante los meses julio y junio de 2018 se construyó un árbol de decisiones definido por las siguientes condiciones:

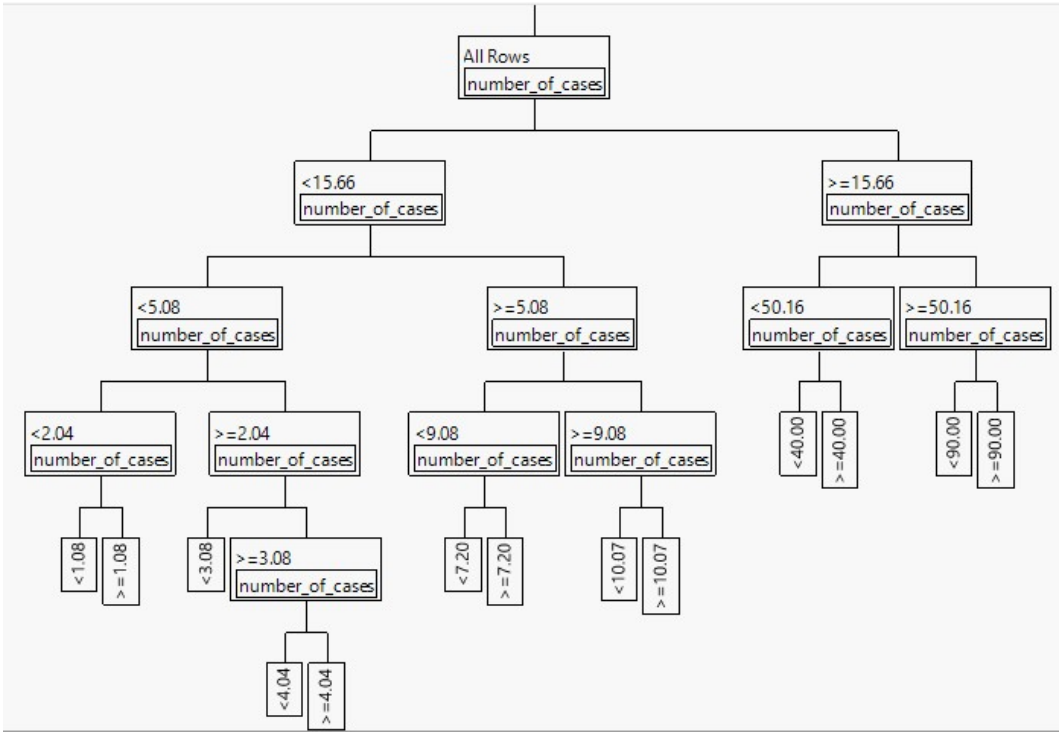

Figura 4.15 Árbol de Decisión por cantidad de paquetes para Tarija. Fuente: Elaboración propia.

.

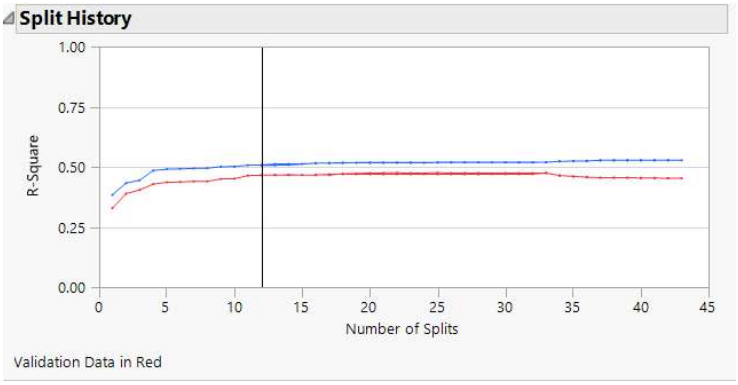

Figura 4.16 R^2 en relación a número de particiones Árbol de Decisión por cantidad de paquetes para Tarija. Fuente: Elaboración propia.

La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,509 con 13 valores finales lo cual se mantiene constante al aumentar la cantidad de particiones hasta empeorar después de 33 candidatos (ver Figura 4.16).

Obteniendo al final 13 posibles valores finales o "candidatos":

number_of_cases<1.08	
Count	2607
Mean	3.2534701
Std Dev	3.0197905

number_of_cases>=1.08	
Count	2710
Mean	4.0946988
Std Dev	3.2031974

number_of_cases<3.08	
Count	1705
Mean	4.9663338
Std Dev	3.6138762

number_of_cases<4.04	
Count	1047
Mean	5.8869936
Std Dev	4.0984329

number_of_cases>=4.04	
Count	736
Mean	6.4173231
Std Dev	4.1821832

number_of_cases<7.20	
Count	749
Mean	7.7358885
Std Dev	4.953882

number_of_cases>=7.20	
Count	347
Mean	9.0084056
Std Dev	5.5339505

number_of_cases<10.07	
Count	178
Mean	10.12832
Std Dev	6.3031878

number_of_cases>=10.07	
Count	291
Mean	12.190723
Std Dev	8.0973572

number_of_cases<40.00	
Count	252
Mean	16.368905
Std Dev	10.189326

number_of_cases>=40.00	
Count	175
Mean	22.197627
Std Dev	12.844892

number_of_cases<90.00	
Count	161
Mean	27.360115
Std Dev	14.178542

number_of_cases>=90.00	
Count	142
Mean	33.998279
Std Dev	18.954637

Figura 4.17 Valores Candidatos en Árbol de Decisión por cantidad de paquetes para Tarija. Fuente: Elaboración propia.

Con los árboles y valores candidatos para los tres CDs los cuales envían productos a la base de datos Foxtrot, se da por finalizada la estimación de la versión 2 del modelo.

4.4. Árbol de Decisiones considerando cantidad de producto por Parada y conductor de camión.

Para realizar la estimación a partir de árbol de decisiones, consideraremos como variable independiente solo el número de productos a ser entregados y el conductor a realizar la entrega, para el caso de Tarija, Rosario y Guarulhos, los productos que se reciben en la base de datos Foxtrot son número de cajas por entregados.

Guarulhos

Para la ciudad de Guarulhos se construyó un árbol de decisiones definido bajo la siguiente estructura observada en la Figura C.1, la cual muestra cada una de las particiones, indicando los conductores que entran en esa rama con su respectiva condición de cantidad de cajas a ser entregadas. Donde los recuadros pequeños son ramas generadas por número de cajas y los recuadros grandes contienen el nombre de los conductores que pertenecen a ese grupo de decisión.

La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,472 con 27 valores finales lo cual se mantiene casi constante al aumentar la cantidad de particiones (ver Figura 4.18). Si se observa el valor de la suma de cuadrados "SS" es posible concluir que aproximadamente el 96% de la construcción del árbol es debido a la varianza del tiempo de servicio a medida que se hacían particiones por número de producto a ser entregado, mientras que la varianza de las particiones por grupos de conductores representa aproximadamente un 4% del total, este comportamiento se repetirá con el resto de los centros de distribución, lo que hace pensar que la variable independiente "Conductor" no impacta significativamente en la predicción del tiempo de servicio.

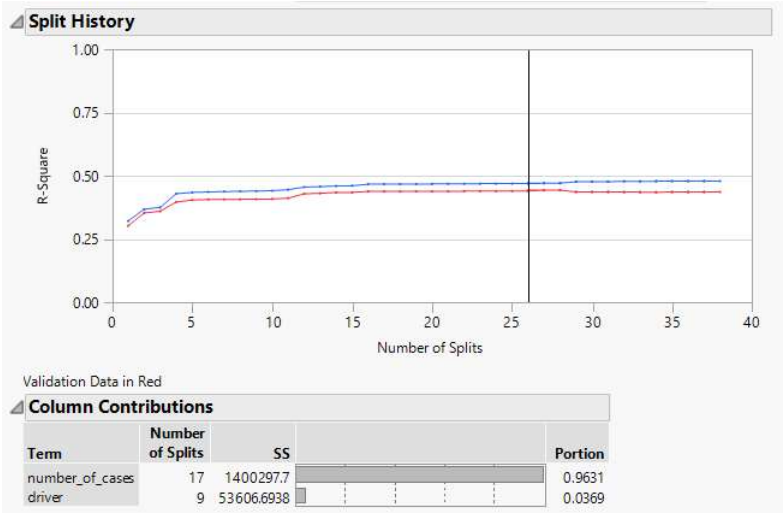

Figura 4.18 R^2 en relación a número de particiones Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos. Fuente: Elaboración propia.

Finalmente se obtienen 26 posibles valores finales o "candidatos" que se muestran en la Figura 4.19. La desviación estándar es alta en la mayoría de los casos, por lo tanto, la confianza de acierto en cada uno de los candidatos no es exacta, esto refleja la dificultad de modelar este tipo de comportamientos que están expuesto a factores aleatorios, además que no todas las variables determinísticas definidas en la sección 3.1 pudieron en ser consideradas como variable independiente en la creación del árbol de decisiones.

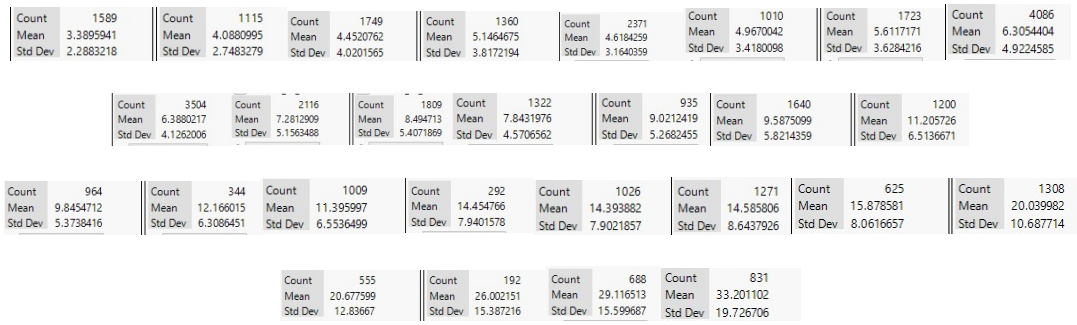

Figura 4.19 Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos. Fuente: Elaboración propia.

Rosario

Para la ciudad de Rosario se construyó un árbol de decisiones definido bajo estructura presentada en la Figura C2. La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,512 con 29 valores finales lo cual se mantiene casi constante al aumentar la cantidad de particiones (Figura 4.19). La contribución de cada variable independiente es casi igual que en el caso Guarulhos, por lo tanto, se demuestra un patrón en la distribución de la suma de cuadrados.

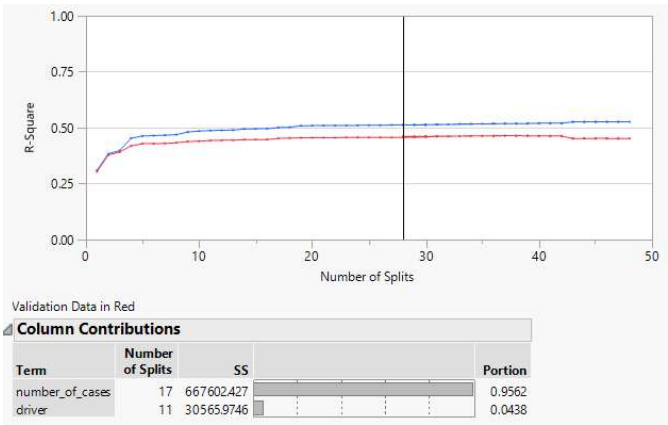

Figura 4.20: R² en relación a número de particiones Árbol de Decisión por cantidad de paquetes Fuente: Elaboración propia.

Finalmente se obtienen 29 posibles valores finales o “candidatos”:

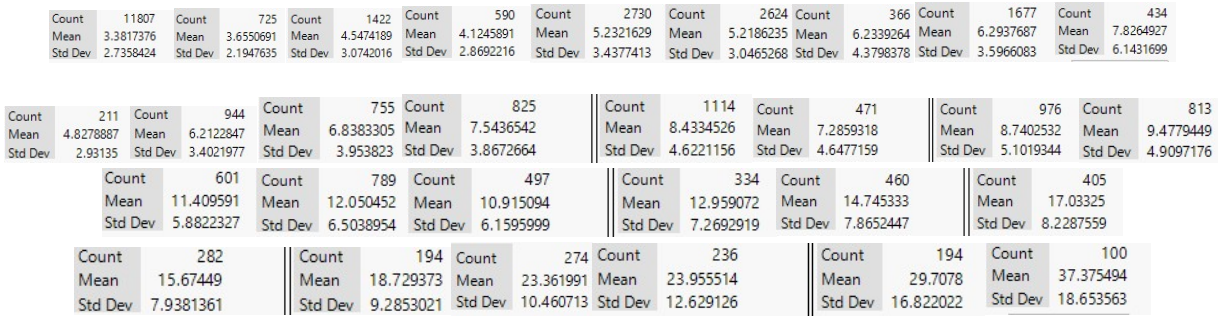

Figura 4.21: Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Rosario. Fuente: Elaboración propia.

Tarija

Para la ciudad de Tarija se construyó un árbol de decisiones definido bajo la siguiente estructura definida en la Figura 4.22.

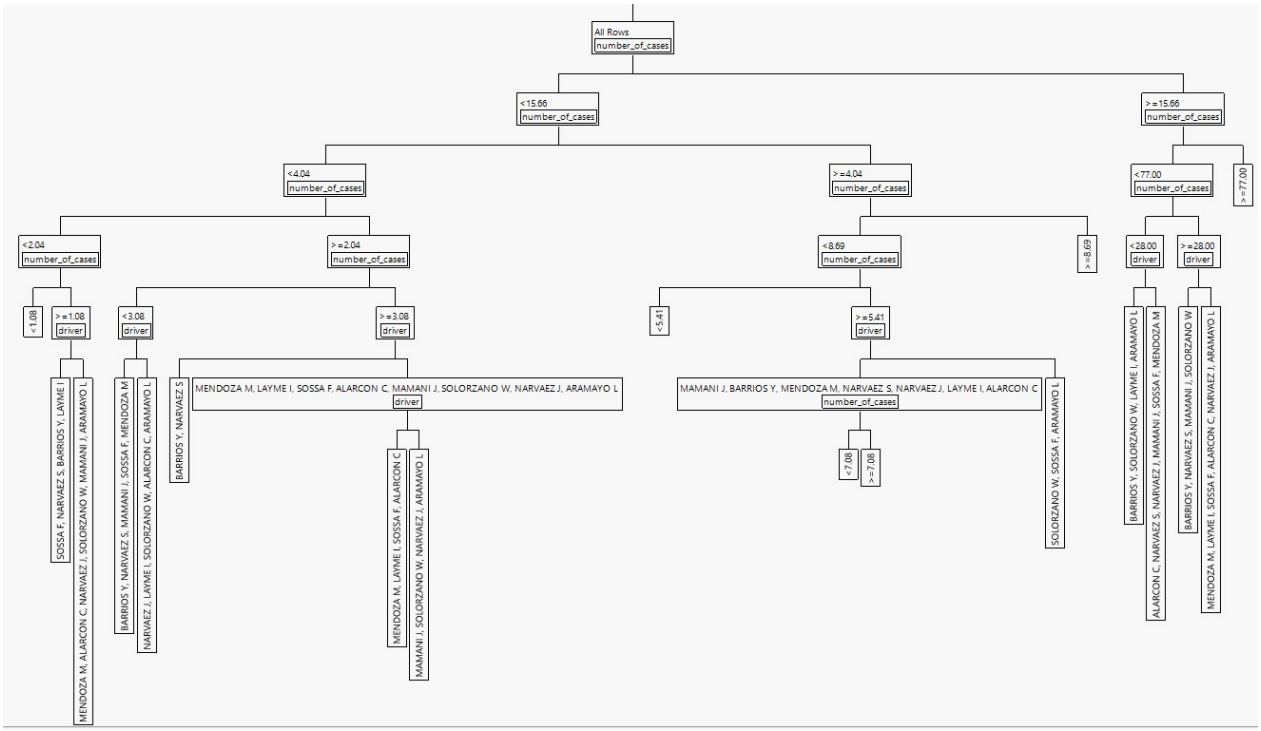

Figura 4.22 Árbol de Decisión por cantidad de paquetes y conductor para Tarija. Fuente: Elaboración propia.

Se puede observar un árbol de decisiones más pequeño que los analizados anteriormente, esto debido a la cantidad de datos disponibles, como se mencionó anteriormente en estos modelos no se admitieron divisiones para definir candidatos a partir de muestras menores 50 y con diferencia de al menos 30 segundos en los dos candidatos de la división.

Por la restricción de las muestras principalmente, es notorio que la complejidad del modelo o la cantidad de ramas que tiene el árbol está fuertemente relacionado con la cantidad de datos que se tenga.

La cantidad de particiones realizadas consiguió alcanzó un R2 = 0,508 con 18 valores finales lo cual se mantiene casi constante al aumentar la cantidad de particiones (ver Figura 4.23). Si bien la cantidad de divisiones por causa de cada variable independiente es casi igual, la suma de cuadrados es casi 99% por el número de cajas, por lo tanto, se reafirma como la variable más significativa.

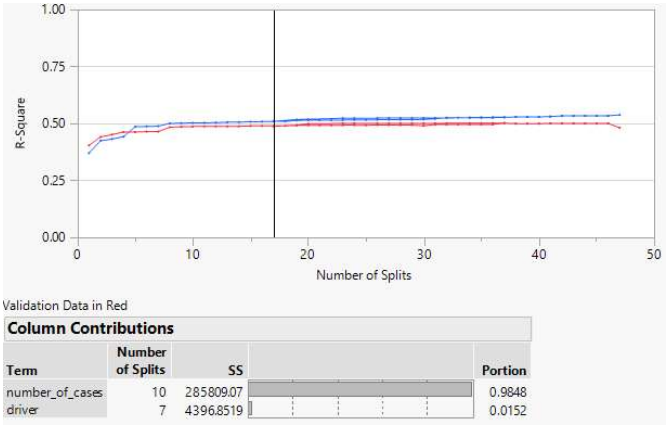

Figura 4.23 R^2 en relación a número de particiones Árbol de Decisión por cantidad de paquetes y conductor para Tarija. Fuente: Elaboración propia.

Finalmente se obtienen 18 posibles valores finales o "candidatos", que se muestran en la Figura 4.24.

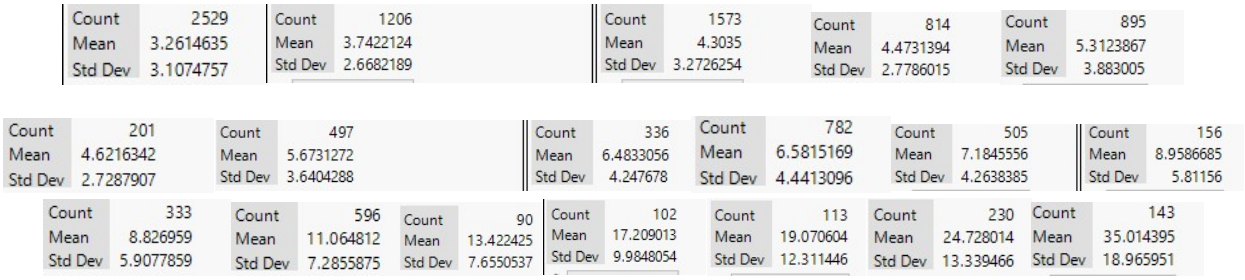

Figura 4.24 Valores Candidatos en Árbol de Decisión por cantidad de paquetes y conductor para Tarija. Fuente: Elaboración propia.

4.5. Promedio de tiempo de parada histórico para cada cliente.

Para esta versión, se consideraron los promedios de los últimos tiempos de parada para cada ID de cliente, dando como resultado un valor estimado para comparar con la nueva base de datos siempre y cuando dichos clientes fueran visitados nuevamente.

Bajo esta premisa se obtuvo una muestra de 340 predicciones (visitas a simular) para comparar para Pittsburgh, 26400 para Guarulhos, 27913 para Rosario y 8320 para Tarija durante los meses de junio y julio. Los resultados fueron obtenidos bajo el formato mostrado en la Figura 4.25 para cada CD comparando la media de la duración de las visitas en la muestra de modelaje con las visitas de prueba del mes.

ID del Cliente	Cantidad De visitas	Predicción: Promedio en Minutos del Tiempo de Parada en Visitas Anteriores
899901	22	12.59
899733	22	7.07
899590	11	5.51
899358	32	7.60
898918	9	5.51
898110	5	5.06
897817	16	7.32
897801	17	5.23
897784	15	5.62
897722	19	5.04
897709	13	5.43
897704	9	9.69
897693	10	7.56
897163	28	6.12
897156	9	5.16
897087	19	7.16
897075	14	8.70
896582	21	5.00
896495	11	6.01
896469	20	7.20
896078	33	31.26
896057	17	6.01
896039	14	5.45
896003	16	5.92

Figura 4.25 Ejemplo de predicciones obtenidas por ID de cliente. Fuente: Elaboración propia.

A continuación, se evaluará cada modelo con respecto a los resultados obtenidos al comparar su Predicción con los datos de prueba de agosto de 2018.

4.6. Evaluación y Comparación

Resultado General:

Para la comparación general de resultados, se calcularon los minutos de dispersión de cada predicción con respecto al valor real de agosto, esta diferencia será referida como un Error de estimación definido como:

$$Error = Tiempo\ de\ Parada\ Real - Tiempo\ de\ Parada\ Estimado\ , \quad (3.4)$$

y su distribución para las operaciones de bebidas es mostrada en la figura XX, el caso de Pittsburgh será analizado aparte debido a que en este Centro de Distribución no fue posible realizar Árbol de Decisiones debido a la falta de datos de entrada. Para entender la oportunidad de mejora con los nuevos modelos presentados, se realizó también la comparación con el tiempo de servicio enviado por el cliente.

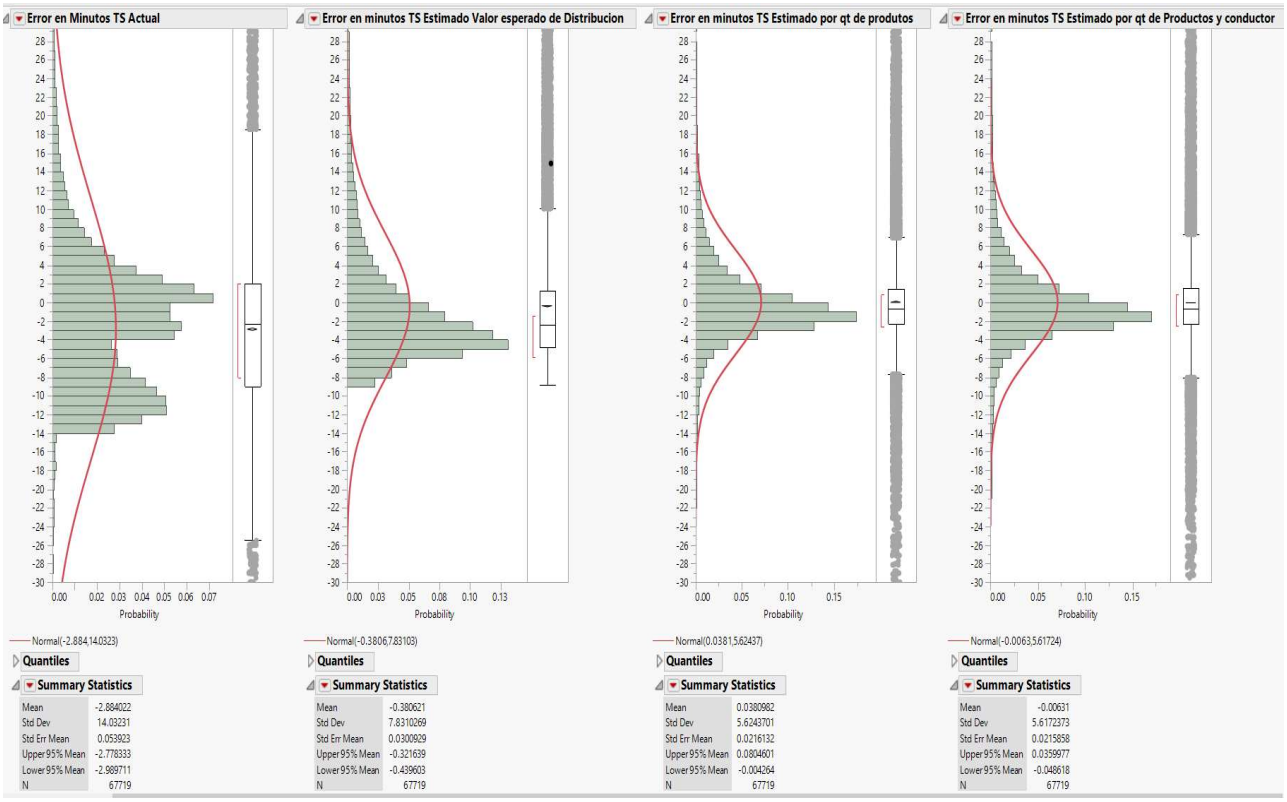

Figura 4.26 Distribución de errores por cliente para la industria de bebidas. Fuente: Elaboración propia.

Observando de lado izquierda a derecha la Figura 4.26, se ve como el modelo se va refinando, mejorando la distribución de errores con cada versión, el objetivo principal sería llevar esta distribución a una distribución con media lo más cercana a cero posible y una desviación estándar mínima.

La primera distribución es el tiempo de servicio enviado a Foxtrot por parte de los Centros de Distribución, es decir, es el tiempo utilizado actualmente para realizar la planificación de ruta, asignar conductores, etc; el mismo tiene un error medio de -2.88 minutos por cliente, quiere decir que en la mayoría de los casos el tiempo de parada está sobre estimando 2.88 minutos por cada cliente (Ecuación 3.4), para un tiempo de parada medio calculado para el área de bebidas que va desde 6 a 9 minutos por parada en un cliente, representa un desvío considerable. Más grave aún es la desviación estándar que se pueda apreciar claramente en la dispersión del gráfico con 14 minutos por cliente, concluyendo así que la situación actual es totalmente imprecisa para el proceso de distribución, tanto en planificación como en ejecución.

En la segunda distribución se evalúa la versión 1 de estimación, que hace referencia la predicción usando los parámetros de la distribución Logarítmica Normal para cada Centro de Distribución, el mismo ya da una mejora significativa con respecto al valor actual utilizado donde la media del error pasa a ser de -0.38 minutos por cliente.

En la tercera distribución de errores se evalúa la versión 2 que hace referencia al árbol de decisión generado a partir de la cantidad de producto a ser entregado, los resultados muestran que el error

fue en General 10 veces mejor que la versión 1 con una media de 0,038 y una reducción de la desviación estándar a 5,62 minutos por cliente.

En la cuarta distribución de errores se evalúa la versión 3 que hace referencia al árbol de decisión generado a partir de la cantidad de productos a ser entregados y el conductor/equipo de reparto que realizará las entregas, este modelo mejoró la posición de la media acercándola al valor cero.

Para finalizar el análisis general tenemos la comparación con el modelo de media por cliente (Figura 4.27), el cual no superó los resultados de los árboles de decisión, pero presenta una mejor estimación que la obtenida a partir de la distribución (versión 1) sin embargo la desviación estándar es bastante grande.

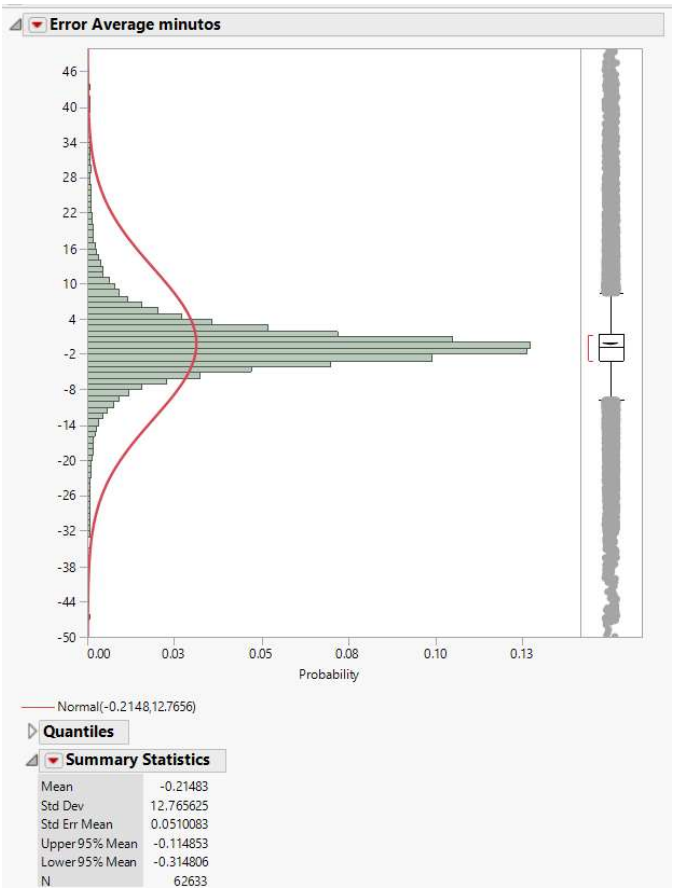

Figura 4.27 Distribución de Errores General por Cliente considerando la 4ta Versión de estimación (Media por cliente). Fuente: Elaboración propia.

Una vez analizados los resultados generales, se presentarán los resultados de la evaluación por Centro de Distribución a continuación:

Guarulhos

Para Guarulhos el resultado fue bastante similar al obtenido en la visión general, siendo la versión 3 la de mejor resultados para la estimación.

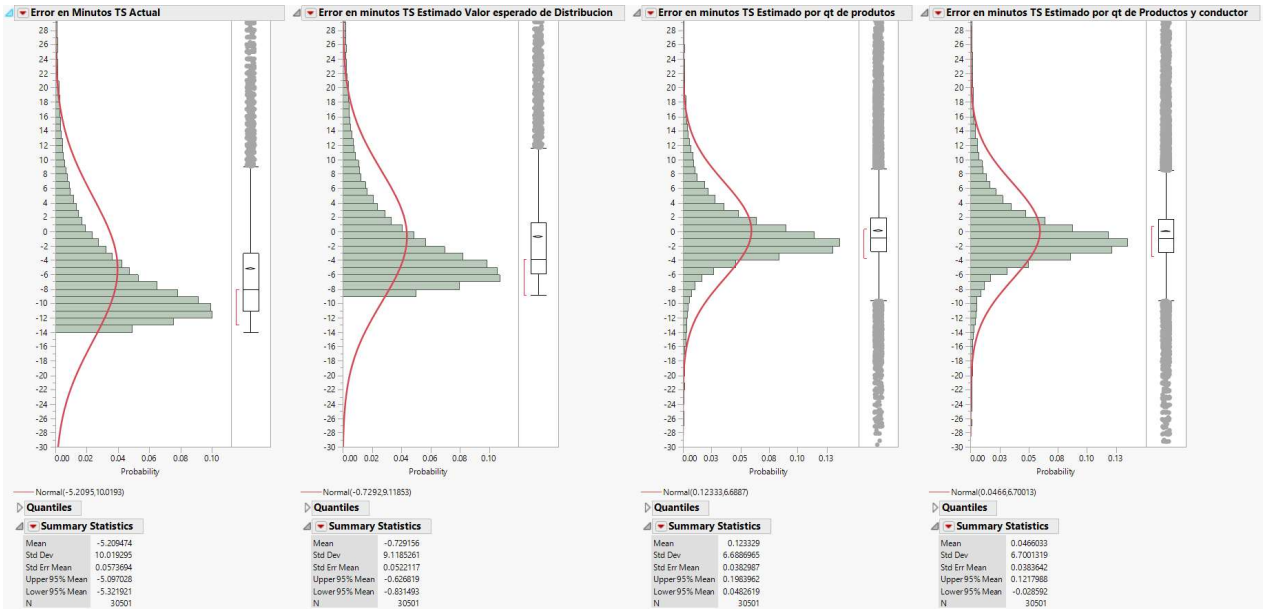

Figura 4.28 Distribución de Errores por Cliente para Guarulhos – de Izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3. Fuente: Elaboración propia.

.

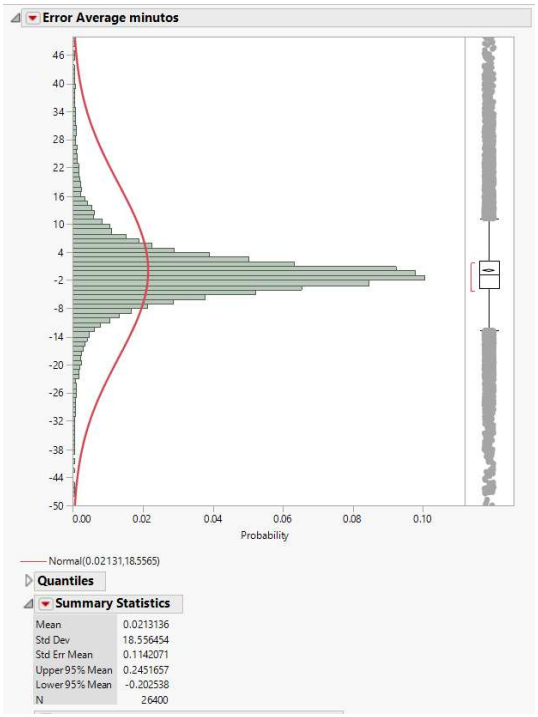

Figura 4.29 Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Guarulhos. Fuente: Elaboración propia.

La desviación estándar sigue siendo considerable, pero la media acercándose al cero a medida que se van refinando los modelos es notorio y representa una mejora bastante significativa con respecto a los tiempos de servicio enviados, que en su mayoría eran de 15 minutos (aparentemente un valor

por defecto) lo cual es más de la media del CD que es de aproximadamente 9 minutos, por lo tanto el resultado es una distribución corrida hacia el lado negativo con una media de -5,2 minutos de exceso lo cual tiene total sentido con los datos obtenidos en la distribución Lognormal para ese CD.

La estimación por media sigue teniendo el mismo problema con la desviación estándar el cual será discutido más adelante.

Rosario

Para el caso Rosario, los resultados son similares, solo que la versión 3 que considera el conductor no tuvo mejores resultados que la versión 2, al menos en la media del error, al considerar solo los productos el resultado fue levemente mejor.

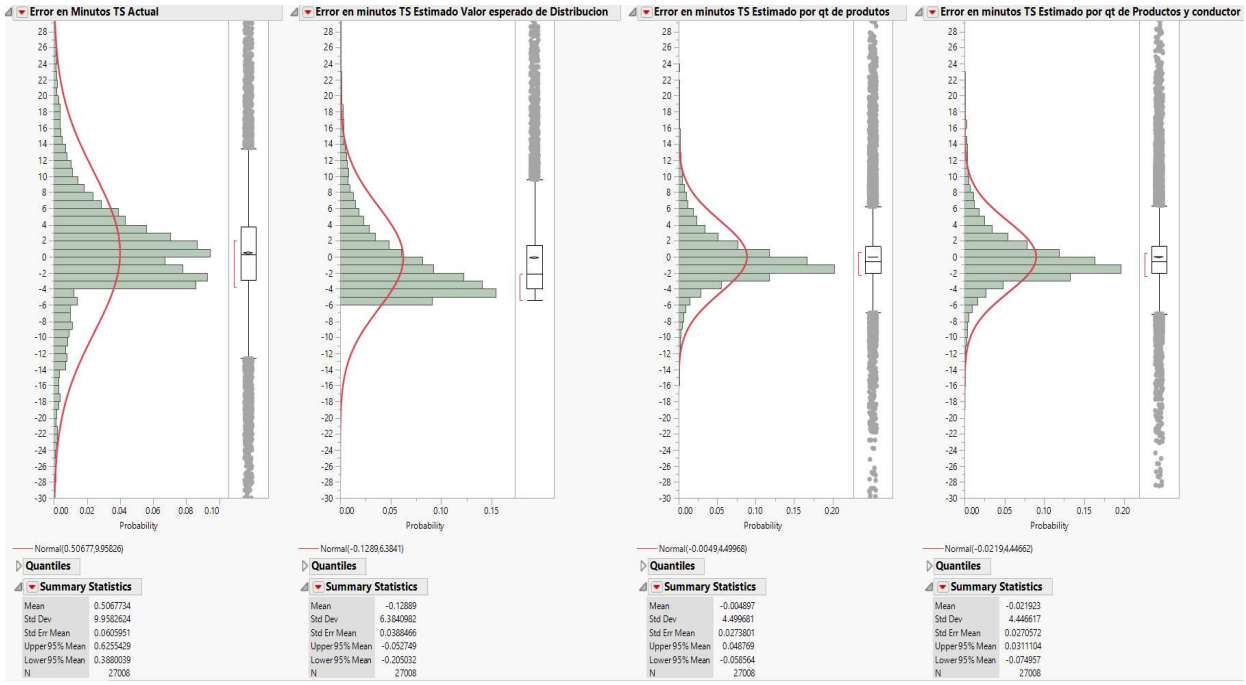

Figura 4.30 Distribución de errores por cliente para Rosario – de izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3. Fuente: Elaboración propia.

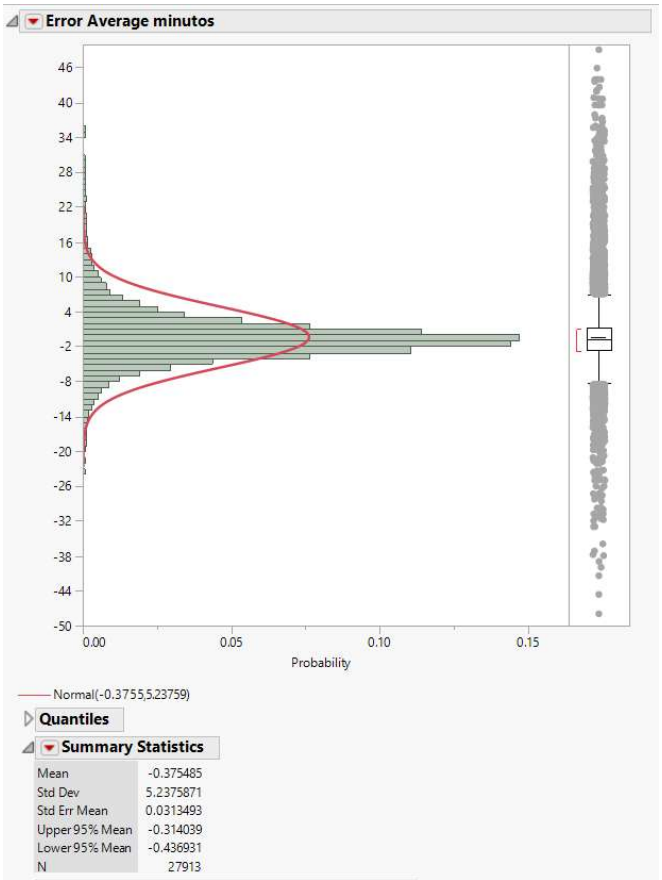

Figura 4.31 Distribución de errores por cliente considerando la 4ta versión de estimación (Media por cliente) para Rosario. Fuente: Elaboración propia.

Tarija

En Tarija se observan los peores resultados de Tiempo de Servicio Actual, con una gran sobreestimación de minutos por parada y carencia de uniformidad en los datos. En este caso los parámetros de la Distribución Lognorm Trabajan bastante bien para dar una estimación, parecen dar el mejor resultado de todas las versiones, sin embargo, su desviación estándar bastante alta en comparación a su tiempo medio de parada, por lo tanto, se pueden considerar los árboles de decisiones como una mejor opción para la estimación.

Con respecto a la media por clientes, observamos el mismo problema que en los otros CDs.

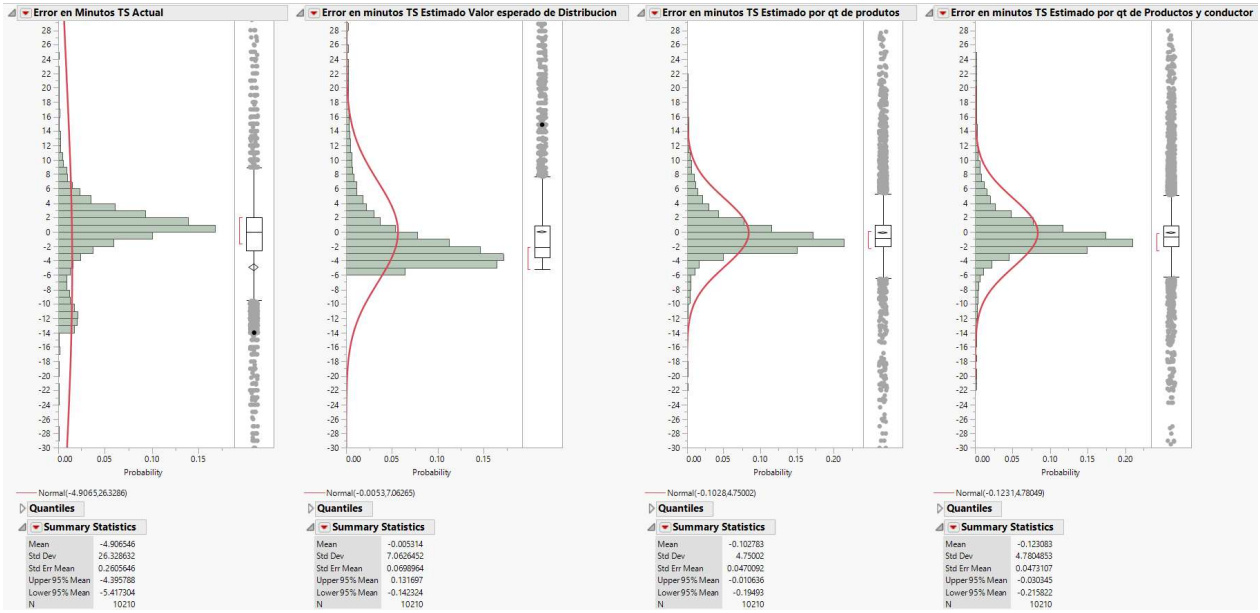

Figura 4.32 Distribución de errores por cliente para Tarija – de Izquierda a derecha, Datos del cliente, Estimación con Versión 1, Estimación con Versión 2 y por último Estimación con Versión 3. Fuente: Elaboración propia.

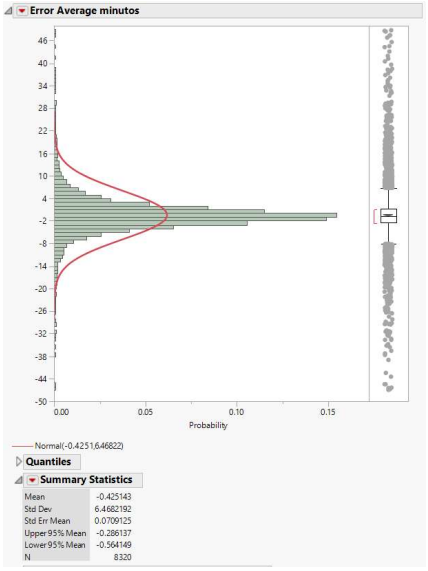

Figura 4.33 Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Tarija. Fuente: Elaboración propia.

4.6.1. Caso Pittsburgh

Para Pittsburg se comparan solo el modelo actual contra la versión 1 y la versión 4.

Es interesante debido a que los tiempos de Servicio Actuales parecen dar buen resultado en comparación a los otros modelos, los mismos son personalizados por los supervisores del Centro de Distribución al momento de registrar un nuevo cliente por primera vez sin ningún análisis cuantitativo previo.

Siendo mejor el Tiempo de Servicio del cliente, es necesario pensar en un posible periodo de evaluación de los tiempos de servicio actuales antes de usar un nuevo cálculo proveniente de Foxtrot, sin embargo, para el caso Pittsburgh no fue posible evaluar las 2 versiones que dieron mejor resultado en las otras ciudades.

Sabiendo que los tiempos de servicios se comportaban con una distribución Lognormal, los inputs actuales parecen más personalizados, con una media menor y una desviación estándar menor. En este caso un simple parámetro de estimación por esperanza de una Lognorm no fue suficiente y la media por clientes dio un mejor resultado.

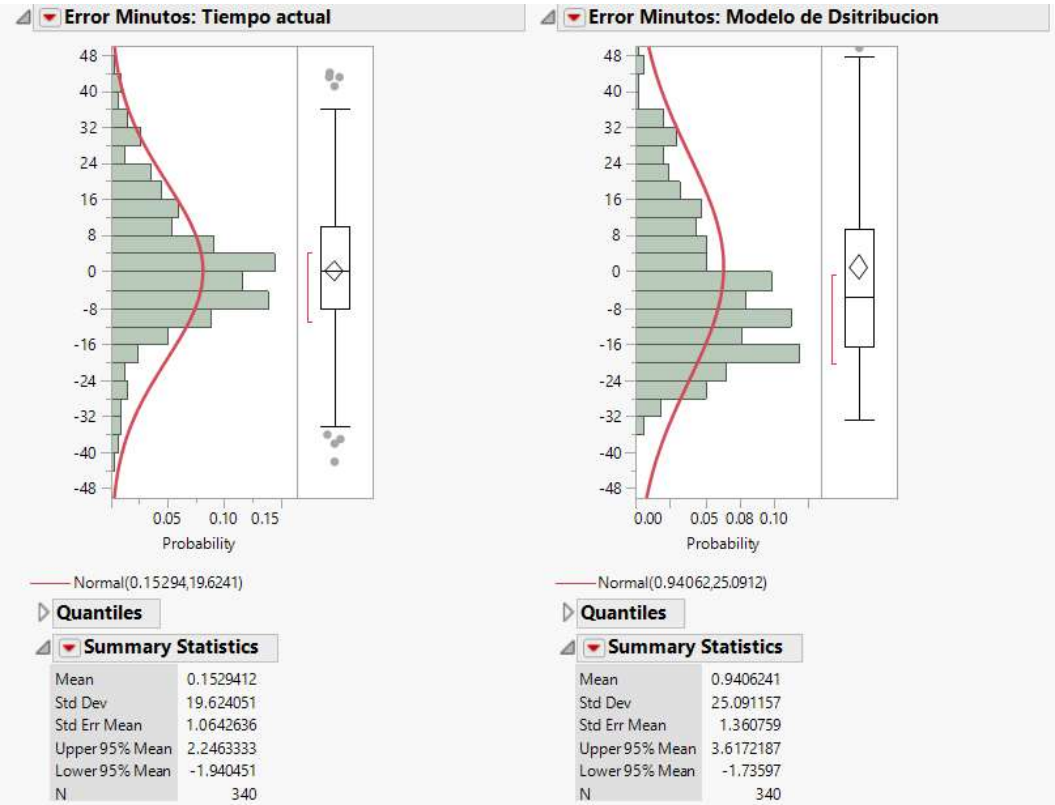

Figura 4.34 Distribución de errores por cliente para Pittsburgh – En la Izquierda para los tiempos enviados por el cliente y a la derecha para la Estimación con la Versión 1. Fuente: Elaboración propia.

El error para el tiempo de servicio enviado por el cliente tiene una media de 0,15 minutos y una alta desviación estándar de 19,62 minutos, mientras que para el modelo de distribución Lognormal presentó una media de 0,94 minutos con una desviación de 25,09 minutos.

Cuando se compara con el modelo de media por cliente, la deviación estándar es menor, pero los valores la media de error es de -1,57 minutos.

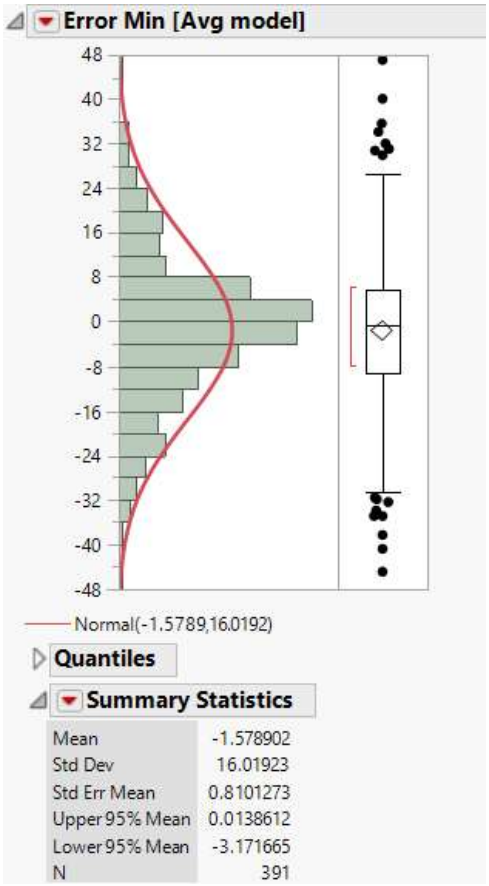

Figura 4.35 Distribución de errores por cliente considerando la 4ta Versión de estimación (Media por cliente) para Tarija. Fuente: Elaboración propia.

.

4.7. Ventaja de Errores Distribuidos normalmente

Cuando se realizan estimaciones de tiempo de servicio, los errores por cliente son acumulativos y terminan generando un error por tiempo total en ruta. Este error en tiempo total en ruta se ve compensado dependiendo de la distribución de errores por cliente como fue discutido en la sección 3.4, por lo tanto, si la distribución de errores es normal, las sobreestimaciones se compensan con las subestimaciones generando un error total de ruta menor. Si consideramos un error que no sea normalizado si no por ejemplo una distribución Lognorm como de la figura 4.4 o cualquiera de los gráficos que definen el tiempo de servicio, la estimación estaría sobreestimando el tiempo de parada por cliente en la mayoría de los casos, generando un error total en ruta que acumula todas las sobrestimaciones. Por ejemplo, imaginando que una ruta tiene 5 clientes y se tienen dos estimaciones diferentes, una con el error normalmente distribuido y la otro en base a una lognorm dando los siguientes errores:

Tabla 3 Ejemplo para entender los beneficios de un error distribuido normalmente.

Cliente	Error Estimación 1 (Normal)	Error Estimación 2 (Lognorm)
A	-3	1
B	2	0,5
C	-1	1,2
D	3	1,3
E	-2	0,7
Total (suma)	-1 minutos	4,7 minutos

En la Tabla 3 se muestra claramente que siendo el error absoluto menor para la estimación 2 por cada cliente, al no estar distribuido normalmente termina afectando mucho más la estimación para el tiempo de parada total por ruta, siendo para la primera estimación un error de -1 minuto y para la segunda estimación un error de 4,7 minutos de sobreestimación.

Por este motivo, se valoriza que la distribución además de tener una media cercana a cero y con una desviación estándar mínima, sea distribuida normalmente para aumentar la precisión por ruta, pudiendo planificar mejor las jornadas diarias de entrega, sabiendo que aproximadamente el 80% del tiempo en ruta es tiempo de parada y solo el 20% es tiempo de manejo en ruta (Foxtrot 2018), una ruta con tiempos de servicio previos precisos tendrá una mayor exactitud en el tiempo estimado de llegada del conductor en el Centro de Distribución.

4.8. Promedio Histórico: Error por efecto látigo o estacionalidad comercial

El promedio histórico por cliente fue la versión 4 evaluada, la cual en ninguno de los casos fue la de mejor resultados. Para el promedio por cliente se esperaban mejores resultados, dado que solo se consideraron frecuencia mínima de 5 visitas para realizar el cálculo, al mínimo 5 visitas en un cliente podría ser una media que caracterice cada cliente.

Para entender en la imprecisión de la estimación se realizó un análisis de frecuencia y su relación con el error medio en minutos. Aparentemente a medida que aumenta la muestra el error es mayor como se muestra en la Figura 4.36.

Para este análisis se evaluó el error para número de visitas con al menos 10 (para tener un espacio muestral mayor), con una frecuencia mínima de 5 clientes para realizar la estimación.

Esta relación muestra a medida que aumenta la cantidad de visitas se observa un efecto látigo relacionado con la cantidad de visitas consideradas para calcular la media. Ya se había determinado la relación entre la cantidad de producto a ser entregado y el tiempo de parada por cliente, en este caso a medida que se consideran más visitas, llega un momento que el error aumenta considerablemente después de tantas visitas, esto podría estar relacionado al efecto látigo o algún tipo de estacionalidad. Los clientes visitados con Foxtrot son clientes finales o distribuidores y no parten de la manufactura (T1), se trata de T2 o last mile, donde se siente el efecto látigo en la cantidad de productos ordenados, lo que podría empieza a generar comportamientos diferentes en el tiempo de servicio.

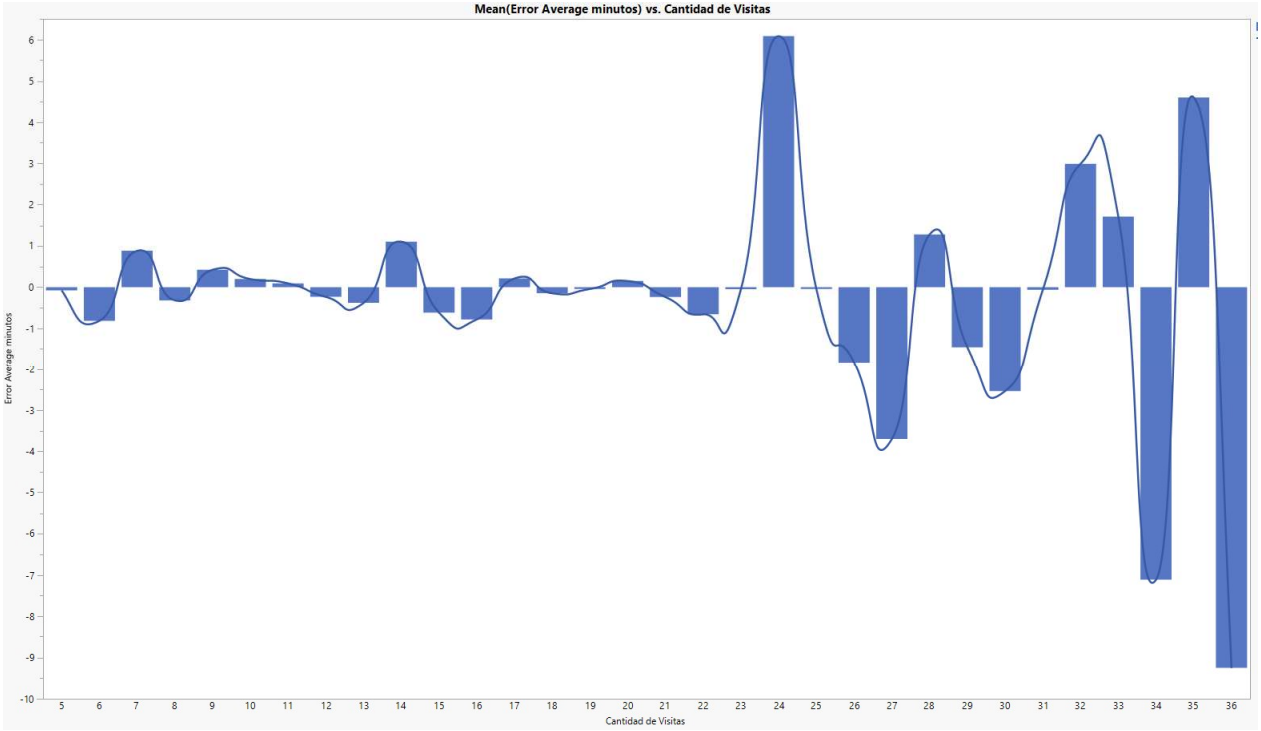

Figura 4.36 Error medio vs número de visitas consideradas para el cálculo de la estimación. Fuente: Elaboración propia.

Interpretando la gráfica de Guarulhos observamos el incremento en la dispersión a partir de la consideración de 24 visitas o más, por lo tanto existe alguna estacionalidad o ruido por lo que no es favorable considerar todas las muestras para calcular la media de tiempo de servicio.

Esta es solo una hipótesis de por qué el tiempo de servicio estimado a través de medias históricas debería tener un horizonte más corto para reducir el efecto látigo o estacionalidad que genera dispersiones en el tiempo de parada.

Si se realiza un análisis más profundo sería posible llegar a una oportunidad de mejora para esta metodología, al considerar una media móvil o algo similar, sin embargo los resultados alcanzados por los árboles de decisiones son aceptables por parte de Foxtrot, por lo tanto no se agregará otro modelo en el estudio.

Finalmente se realizará una comparación de los indicadores principales por Centro de Distribución.

4.9. Tabla comparativa Final

En la Tabla 4 semuestra la comparación en los valores de errores para cada modelo.

Tabla 4 Comparación de resultados por Centro de Distribución.

Centro de Distribución	Media de Tiempo de servicio	Error Medio TS Actual y Desvest [min]	Error Medio TS V1 y Desvest [min]	Error Medio TS V2 y Desvest [min]	Error Medio TS V3 y Desvest [min]	Error Medio TS V4 y Desvest [min]
Guarulhos	9,26	5,21	0,73	0,12	0,04	0,02
		10,02	9,12	6,69	6,70	18,55
Rosario	6,47	5,51	-0,12	-0,005	-0,02	-0,37
		9,96	6,38	4,50	4,44	5,24
Tarija	6,29	-4,90	-0,005	-0,10	-0,12	-0,42
		26,32	7,06	4,75	4,78	6,47
Pittsburg	29,95	0,15	0,94	NA	NA	-1,58
		19,62	25,09	NA	NA	16,02

Observando la Tabla 4 es posible concluir que el mejor modelo es el dos, siendo que el tres tiene más demanda de cálculo y no mejora significativamente con respecto al modelo dos.

La desviación estándar sigue siendo alta para todos los casos, por lo que los modelos genéricos siempre tienen un alto grado de imprecisión, en la mayoría de los casos podría ser más del 50% del valor medio del tiempo de parada, lo cual demuestra que la calidad de predicción no es tan alta, sin embargo es mucho mejor que las estimaciones que se usan actualmente por las 4 empresas evaluadas, donde la Empresa comercializadora de panes en Pittsburgh tiene los mejores tiempos de servicio gracias a la estimación experta de su área de logística, pero solo comparado con los modelos 1 y 4 los cuales no tuvieron buenos resultados en las otras ciudades, por lo que se podría esperar que el mismo modelo de productos consiga resultados mejores; como próximo paso se solicitarán la cantidad de productos para comenzar a crear una base de datos para la construcción del árbol de decisión para Pittsburgh también.

Por el caso Pittsburgh se considerará un periodo de evaluación para los datos del cliente, para escoger el método más preciso. Con estas premisas se define la lógica para utilizar tiempo de servicio en ruta para Foxtrot.

4.10. Flujograma de Decisiones para Implementación de Tiempo de Servicio Genérico:

Los valores para definir el tamaño de muestra para utilizar cada modelo fueron definidos por el área de ingeniería de Foxtrot, en una segunda versión del algoritmo se espera que los mismos sean personalizados por Centro de Distribución considerando intervalos de confianza necesarios.

Partiendo de los resultados obtenidos se realizó la propuesta de la construcción de la versión 1 del algoritmo bajo la lógica descrita en la Figura 4.37.

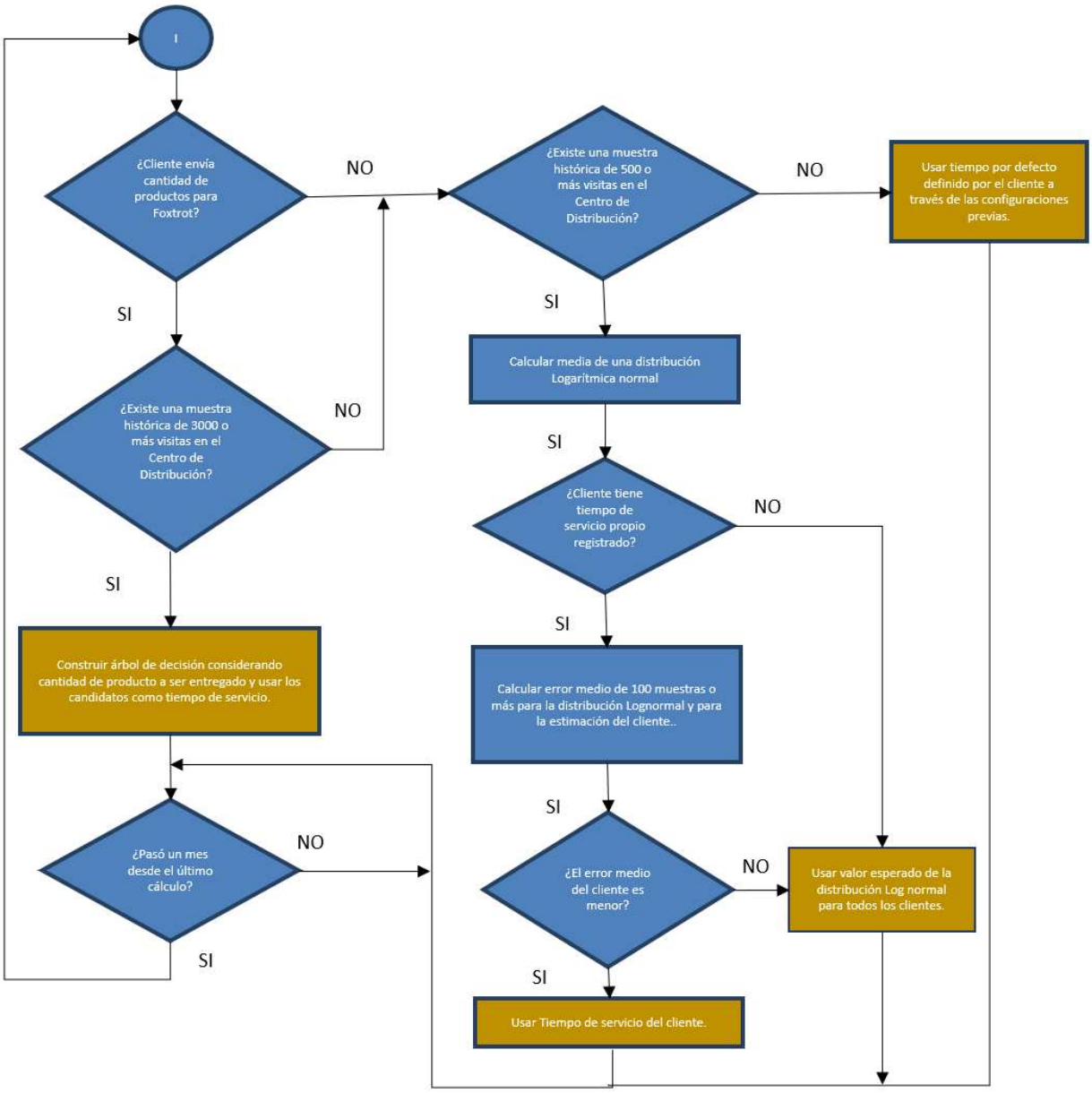

Figura 4.37 Lógica de Algoritmo de Utilización y Aprendizaje de Tiempos de Servicio. Fuente: Elaboración propia.

CONCLUSIONES

El Tiempo de Servicio para una ruta de reparto es estimable por cliente con un cierto grado de imprecisión, se trata de un proceso difícil de predecir.

Se presentaron varios modelos que fueron evaluados a partir de su precisión, donde la mejor opción para la estimación de Tiempo de Servicio fue mediante el método de partición y definición de árboles de decisión basado en cantidad de producto a ser entregada.

El Tiempo de Servicio de las operaciones de distribución urbana analizadas sigue una distribución logarítmica normal, con diferentes parámetros dependiendo el tipo de producto a ser entregado, por lo tanto, es posible establecer probabilidades de un tiempo de servicio alcanzar un valor u otro, lo cual es un gran avance para la generación de alertas. En esta distribución se definen claramente dos grupos: muchos clientes relativamente pequeños y pocos clientes relativamente grandes. Otras distribuciones como la normal, Weibull, logarítmica y exponencial no generan buenos resultados para el modelaje de la curva de densidad para el proceso de entrega de bienes y servicios.

La operación de entrega de panes representa un proceso más largo debido a que conlleva también, preventa, inventario y un relacionamiento con el cliente diferente.

La base de datos Foxtrot tiene algunas limitaciones y existen algunas variables que no se pudieron evaluar en este estudio por falta de disponibilidad de datos.

Para las empresas de bebidas se logró con la versión 2 una mejora en la estimación de tiempo servicio por cliente de 98% (-2.88 min a -0.038 min) con respecto a la media del error del valor de tiempo de servicio actualmente utilizado y de 59% (14.03 min a 5.62 min) con respecto a la desviación estándar. El caso Pittsburgh queda pendiente para aplicar la misma metodología de árbol de decisiones que dio los mejores resultados en los otros casos.

La media total por cliente es una medida que, debido a factores estacionales o a efectos látigo en la cadena de suministro, pierde precisión con el paso del tiempo. La causa exacta de este efecto no fue parte del análisis detallado del presente trabajo.

La utilización de una fuente de datos como un teléfono celular para analizar comportamientos y realizar predicciones es la base fundamental para implementar métodos de inteligencia artificial. Esta experiencia ha demostrado que el buen uso de los datos ayuda a planificar de forma más inteligente. La lógica para implementar el flujo del algoritmo de estimación de tiempo de servicio quedó bien definida y aceptada para su próxima implementación.

El Ambiente de trabajo en una “StartUp” es informal y dinámico fomentando la innovación, donde las decisiones deben tomarse rápido y con responsabilidad. Es un ambiente desafiante donde el crecimiento de la empresa depende del desempeño de cada funcionario en la más mínima tarea.

Foxtrot Systems, siendo una empresa pequeña, tiene como valor principal a sus funcionarios, quienes tienen gran potencial de innovación y resolución de problemas que permiten realizar mejoras continuas en sus productos, y adaptarse al mercado en distintos países.

La propuesta de estimación de tiempos de servicio será una característica principal y diferenciadora en comparación de otras empresas, debido a que no es un producto que exista en el mercado actualmente. Las expectativas fueron cumplidas y se espera una mejora en el funcionamiento del algoritmo de toma de decisiones en ruta debido a que se podrá contar con un valor más realista de parada estimada en cada cliente.

RECOMENDACIONES

El tiempo de servicio por cliente es un proceso difícil de caracterizar y algunas recomendaciones y estudios podrían mejorar los resultados obtenidos actualmente.

1. Exigir la mayor cantidad de datos por parte del cliente que puedan estar relacionados con el tiempo de servicio, de esta forma es posible hacer una estimación del árbol de decisiones con mayor cantidad de variables como input.

2. Realizar un estudio con Promedio Móvil o cualquier otro método para evaluar mejor el modelo basado en datos históricos por cliente, de esta forma será posible considerar efectos estacionales.

3. Conseguir integrar sistemas con todos los clientes Foxtrot para recibir al menos la cantidad de productos a ser entregados para poder construir el árbol de decisión para cada cliente.

4. Idear un proceso para juntar los tiempos de parada relacionados a un cliente y no al evento generado al presionar el botón de entrega. Mientras no se realice, exhortar el comportamiento de los choferes a realizar el "click" de visita dentro del aplicativo durante la parada de descarga, y no realizar dos paradas consecutivas en el mismo cliente debido a que los eventos de parada no se juntarán.

5. Evaluar a través de análisis de sensibilidad e intervalos de confianza para entender los valores específicos para el número de muestras necesarias para recorrer la lógica de cálculo.

6. Analizar la factibilidad de utilizar información de diferentes clientes que atienden el mismo punto de venta para tener más información y aprender más rápido.

7. Utilizar la estimación de Tiempo de Servicio para mejorar la planificación de rutas y no solo la ejecución con Foxtrot, debido que se demostró que los valores utilizados para planificar las rutas son bastante imprecisos, lo cual podría resultar jornadas subestimadas (poco aprovechamiento de los equipos de reparto) o con demasiado tiempo de entrega para entrar dentro de la jornada de trabajo.

8. Establecer un sistema de alertas y seguimiento de tiempos de parada en base a la probabilidad de la distribución logarítmica normal.

REFERENCIAS BIBLIOGRÁFICAS

[1] B. Bian, N. Zhu, S. Ling, and S. Ma, "Bus service time estimation model for a curbside bus stop," *Transp. Res. Part C Emerg. Technol.*, vol. 57, pp. 103–121, 2015.

[2] "Foxtrot | Better Driver Decisions on Route." [Online]. Available: https://foxtrotsystems.com/. [Accessed: 19-Aug-2018].

[3] M. D. Arango-Serna, C. G. Gómez-Marín, and C. A. Serna-Urán, "Modelos logísticos aplicados a la distribución urbana de mercancías," *Rev. EIA*, vol. 14, no. 28, pp. 57–76, 2017.

[4] "Modelo Determinístico y Probabilístico." [Online]. Available: http://proyectoeypii.blogspot.com/2016/10/estadisticas-probabilidades-modelos.html. [Accessed: 19-Dec-2018].

[5] "Distribución Weibull." [Online]. Available: http://distweibull.blogspot.com/. [Accessed: 19-Dec-2018].

[6] F. Ríus Díaz, *Bioestadística : métodos y aplicaciones*. Universidad de Málaga, 1997.

[7] "¿Qué es el método de estimación de máxima verosimilitud y cómo se interpreta? - Seh-lelha - Sociedad Española de Hipertensión Liga Española para la Lucha contra la Hipertensión Arterial." [Online]. Available: https://www.seh-lelha.org/que-es-el-metodo-de-estimacion-de-maxima-verosimilitud-y-como-se-interpreta/. [Accessed: 19-Aug-2018].

[8] "Bondad de ajuste - Wikipedia, la enciclopedia libre." [Online]. Available: https://es.wikipedia.org/wiki/Bondad_de_ajuste. [Accessed: 19-Dec-2018].

[9] "1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test." [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm. [Accessed: 19-Dec-2018].

[10] M. Hazewinkel, *Encyclopaedia of mathematics*. Springer-Verlag, 2002.

[11] H. A. Chipman, "Recursive Partitioning."

[12] "CHAID (Chi-square Automatic Interaction Detector) - Select Statistical Consultants." [Online]. Available: https://select-statistics.co.uk/blog/chaid-chi-square-automatic-interaction-detector/. [Accessed: 19-Dec-2018].

[13] H. Byeon, "Chi-Square Automatic Interaction Detection Modeling for Predicting Depression in Multicultural Female Students," 2017.

[14] H. Byeon and S. Cho, "The Factors of Subjective Voice Disorder Using Integrated Method of Decision Tree and Multi-Layer Perceptron Artificial Neural Network Algorithm," 2016.

[15] R. P. and M. K. Khandelwal, “JMP Statistical Discovery Software: An Overview JMP,” *Radiat. Heat Transf. (Second Ed.*, vol. 4c1, pp. 93–94, 2003.

[16] S. D. Schlotzhauer, *Elementary Statistics Using JMP*. 2007.

[17] “¿Qué es Cloud Computing? | Salesforce.” [Online]. Available: https://www.salesforce.com/mx/cloud-computing/. [Accessed: 19-Dec-2018].

[18] “Amazon Web Services: grandes posibilidades en la nube. | InGenio Learning.” [Online]. Available: https://ingenio.edu.pe/amazon-web-services-grandes-posibilidades-en-la-nube/. [Accessed: 19-Aug-2018].

[19] “Definición de GPS - Qué es, Significado y Concepto.” [Online]. Available: https://definicion.de/gps/. [Accessed: 19-Dec-2018].

APENDICE A: Distribuciones Estadísticas por Centro de Distribución

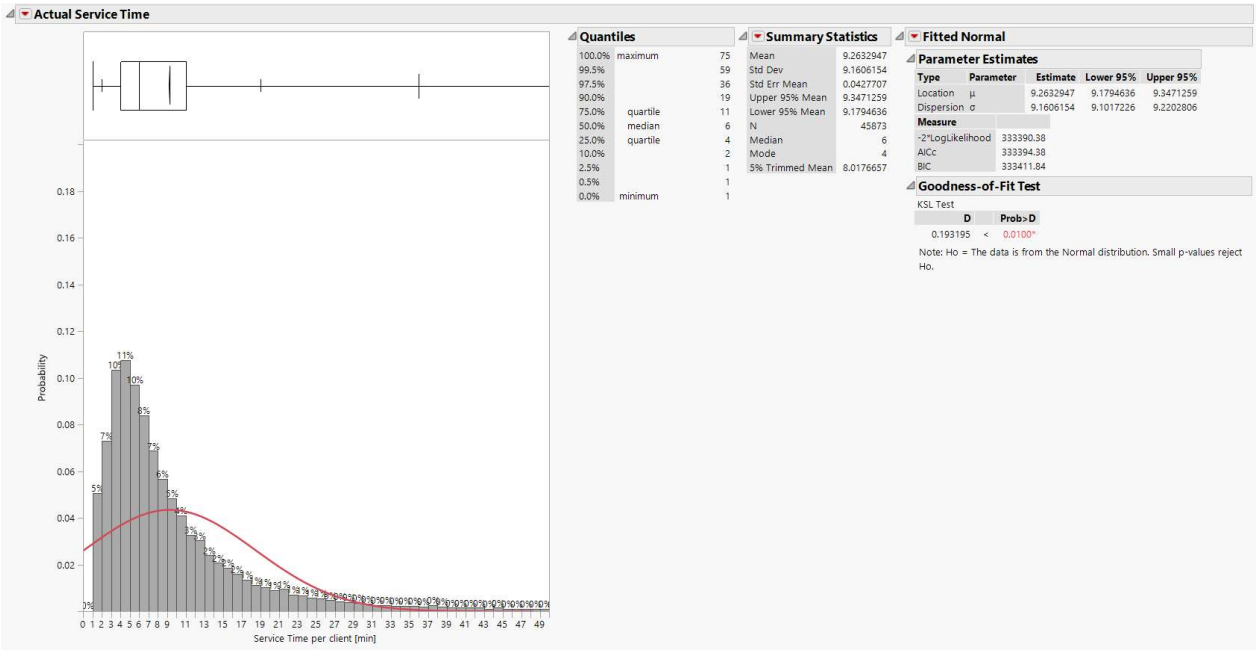

Figura A.1 Distribución normal para el Tiempo de Servicio en Guarulhos. Fuente: Elaboración propia.

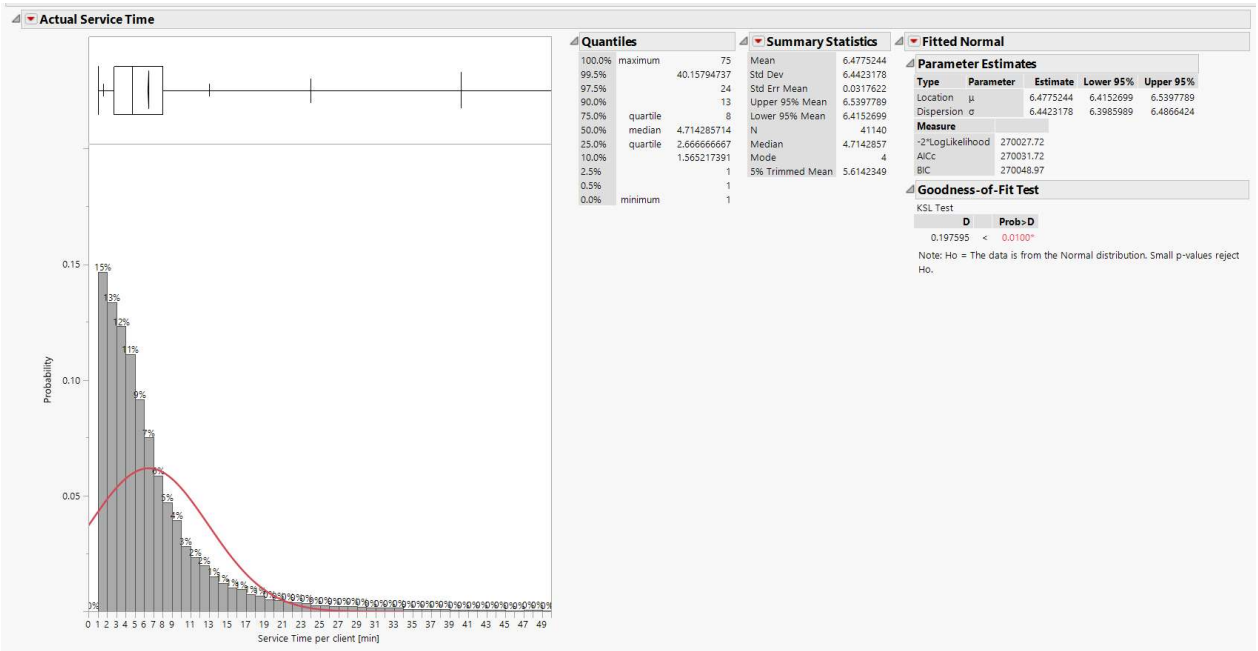

Figura A.2 Distribución normal para el Tiempo de Servicio en Rosario. Fuente: Elaboración propia.

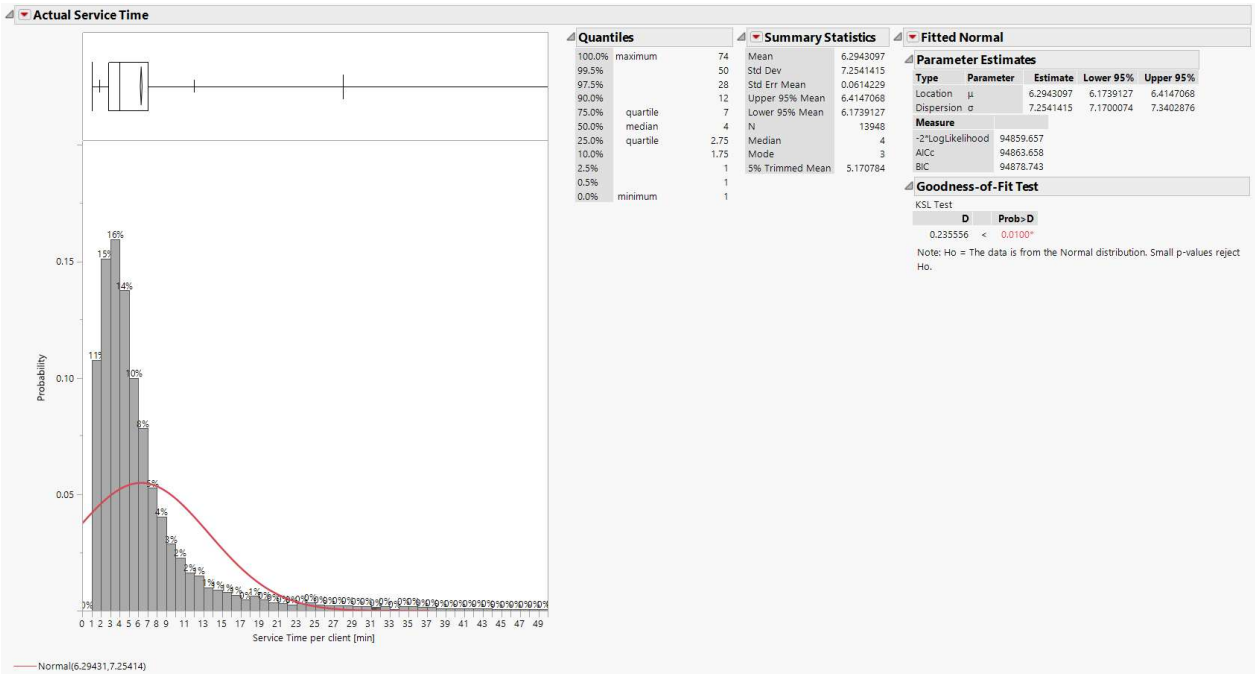

Figura A.3 Distribución normal para el Tiempo de Servicio en Guarulhos. Fuente: Elaboración propia.

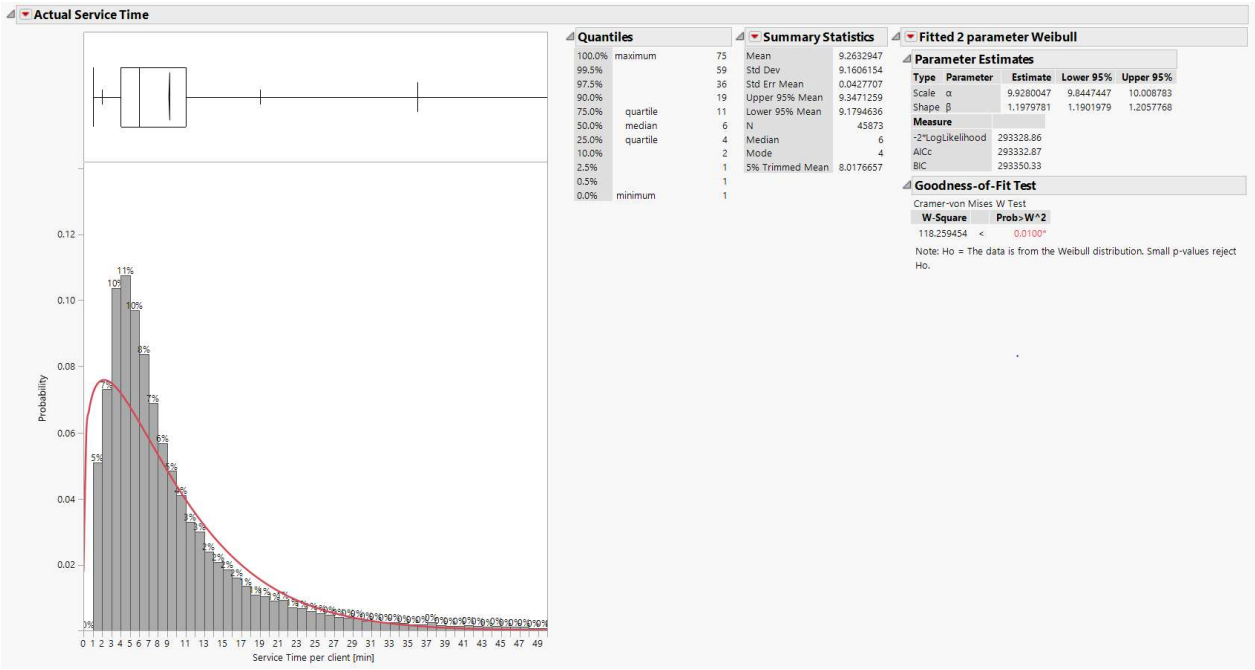

Figura A.4 Distribución Weibull para el Tiempo de Servicio en Guarulhos. Fuente: Elaboración propia.

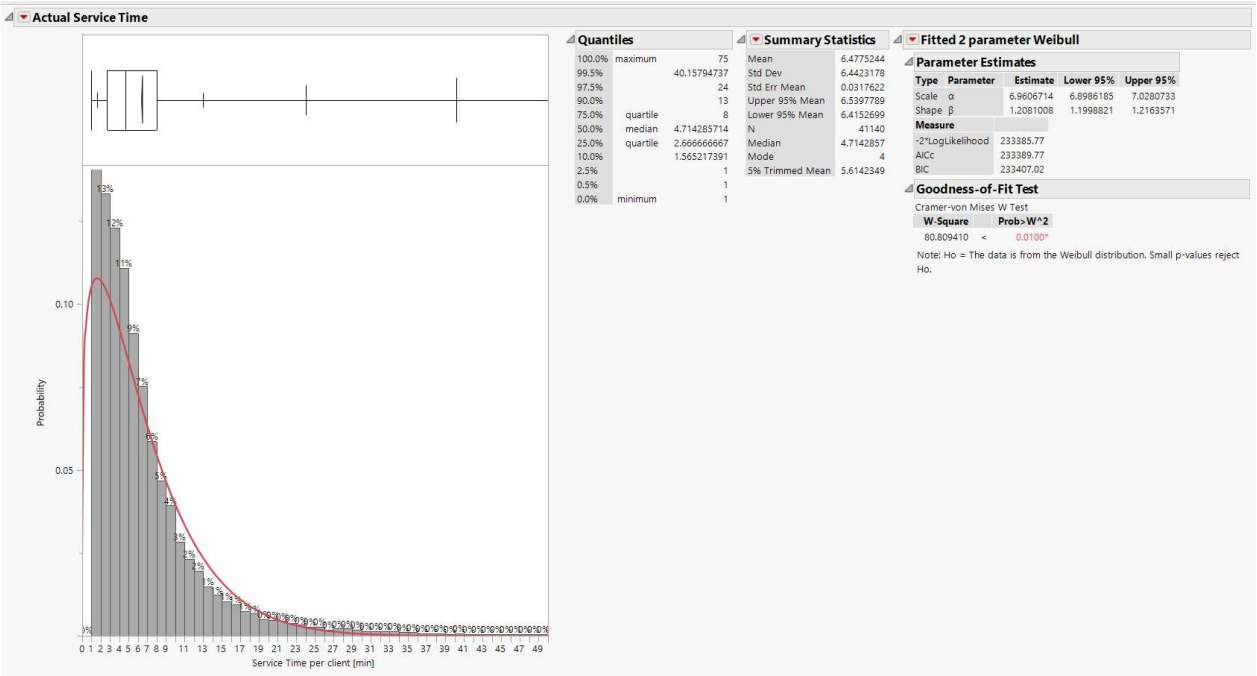

Figura A.5 Distribución Weibull para el Tiempo de Servicio en Rosario. Fuente: Elaboración propia.

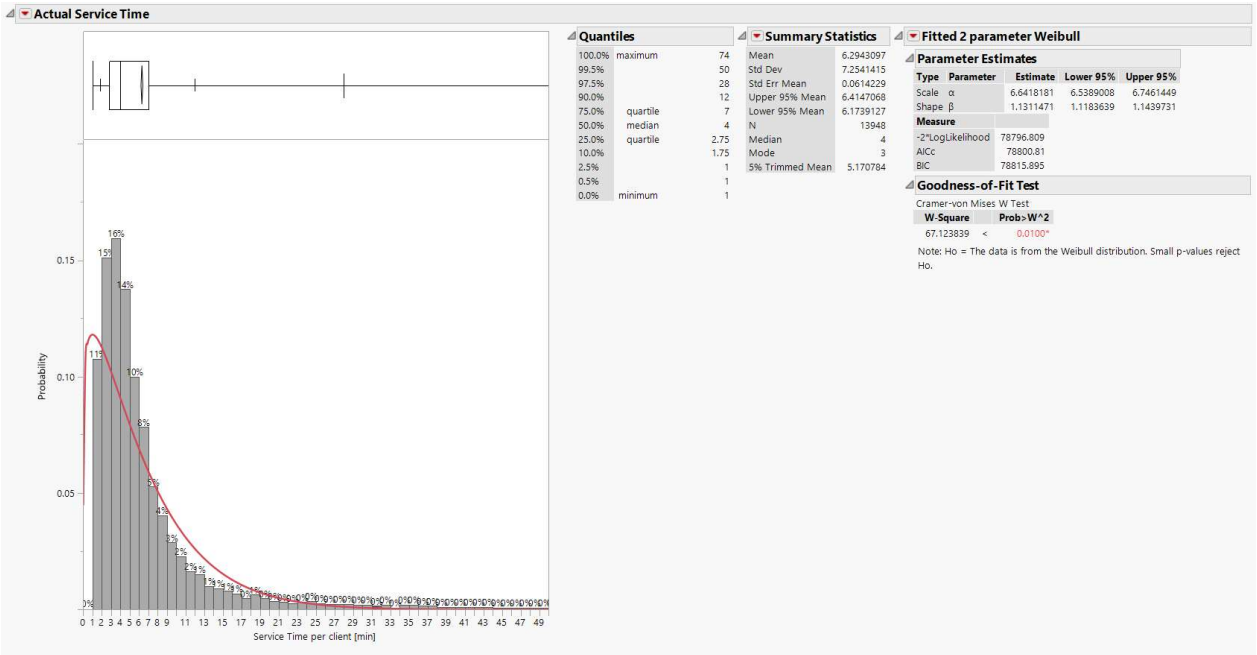

Figura A.6 Distribución Weibull para el Tiempo de Servicio en Tarija. Fuente: Elaboración propia.

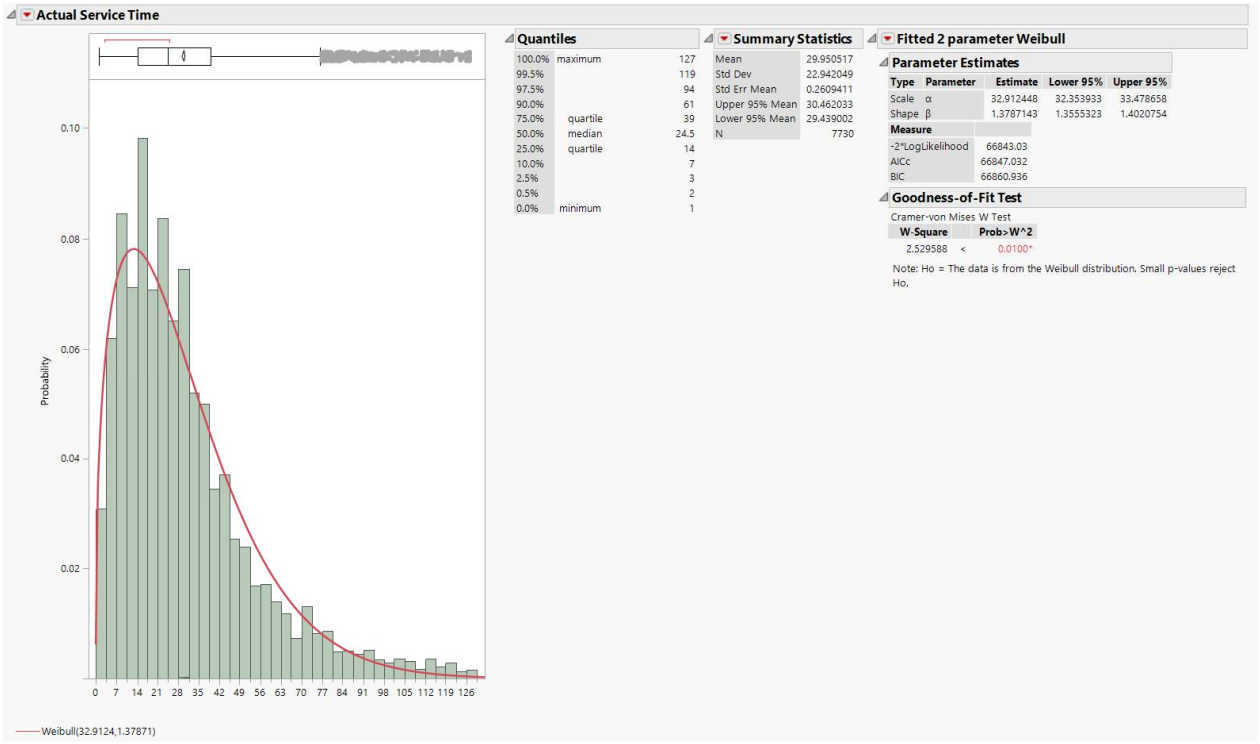

Figura A.7 Distribución Weibull para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

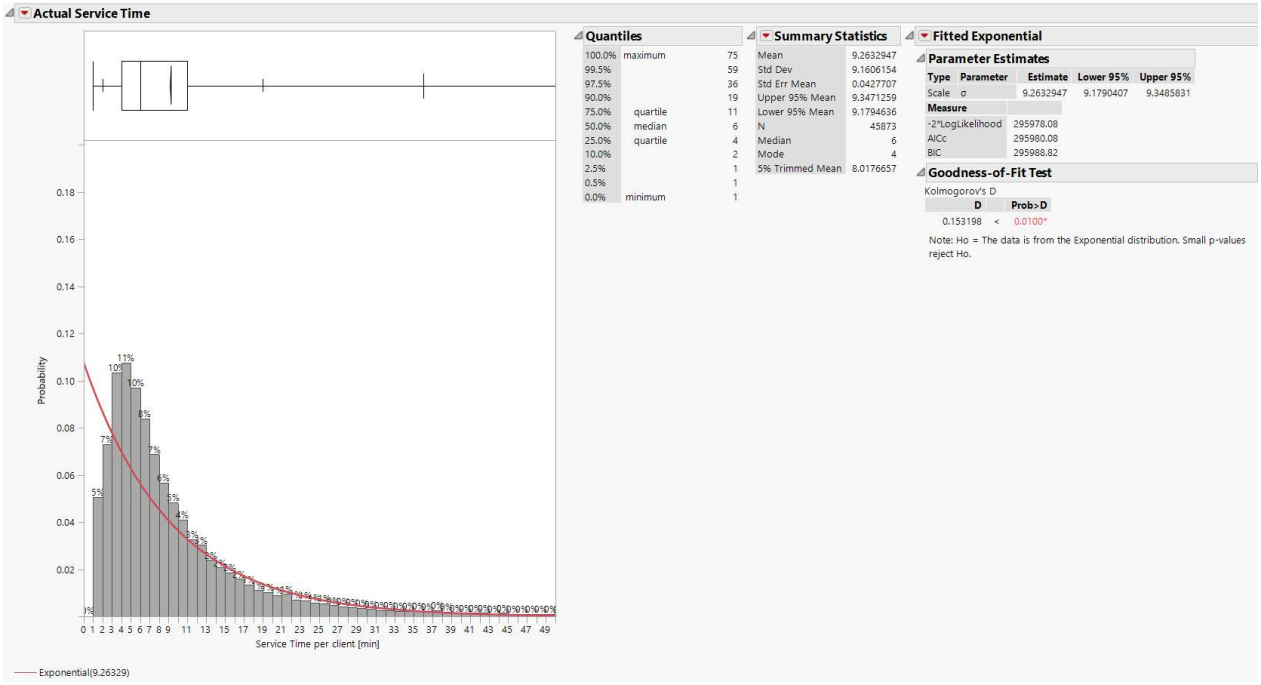

Figura A.8 Distribución Exponencial para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

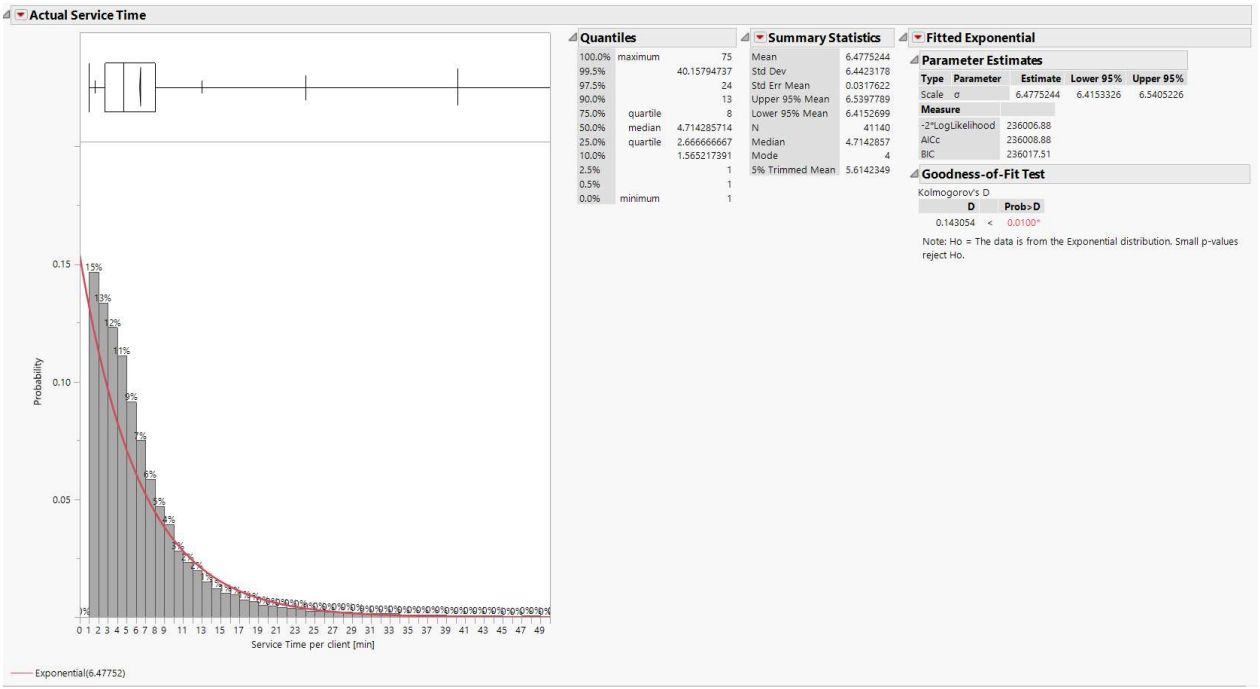

Figura A.9 Distribución Exponencial para el Tiempo de Servicio en Rosario. Fuente: Elaboración propia.

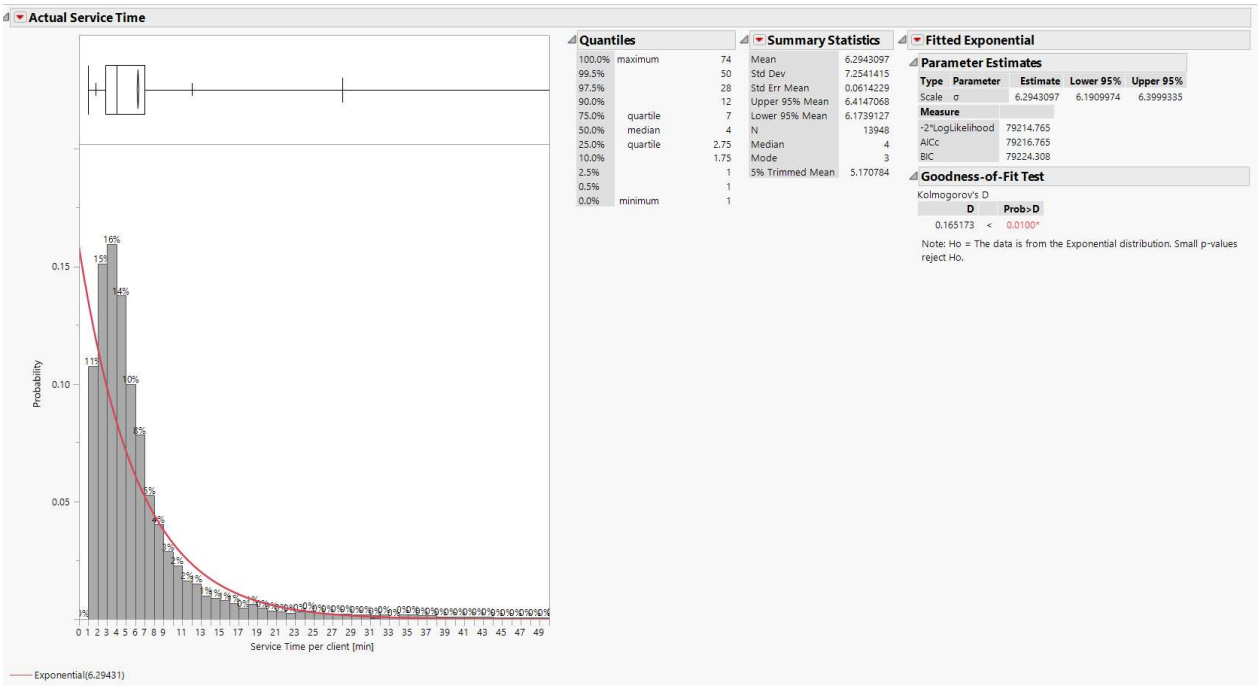

Figura A.10 Distribución Exponencial para el Tiempo de Servicio en Tarija. Fuente: Elaboración propia.

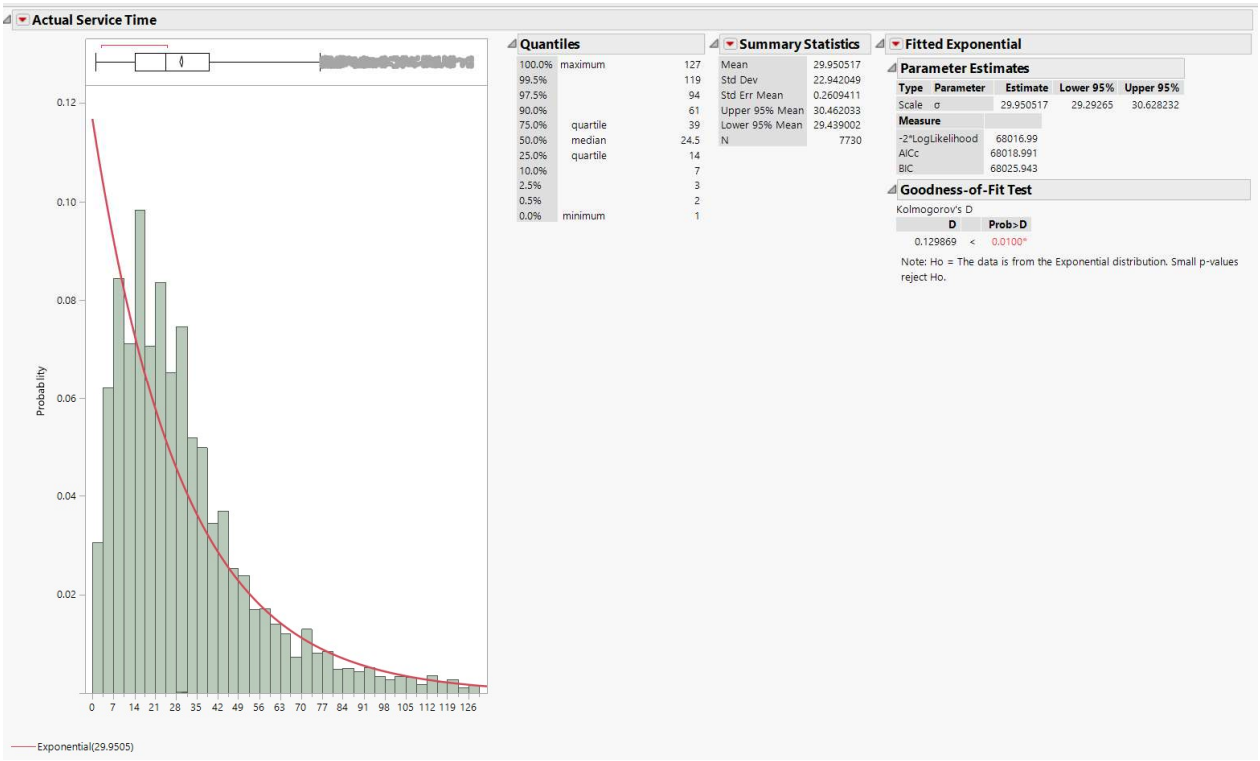

Figura A.11 Distribución Exponencial para el Tiempo de Servicio en Pittsburgh. Fuente: Elaboración propia.

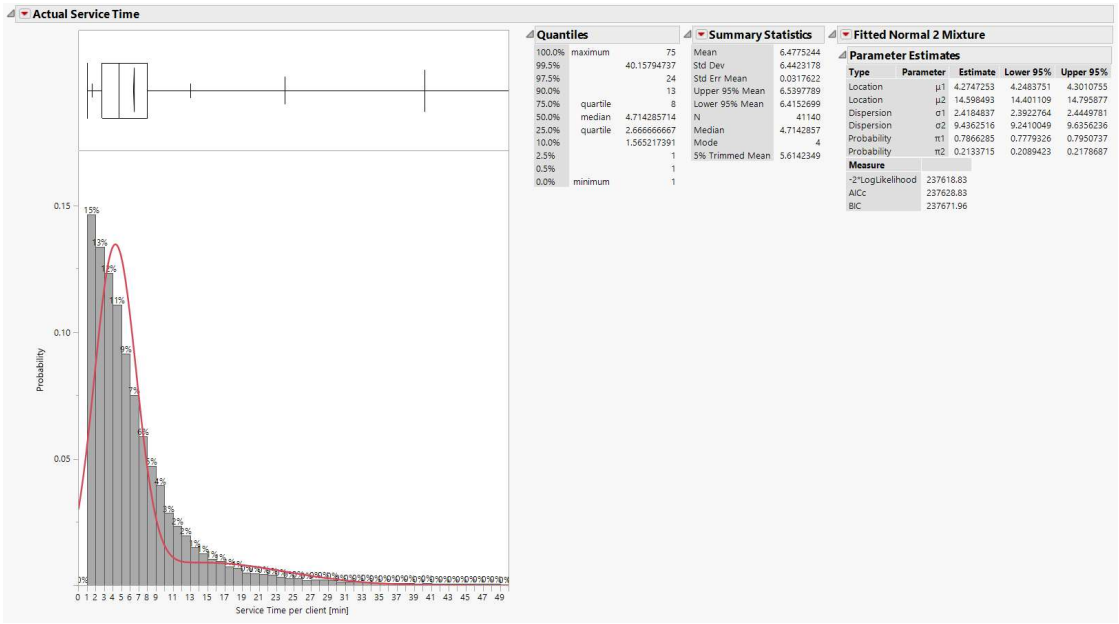

Figura A.12 Distribución Normal Doble para el Tiempo de Servicio en Rosario. Fuente: Elaboración propia.

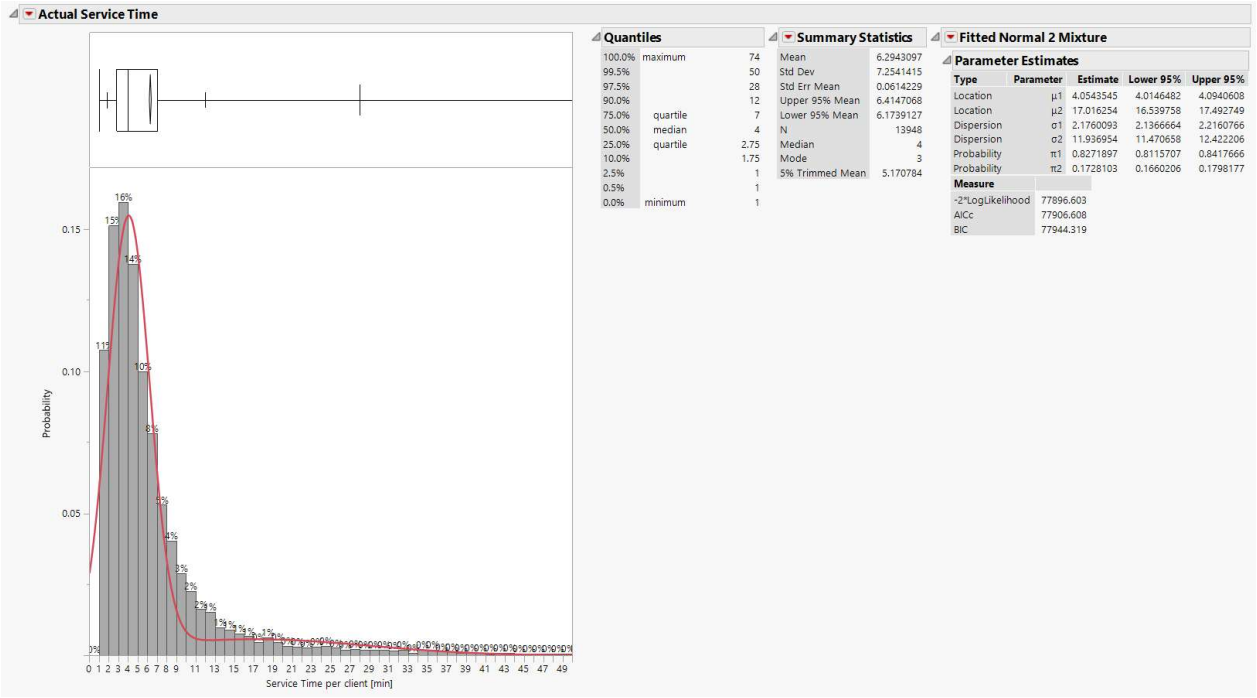

Figura A.13 Distribución Normal Doble para el Tiempo de Servicio en Tarija. Fuente: Elaboración propia.

APENDICE B: Ejemplos de distintos locales de entrega.

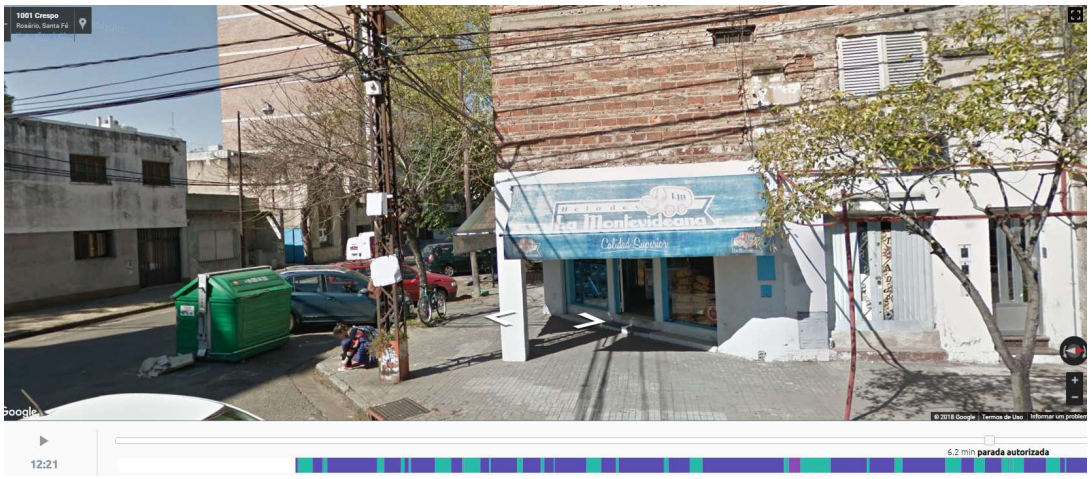

Figura B.1: Local de parada corta autorizada en Rosario. Fuente: Foxtrot Systems 2018.

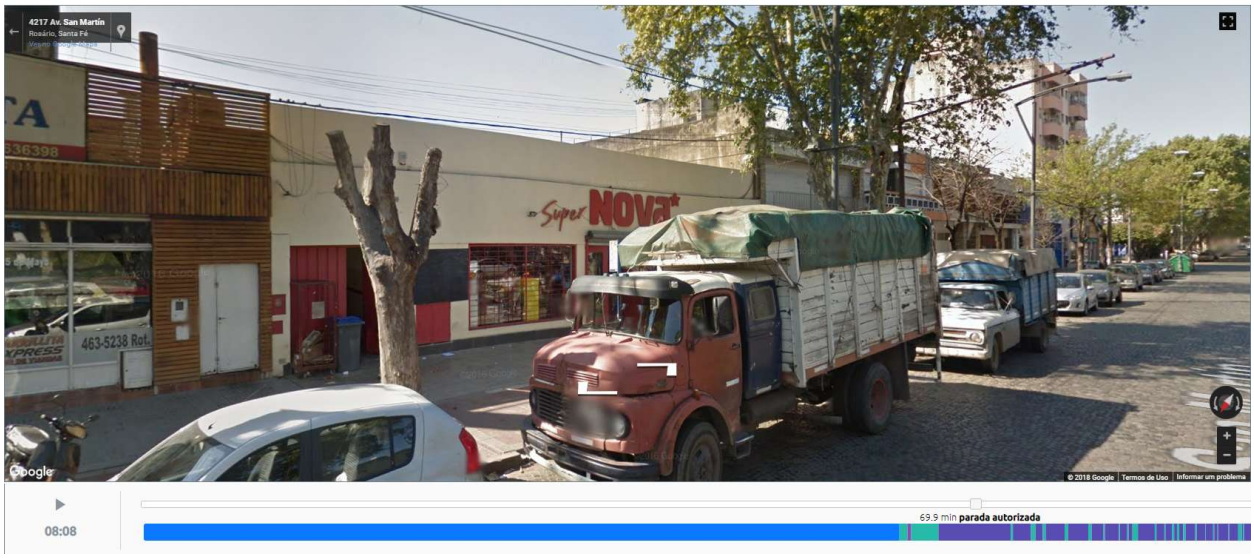

Figura B.2 Local de parada larga autorizada en Rosario. Fuente: Foxtrot Systems 2018.

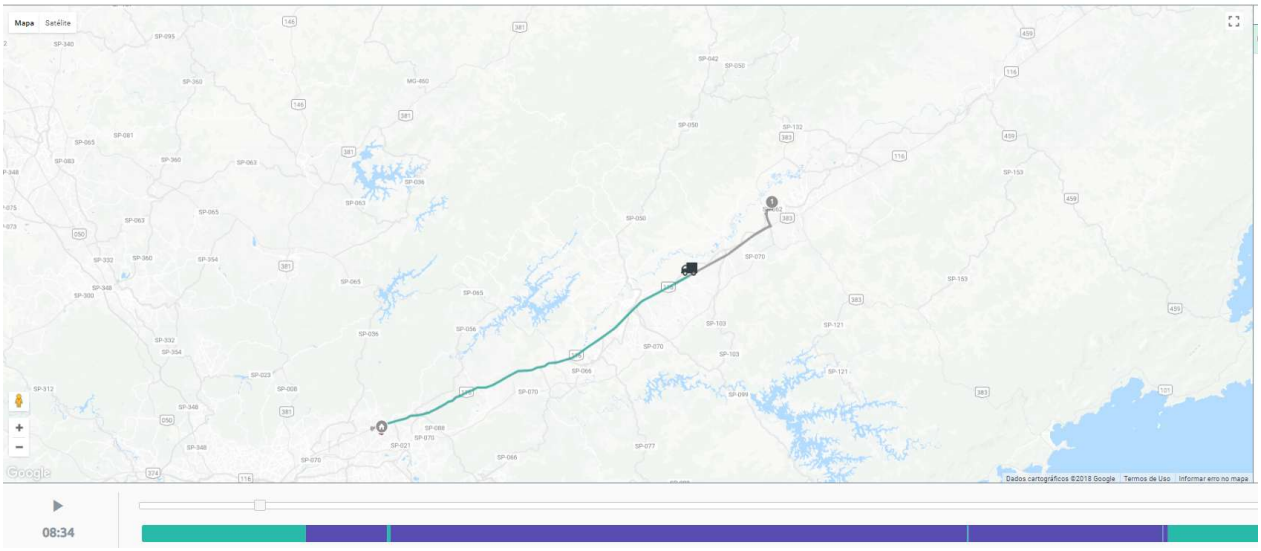

Figura B.3 Route Inspector de una ruta con un solo cliente (Walmart-Pittsburgh). Fuente: Foxtrot Systems 2018.

Figura B.4 Route Inspector en parada (Walmart-Pittsburgh). Fuente: Foxtrot Systems 2018.

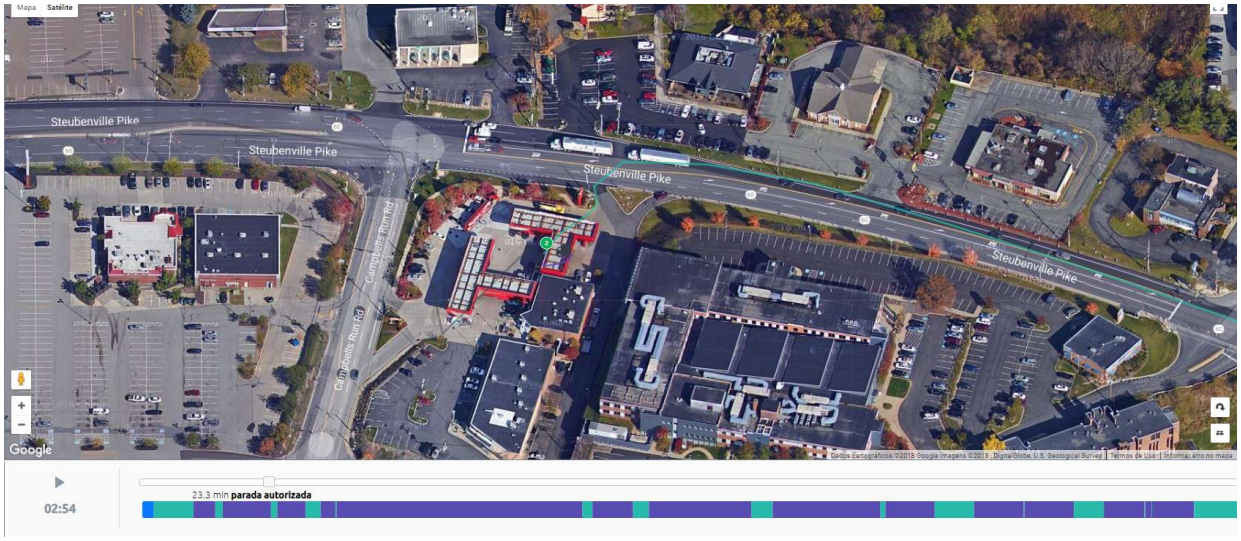

Figura B.5 Route Inspector en tienda de conveniencia en una estación de Gasolina en Pittsburgh. Fuente: Foxtrot Systems 2018.

Figura B.6 Route Inspector en parada en un mayorista en Guarulhos. Fuente: Foxtrot Systems 2018.

Figura B.7 Route Inspector en parada en un abasto en Guarulhos. Fuente: Foxtrot Systems 2018.

APENDICE C: Árboles de Decisión.

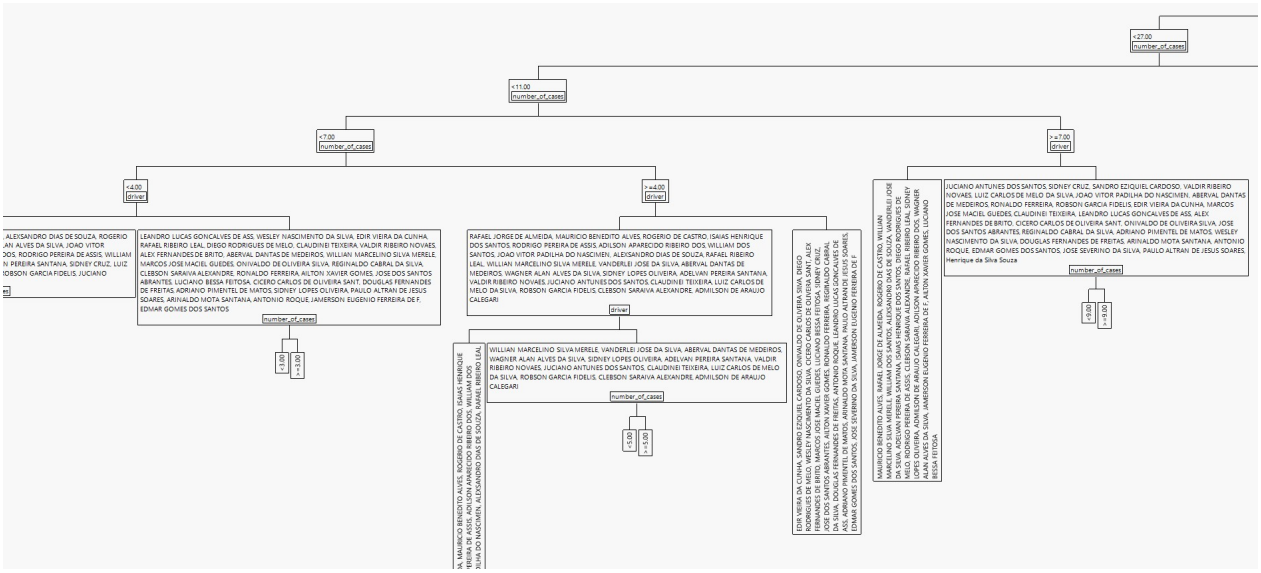

Figura C.1.1 Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte superior izquierda). Fuente: Elaboración propia.

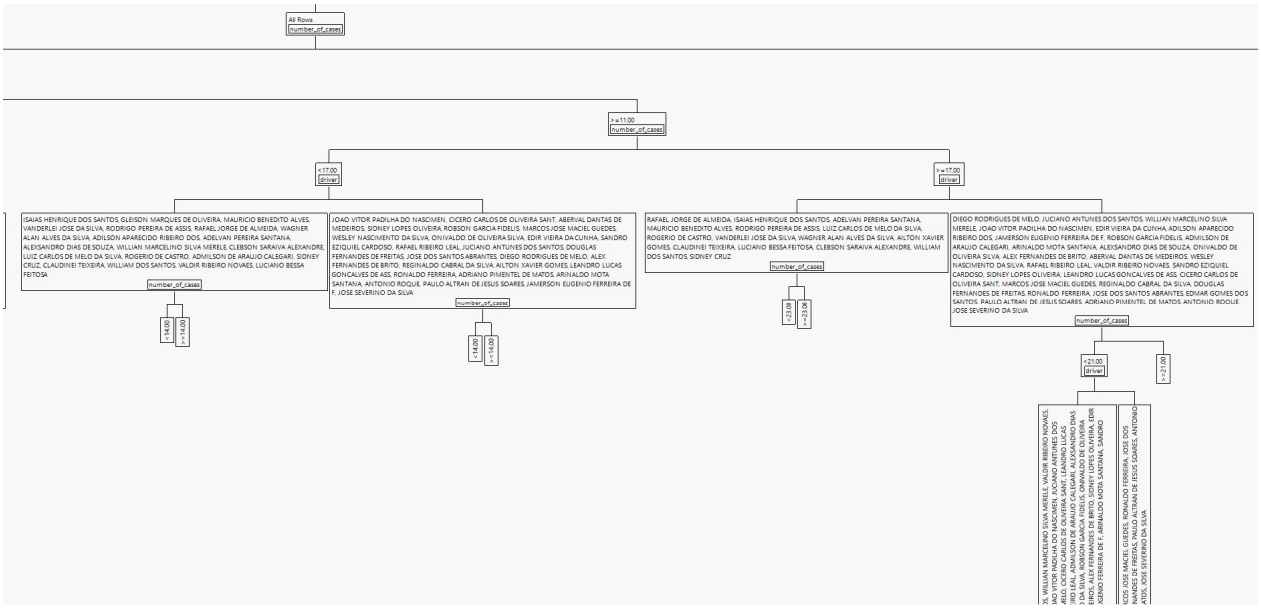

Figura C.1.2 Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte superior derecha). Fuente: Elaboración propia.

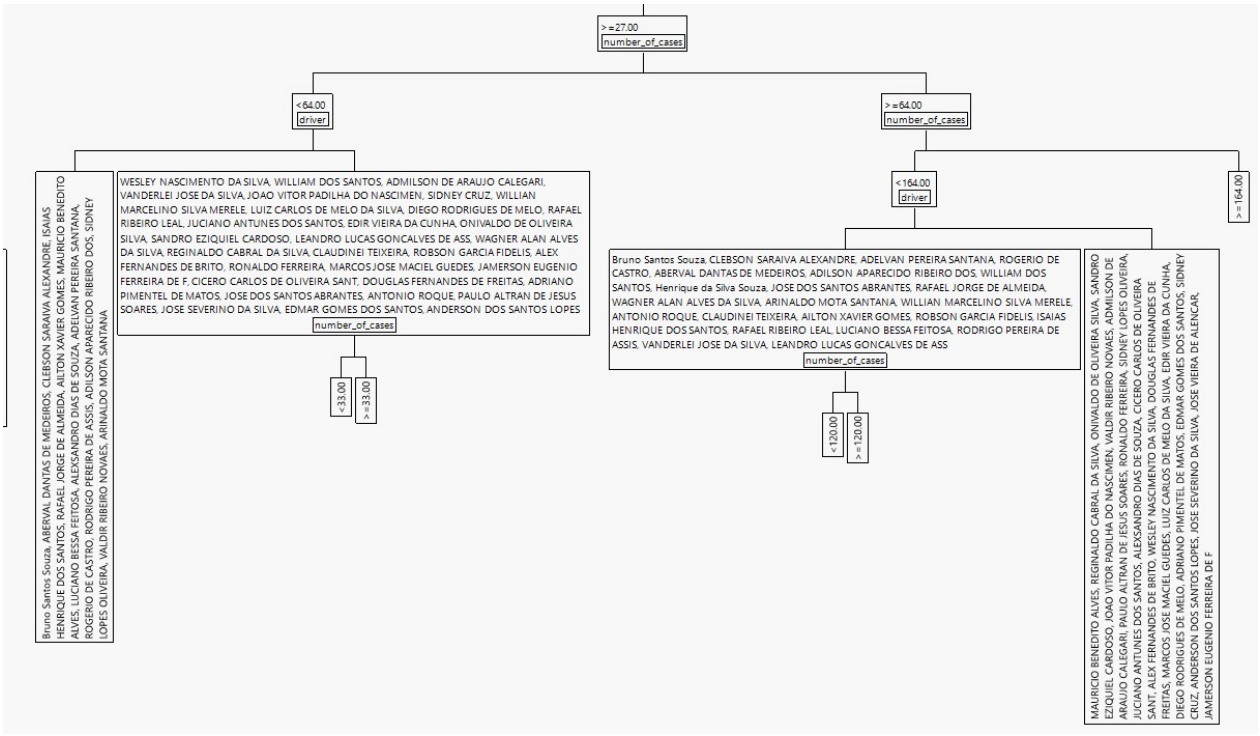

Figura C.1.3 Árbol de Decisión por cantidad de paquetes y conductor para Guarulhos (parte inferior). Fuente: Elaboración propia.

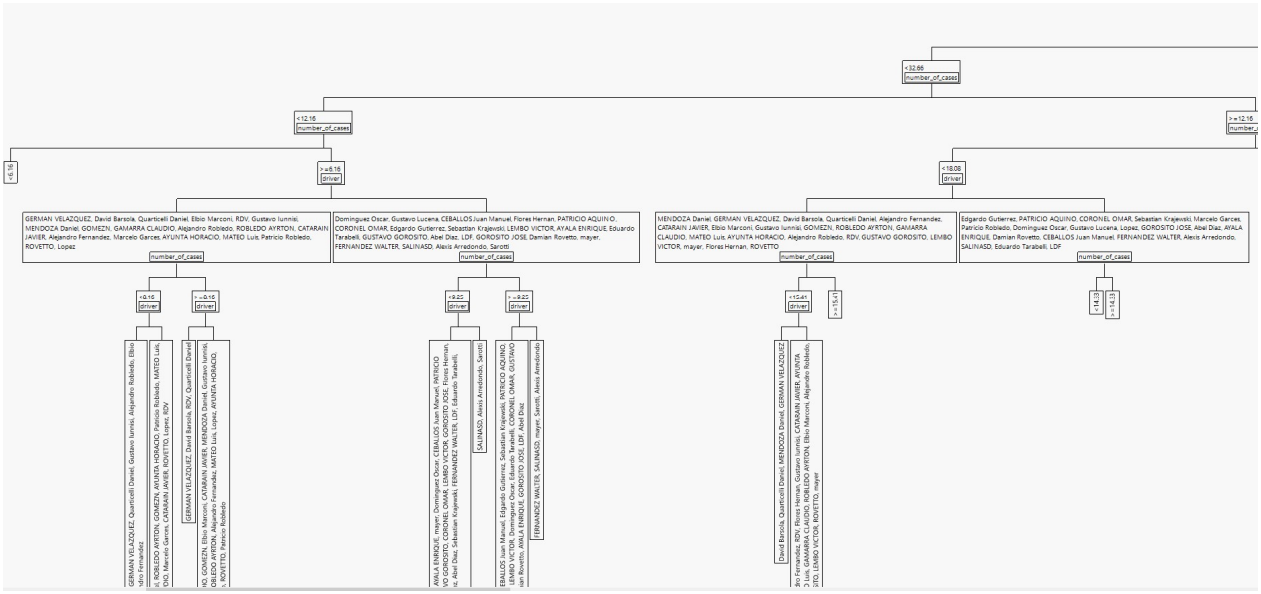

Figura C.2.1 Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte izquierda). Fuente: Elaboración propia.

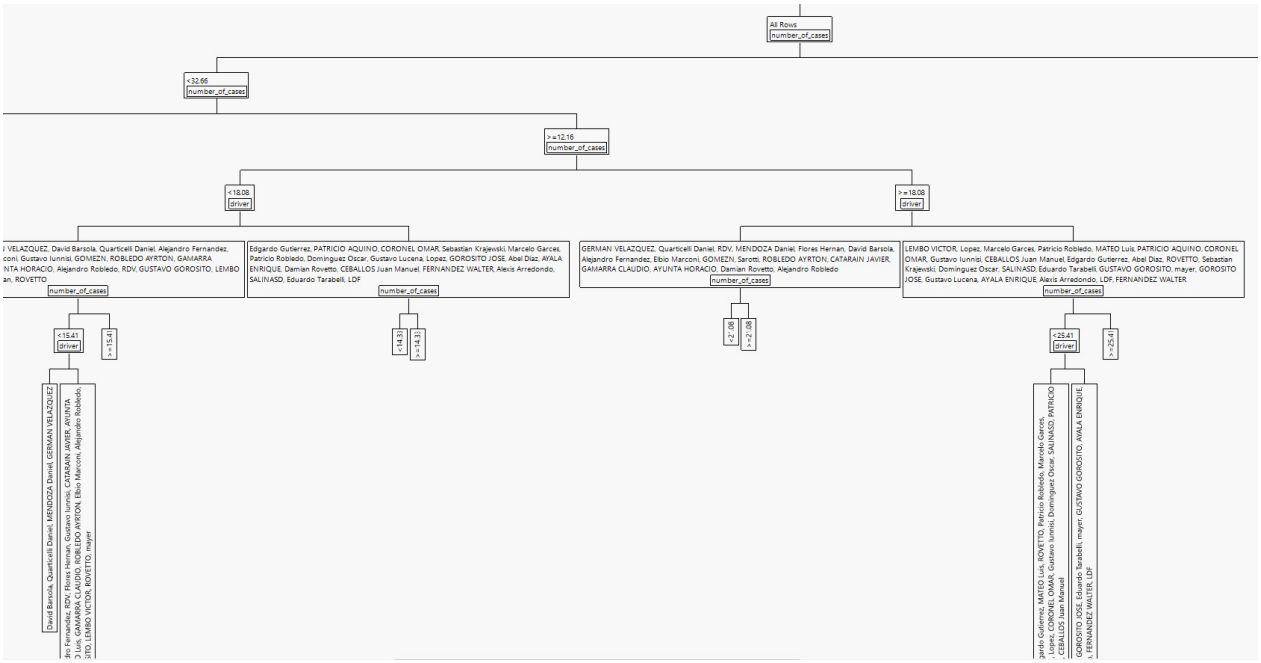

Figura C.2.2 Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte central). Fuente: Elaboración propia.

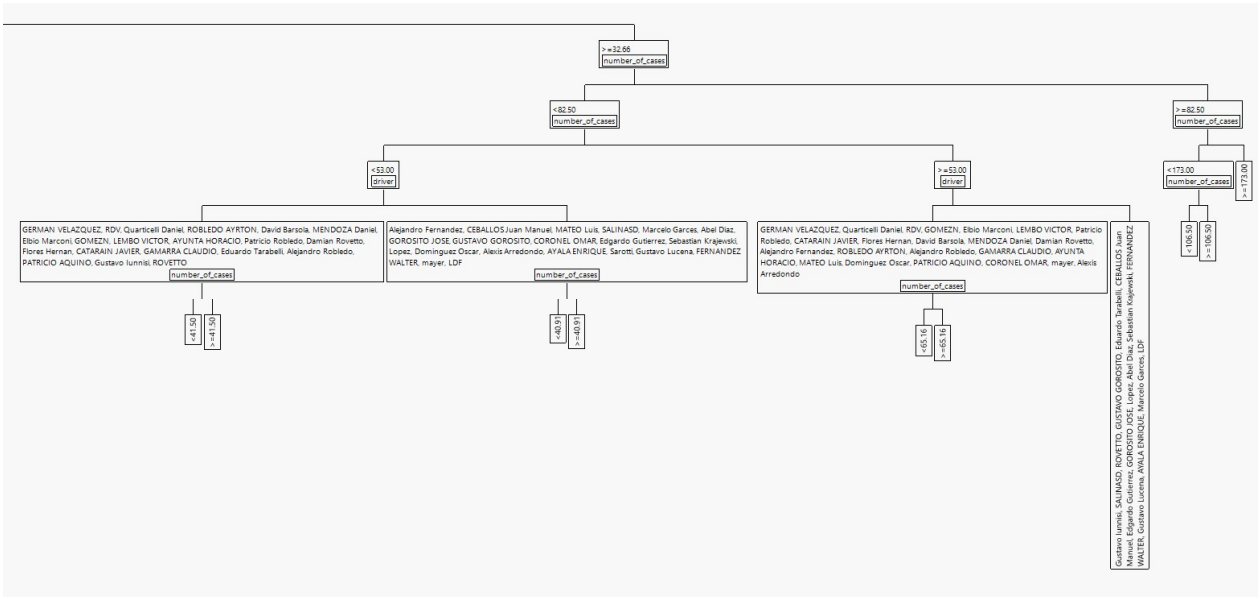

Figura C.2.3 Árbol de Decisión por cantidad de paquetes y conductor para Rosario (parte derecha). Fuente: Elaboración propia.

APENDICE D: Error en promedio vs cantidad de muestras consideradas.

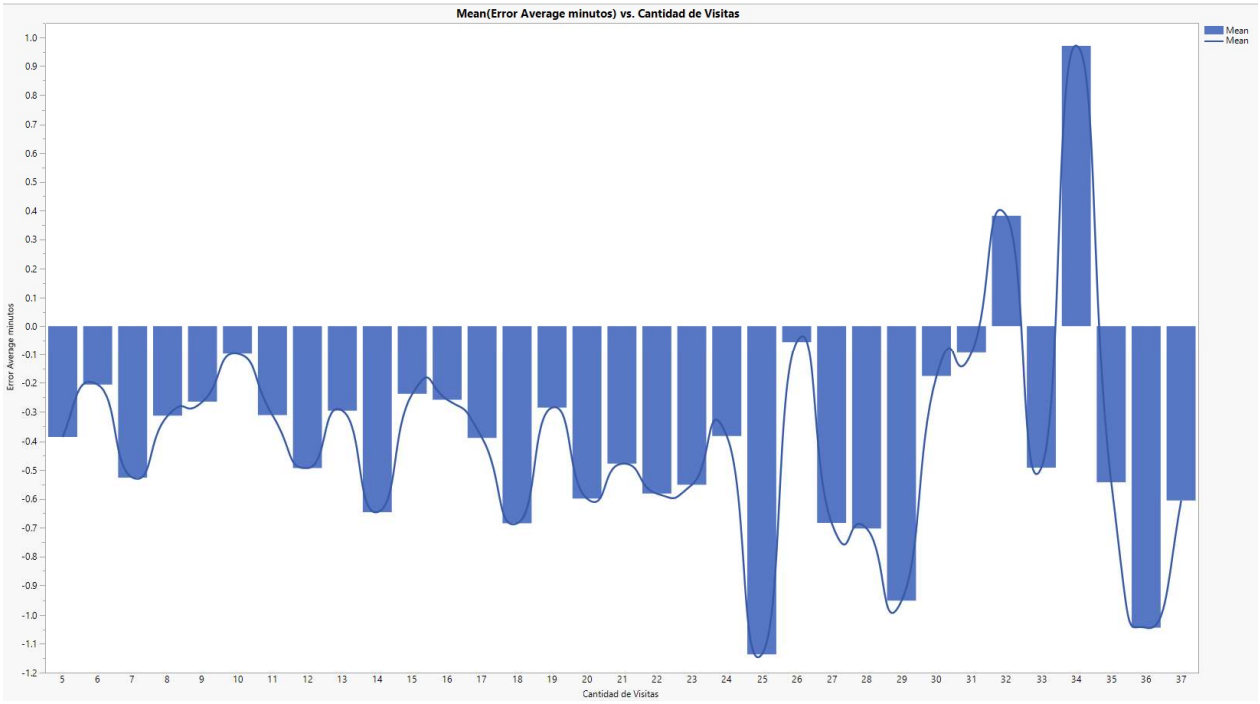

Figura D.1 Error medio vs número de visitas consideradas para el cálculo de la estimación - Rosario. Fuente: Elaboración propia.

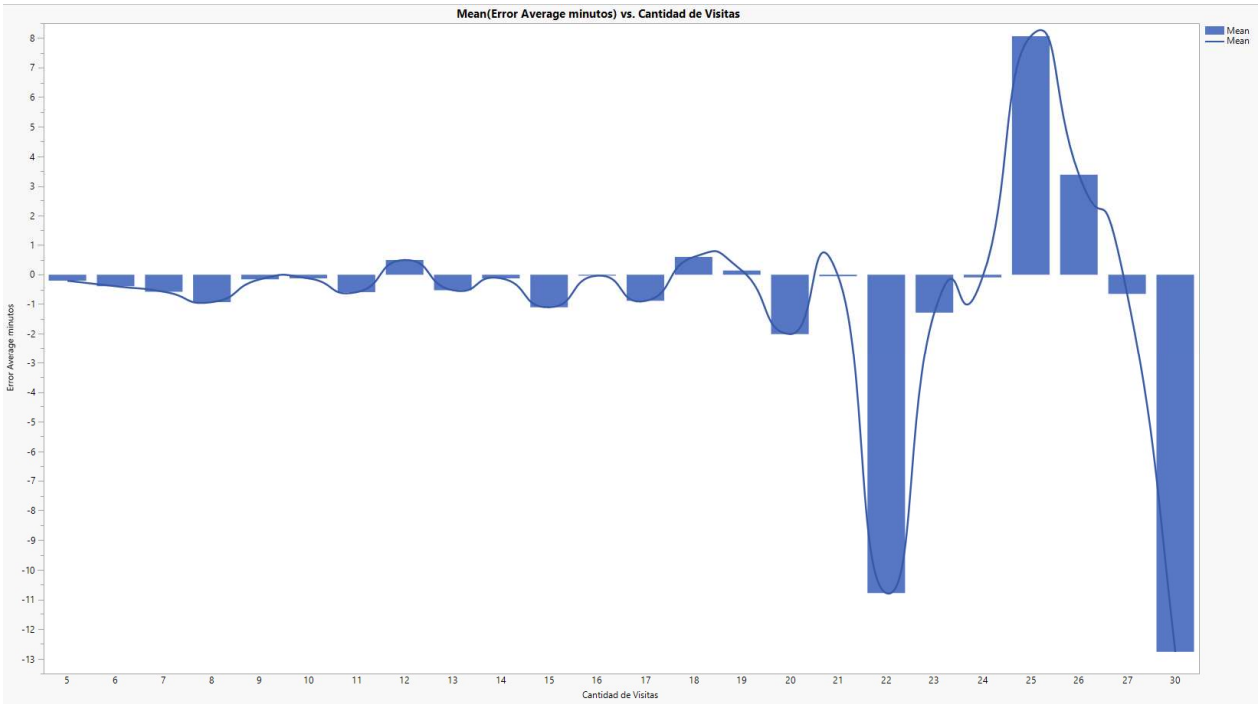

Figura D.2 Error medio vs número de visitas consideradas para el cálculo de la estimación - Tarija. Fuente: Elaboración propia.

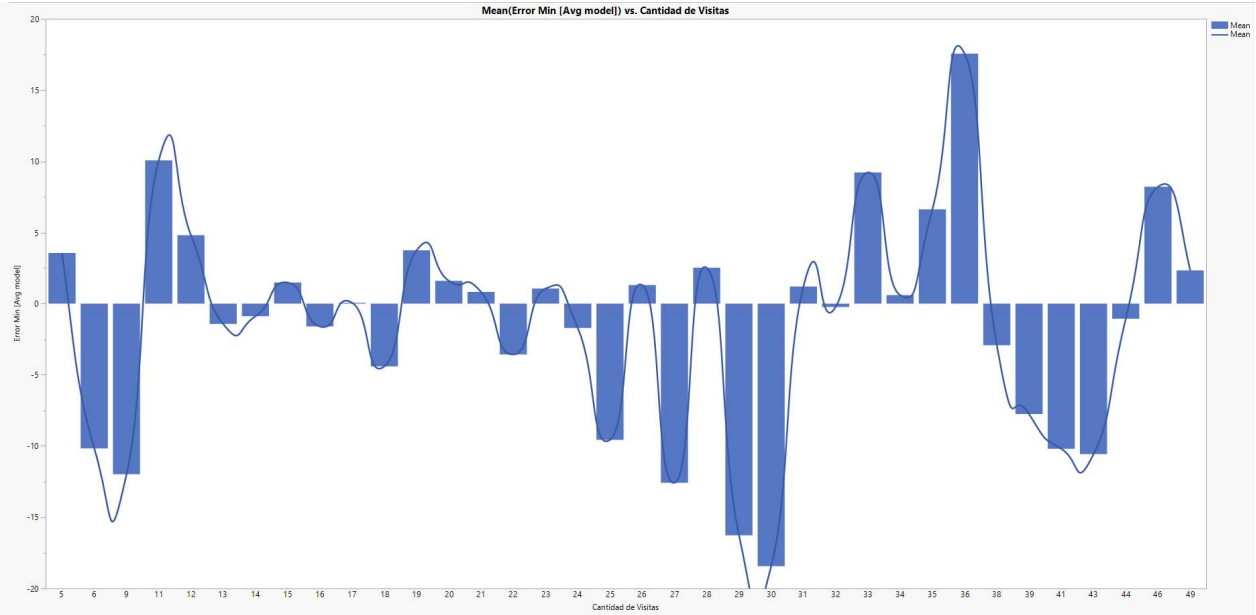

Figura D.3 Error medio vs número de visitas consideradas para el cálculo de la estimación - Pittsburgh. Fuente: Elaboración propia.