Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Cálculo intradiario de flota para la industria del delivery mediante un enfoque estocástico

Autoría: Lezcano, Gabriela
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

CÁLCULO INTRADIARIO DE FLOTA PARA LA INDUSTRIA DEL DELIVERY MEDIANTE UN ENFOQUE ESTOCÁSTICO

TESIS

Gabriela Lezcano

Mayo 2025

Tutor: Javier Marenco

Resumen

En la industria del reparto de comida a domicilio, disponer de una metodología de planificación robusta es clave para asegurar niveles adecuados de servicio y eficiencia operativa. En los últimos años, las plataformas de entrega a demanda han ganado presencia en los centros urbanos de América Latina y el mundo, integrándose como un canal habitual para acceder a bienes y servicios. Este fenómeno, impulsado por el crecimiento del comercio electrónico y los cambios en los hábitos de consumo, ha transformado los patrones logísticos tradicionales e intensificado la presión sobre los sistemas de planificación operativa. Una planificación subóptima en este contexto puede derivar rápidamente en tiempos de espera prolongados, costos operativos excesivos y una consecuente pérdida de clientes.

Frente a este escenario, esta tesis desarrolla un modelo estocástico de planificación intradiaria basado en teoría de colas y simulación discreta de eventos. A diferencia del método heurístico tradicional, que estima la dotación de flota a partir de tasas promedio, el enfoque propuesto incorpora explícitamente la incertidumbre del sistema. Utilizando simulaciones calibradas con datos históricos, se implementa un optimizador secuencial que determina, para cada intervalo de tiempo, la dotación de flota que minimiza el tiempo de espera, penalizando simultáneamente las desviaciones respecto al valor objetivo de la tasa de utilización (UTR).

El modelo fue evaluado en tres escenarios definidos por su rango de exploración y penalización al desvío del UTR. En todos los casos, la demanda fue completamente atendida por ambas configuraciones (heurística y optimizada), con niveles similares de rendimiento. Sin embargo, el modelo optimizado mostró una mejor alineación con el UTR objetivo y redujo el tiempo de espera en tramos críticos, especialmente cuando contaba con mayor flexibilidad de búsqueda. Cuando las restricciones eran más estrictas, las mejoras fueron menores y el modelo tendió a validar la configuración heurística.

Esta investigación contribuye al desarrollo de una herramienta más realista y adaptable para la planificación logística urbana, destacando el potencial de calibrar los hiperparámetros del modelo de forma dinámica según el momento del día y las condiciones de operación. Finalmente, se discuten posibles líneas futuras de mejora, incluyendo reglas adaptativas de penalización y la extensión hacia esquemas de planificación en tiempo real que permitan anticiparse con mayor eficacia a situaciones imprevistas.

Abstract

In the food delivery industry, robust planning methodologies are essential to ensure adequate service levels and operational efficiency. Over recent years, on-demand delivery platforms have significantly expanded their presence in urban centers across Latin America and worldwide, becoming integral channels for accessing goods and services. This phenomenon, driven by the rise of e-commerce and shifting consumer habits, has altered traditional logistics patterns and increased pressure on operational planning systems. Suboptimal planning under these conditions can quickly result in prolonged waiting times, excessive operating costs, and consequent customer dissatisfaction.

To address this challenge, this thesis develops a stochastic intraday planning model based on queueing theory and discrete-event simulation. Unlike traditional heuristic methods that estimate fleet staffing from average utilization rates, the proposed approach explicitly incorporates operational uncertainty. Using simulations calibrated with historical data (non-homogeneous Poisson demand and Gamma-distributed service times), the model implements a sequential optimizer that determines the optimal fleet size for each time interval. The goal is to minimize waiting times while simultaneously penalizing deviations from the targeted resource utilization (UTR).

The model was evaluated across three scenarios, each defined by its exploration range and penalty applied to UTR deviations. In all cases, demand was fully met by both heuristic and optimized configurations, resulting in similar throughput levels. Nevertheless, the optimized model achieved better alignment with the UTR target and reduced waiting times in critical periods, particularly when it had greater search flexibility. Under stricter constraints, improvements were more modest, and the model tended to validate the heuristic configuration.

This research contributes to developing a more realistic and adaptable tool for urban logistics planning, highlighting the potential of dynamically calibrating model hyperparameters according to the time of day and operational conditions. Finally, future improvement opportunities are discussed, including adaptive penalty rules and extending the model towards real-time planning schemes capable of better anticipating unforeseen circumstances.

Índice

Índice de tablas

Índice de figuras

1 Introducción

1.1 Contexto y motivación

En los últimos años, el crecimiento del comercio electrónico y los cambios en los hábitos de consumo han transformado profundamente la forma en que las personas acceden a productos y servicios. Uno de los sectores que más intensamente ha experimentado esta transformación es el de la entrega de productos a domicilio, en particular la entrega de comida a domicilio, cuyo desarrollo ha sido significativo a nivel global.

A diferencia de otras innovaciones digitales, las plataformas que surgieron ofreciendo este tipo de servicio no crearon el hábito de pedir comida a domicilio, ya que este ha estado presente en los centros urbanos desde hace décadas. Lo que sí lograron fue transformarlo, ampliarlo y estandarizarlo a nuevas escalas. El impacto de estas plataformas se hizo visible no solo en la cantidad de comercios y usuarios conectados, sino también en la redefinición de procesos logísticos, formas de empleo y dinámicas urbanas asociadas a la última milla.

A nivel global, el segmento de entrega de comida a domicilio ha sido uno de los más dinámicos dentro de la economía digital en el período reciente. Según un informe de McKinsey & Company (2021), el tamaño de esta industria se duplicó durante la pandemia, y en Europa, incluso, se quintuplicó entre 2018 y 2021. Este crecimiento fue impulsado por el avance de soluciones tecnológicas que comenzaron a integrar múltiples comercios, sistemas de pago, promociones y funcionalidades como el seguimiento en tiempo real del pedido, lo que sentó las bases para su rápida expansión.

Este fenómeno no ha sido exclusivo de las grandes economías desarrolladas. En América Latina, las plataformas de entrega a domicilio también han tenido un crecimiento exponencial, impulsado por una alta penetración de smartphones, cambios en los patrones de consumo y una creciente urbanización. En ciudades como Bogotá, Ciudad de México, Lima, Montevideo o Buenos Aires, la entrega a domicilio se consolidó como un canal habitual y competitivo, integrándose como parte central del ecosistema gastronómico local (Banco Interamericano de Desarrollo, 2022).

En el caso particular de Argentina, el crecimiento de las compañías como PedidosYa y Rappi ha sido notable, especialmente tras la pandemia de COVID-19, que actuó como catalizador de los hábitos de consumo digitales. Este proceso amplió el alcance de la entrega a domicilio

tradicional que antes se encontraba concentrado en ciertos productos y restaurantes, incorporando una oferta mucho más diversa y accesible, y redefiniendo el estándar de consumo de alimentos preparados.

A modo de ejemplo, se estima que para 2023 existían más de 40.000 repartidores activos en Argentina operando en plataformas digitales, representando un volumen significativo dentro del ecosistema laboral urbano (Banco Interamericano de Desarrollo, 2023).

Este nuevo escenario plantea desafíos significativos en términos de gestión operativa. Tanto la experiencia del usuario como la sostenibilidad del modelo de negocio dependen, en gran medida, de la capacidad de las plataformas para utilizar su flota de forma eficiente ante una demanda fluctuante, estacional e impredecible. Una planificación inadecuada puede traducirse en una sobreoferta de repartidores, con tiempos ociosos y aumento de costos, o en una subasignación que impacta negativamente en los tiempos de espera, las cancelaciones de pedidos y la satisfacción del cliente.

Ante esta nueva configuración del servicio urbano de entrega a domicilio, comprender los desafíos operativos que enfrentan las plataformas y explorar alternativas más sólidas para la planificación de recursos no solo es relevante desde una perspectiva operativa, sino también fundamental para avanzar hacia un modelo más eficiente, sostenible y alineado con las dinámicas reales del sistema.

1.2 El problema

1.2.1 Los desafíos de la planificación operativa en la entrega a domicilio

La planificación operativa de sistemas de entrega de comida a domicilio presenta una serie de desafíos particulares, que derivan de la combinación de tres elementos clave: una demanda altamente volátil, tiempos de entrega muy limitados y un entorno operativo con condiciones cambiantes y difícilmente controlables.

En primer lugar, el marco temporal disponible para cumplir con un pedido es sumamente acotado. Los clientes suelen esperar recibir su pedido en un lapso total de 30 a 45 minutos, contemplando un tiempo promedio de preparación en cocina de entre 10 y 15 minutos. Esto deja solo entre 20 y 30 minutos efectivos para concretar la entrega. Este margen tan estrecho convierte cualquier desvío o imprevisto en un factor potencial de incumplimiento.

A la presión temporal inherente al modelo de entrega a domicilio bajo demanda se suma una demanda altamente concentrada en franjas horarias específicas, particularmente durante los períodos tradicionales de almuerzo y cena. Esta concentración genera picos operativos que

desafían la capacidad de respuesta del sistema. Además, la operación se ve afectada por una serie de factores externos no controlables, como el estado del tránsito, las condiciones meteorológicas, eventos urbanos imprevistos o campañas promocionales impulsadas por la propia plataforma, todos con potencial para alterar significativamente el desempeño del sistema.

Dentro de esta cadena operativa, un eslabón especialmente crítico lo constituyen los socios comerciales, en particular los restaurantes en el caso del servicio de reparto a domicilio de comida. La velocidad y regularidad con la que preparan los pedidos resulta clave para sostener un flujo eficiente de entregas. Retrasos en cocina, variabilidad en los tiempos de preparación o descoordinaciones con los repartidores pueden convertirse rápidamente en cuellos de botella que comprometen la estabilidad operativa y la experiencia del usuario final.

En este contexto, planificar adecuadamente la cantidad de repartidores activos en cada momento no solo requiere anticipar cuántos pedidos se recibirán, sino también incorporar una comprensión más profunda del funcionamiento real del sistema. Diseñar esquemas de turnos cortos (habitualmente en bloques de 15 o 30 minutos) exige un alto nivel de granularidad en la toma de decisiones, que contemple múltiples factores operativos relevantes, tales como la posibilidad de asignación de múltiples pedidos a un mismo repartidor, la geografía y densidad de cada zona, las distancias a recorrer y la disponibilidad efectiva de repartidores en cada intervalo.

Aun cuando la planificación previa sea adecuada, el rendimiento operativo del sistema está fuertemente influido por el comportamiento de los repartidores durante la ejecución. Factores como el conocimiento del territorio, la experiencia con puntos de entrega específicos, las decisiones para estacionar o caminar, y la capacidad de adaptación ante imprevistos introducen una variabilidad difícil de modelar. Esta complejidad se ve acentuada en el contexto de la *gig economy*, donde muchos repartidores no tienen horarios ni rutas fijas, lo que reduce su familiaridad con ciertas zonas e incrementa la dispersión en su desempeño operativo.

En definitiva, la entrega de comida a domicilio implica una cadena operativa atravesada por múltiples fuentes de incertidumbre: desde la llegada de pedidos y el comportamiento de los restaurantes, hasta las condiciones externas y la ejecución a cargo de los repartidores. En este escenario, los métodos deterministas o basados en promedios, si bien útiles como aproximación inicial, resultan insuficientes para capturar la complejidad del sistema. Por ello, cobran relevancia enfoques alternativos que incorporan explícitamente la estocasticidad de los procesos, ofreciendo herramientas más robustas para una planificación operativa eficiente.

1.2.2 Revisión conceptual y antecedentes

Dado el contexto mencionado en la problemática operativa del servicio de reparto a domicilio, han emergido diferentes enfoques analíticos orientados a mejorar la planificación y asignación de recursos en sistemas logísticos dinámicos, especialmente en operaciones de última milla y servicios urbanos bajo demanda.

Tradicionalmente, la planificación operativa en logística urbana se ha basado en enfoques deterministas y heurísticos, que utilizan datos históricos promedio, tasas de utilización objetivo y proyecciones simples de demanda. Métodos como la planificación por ventanas de tiempo y los algoritmos clásicos de ruteo (Vehicle Routing Problem - VRP) han sido ampliamente utilizados en este tipo de contextos. Sin embargo, su principal limitación radica en la incapacidad de capturar la alta variabilidad e incertidumbre de sistemas con demanda fluctuante y condiciones operativas cambiantes (Agatz et al., 2013).

Para abordar estas limitaciones, se ha propuesto el uso de herramientas basadas en simulación y teoría de colas, capaces de representar de forma explícita la interacción entre la aleatoriedad en la llegada de órdenes y la capacidad limitada del sistema. Un ejemplo destacado en esta línea es el trabajo de Crainic, Ricciardi y Storchi (2009), quienes desarrollan modelos para evaluar y planificar sistemas logísticos urbanos mediante programación estocástica y simulación discreta, con énfasis en redes de distribución de dos niveles.

Complementariamente, Zhang y Pavone (2014) proponen un enfoque de redes de colas cerradas para modelar y controlar sistemas de movilidad bajo demanda, capturando el flujo dinámico de recursos y la necesidad de reequilibrio en contextos urbanos. Su trabajo demuestra que el diseño operativo (incluyendo decisiones sobre flota y ubicación) puede beneficiarse significativamente del análisis estocástico.

Desde una mirada más transversal, Buldeo Rai et al. (2022) ofrecen una revisión exhaustiva del fenómeno del crowdsourced entrega a domicilio, analizando tanto las plataformas existentes como la literatura académica que aborda sus desafíos logísticos. Este trabajo enmarca el servicio de entrega a domicilio como un sistema complejo donde confluyen decisiones tecnológicas, operativas y organizacionales, subrayando la necesidad de enfoques flexibles para abordar la variabilidad de la operación.

Por otro lado, en los últimos años, la investigación operativa ha comenzado a ganar mayor visibilidad fuera del ámbito estrictamente académico, posicionándose como una herramienta clave para entender fenómenos complejos en plataformas digitales. En este marco, diversos medios especializados han abordado los desafíos logísticos del servicio de reparto de comida a

domicilio desde una perspectiva cuantitativa y sistémica. Un ejemplo relevante es el artículo de Srihita et al. (2019), publicado en FactorDaily, que ofrece un análisis accesible y bien fundamentado sobre la economía del reparto a domicilio bajo demanda. El texto examina cómo decisiones sobre tamaño de flota, tiempos de espera y estructuras tarifarias impactan simultáneamente en la experiencia del cliente, la eficiencia operativa y la rentabilidad del sistema. Si bien no se trata de una publicación académica, este artículo resultó particularmente útil durante las etapas iniciales del trabajo, sirviendo como disparador conceptual para el planteo del modelo y la exploración de sus principales variables.

En síntesis, los antecedentes revisados coinciden en señalar que los enfoques tradicionales de planificación son insuficientes frente a la complejidad y volatilidad del servicio entrega a domicilio. La incorporación de modelos estocásticos, y en particular la simulación basada en teoría de colas, ofrece una vía más robusta y adaptable para dimensionar recursos, evaluar configuraciones y mejorar la calidad del servicio en sistemas operativos dinámicos.

1.2.3 Definición del problema

En línea con la problemática expuesta y los antecedentes conceptuales mencionados, esta tesis aborda específicamente el problema de la planificación operativa intradiaria de la flota de repartidores en una plataforma urbana de entrega a domicilio. La investigación se centra en una operación real que organiza su servicio en ciudades divididas en zonas logísticas, estructuradas a partir de áreas geográficas predefinidas. Estas zonas operativas se gestionan mediante la asignación de turnos específicos a distintos puntos de inicio distribuidos estratégicamente en cada zona.

La operación diaria se planifica en intervalos muy cortos denominados *buckets*, con una duración estándar de 15 minutos cada uno. Para cada uno de estos buckets, comprendidos dentro de un horario operativo amplio que abarca desde las 07:00 hasta las 02:00 (un total de 76 buckets diarios), la plataforma estima la cantidad necesaria de repartidores utilizando un enfoque determinista y heurístico, cuya descripción detallada será abordada en la próxima sección.

El método actual empleado para dimensionar la flota requerida a lo largo de cada jornada se basa principalmente en dos variables clave: la proyección de órdenes (calculada en intervalos de 30 minutos mediante técnicas de series temporales y análisis histórico) y la tasa de utilización (UTR), definida como el número promedio de órdenes que un repartidor puede completar en una hora, incorporando implícitamente el efecto del apilado de órdenes (capacidad de un repartidor para recoger y entregar múltiples pedidos en un mismo recorrido). La fórmula básica empleada para estimar la demanda de repartidores es la siguiente:

$$Repartidores\ necesarios = \frac{Proyección\ de\ Órdenes}{Tasa\ de\ Utilización\ * 2} \quad (1)$$

El factor 2 surge para ajustar la diferencia en unidades temporales utilizadas (proyección cada 30 minutos vs la tasa de utilización expresada en horas). Además, se realiza una interpolación adicional para obtener estimaciones en intervalos de 15 minutos, utilizando valores adyacentes para suavizar la curva de demanda. Finalmente, el modelo requiere ajustes manuales constantes realizados por los equipos operativos locales, para compensar fluctuaciones inesperadas en la disponibilidad efectiva de repartidores.

Aunque el método heurístico actual permite una implementación sencilla y práctica, los equipos operativos locales enfrentan cotidianamente diversas dificultades debido a las limitaciones inherentes a este enfoque:

- Falta de captura de la variabilidad real en los tiempos de servicio:
  La Tasa de Utilización (UTR) utilizada como referencia es un valor promedio que presupone estabilidad en los tiempos de servicio. En la operación real, factores como variaciones en la preparación de órdenes, a que niveles se realiza la asignación de múltiples órdenes a un mismo repartidor, o cambios en las condiciones del tráfico generan desviaciones importantes que no quedan reflejadas adecuadamente, ocasionando frecuentes subestimaciones en los momentos de mayor variabilidad.
- Incapacidad para modelar adecuadamente la aleatoriedad en la llegada de pedidos:
  El cálculo de flota se basa en datos históricos agregados por intervalos, lo que limita significativamente su capacidad para anticipar situaciones repentinas de aumento en la demanda o cambios operativos inesperados.
- Asignación múltiple de órdenes implícita mediante promedios:
  El efecto del despacho de más de una orden a un mismo repartidor se incorpora indirectamente a través del valor promedio de la tasa de utilización, ocultando la considerable variabilidad que puede existir entre intervalos específicos. Esto conduce a errores frecuentes en la estimación efectiva de la flota necesaria, particularmente en períodos donde los agrupamientos de pedidos para un mismo repartidor se dan a niveles irregulares.

Estas limitaciones conducen a una revisión constante de las sugerencias generadas automáticamente por el modelo actual, exigiendo ajustes manuales frecuentes por parte de los equipos operativos locales, lo que implica una alta dedicación de tiempo y recursos.

En definitiva, el problema central abordado por esta tesis es superar las restricciones del enfoque determinista y heurístico actual, proponiendo una metodología más robusta basada en la teoría de colas que permita incorporar explícitamente la variabilidad observada en la operación. Este trabajo no abordará la lógica específica de asignación de pedidos, y utilizará en cambio un supuesto simplificado de asignación FIFO (primero en llegar, primero en ser atendido). El objetivo último es lograr una estimación más precisa de la cantidad de repartidores requerida por franja horaria, reducir la necesidad de ajustes manuales constantes, y mejorar así la eficiencia operativa general del sistema.

1.3 Objetivo

El objetivo principal de esta tesis es desarrollar un modelo analítico basado en teoría de colas, calibrado empíricamente con datos históricos de una operación real de entrega a domicilio en entornos urbanos, con el fin de mejorar la precisión en la estimación intradiaria del número de repartidores necesarios.

En particular, esta investigación busca alcanzar los siguientes objetivos específicos:

- Caracterizar estadísticamente las principales variables operativas (tasa de arribo de pedidos y tiempos efectivos de servicio), determinando las distribuciones probabilísticas que representen más adecuadamente la dinámica real del sistema.
- Desarrollar y calibrar un modelo de colas con múltiples servidores que permita estimar con mayor precisión la demanda de repartidores para cada intervalo horario (bucket), incorporando de forma explícita la variabilidad observada en la operación diaria.
- Evaluar comparativamente el desempeño del modelo propuesto frente al enfoque heurístico actual, identificando mejoras específicas en métricas críticas como el tiempo de espera, la utilización efectiva de la flota y la reducción del riesgo de saturación del sistema.
- Brindar recomendaciones concretas basadas en los resultados obtenidos, orientadas a facilitar una planificación operativa diaria más eficiente, robusta y precisa, que minimice la necesidad de revisiones y ajustes manuales frecuentes por parte de los equipos locales.

Mediante el logro de estos objetivos específicos, esta tesis busca contribuir directamente al desarrollo de herramientas analíticas más adecuadas para la gestión de la flota requerida a nivel intradiario en plataformas de entrega a domicilio urbano, generando un impacto positivo tanto en la eficiencia operativa como en la calidad percibida por los usuarios finales.

2 Datos

2.1 El conjunto de datos

Esta sección presenta las características de los datos utilizados, su estructura, el tratamiento previo aplicado y los principales atributos relevantes para el análisis y modelado del sistema de entrega a domicilio.

2.1.1 Fuente y alcance

Los datos fueron provistos por una plataforma de entrega a domicilio en línea con presencia operativa en Argentina, correspondientes a operaciones realizadas en una de las zonas logísticas de mayor volumen del país. Este nivel representa la unidad mínima de planificación local empleada por la compañía: múltiples zonas conforman una ciudad, y varias ciudades componen la operación de un país. La elección de esta zona se justifica por su densidad operativa, lo cual permite observar dinámicas representativas en contextos de alta exigencia logística.

El período analizado abarca desde el 1 de septiembre hasta el 30 de noviembre de 2024, incluyendo un total de 1.334.183 órdenes distribuidas en 2.736 buckets de 15 minutos a lo largo de los 91 días del trimestre. Este intervalo temporal fue seleccionado para capturar patrones estables de demanda, evitando semanas con eventos estacionales o irregularidades marcadas.

Se descartaron únicamente aquellos registros con inconsistencias temporales (como registros de tiempos negativos o fuera de orden), órdenes incompletas o canceladas antes de la aceptación por parte del repartidor, y variables vacías o redundantes sin aporte al análisis. Estas depuraciones representaron un porcentaje marginal del total (< 1,5 %) y no alteraron la estructura general de los datos.

2.1.2 Estructura

El conjunto de datos posee una estructura tabular, donde cada fila representa una orden individual con información completa sobre su ciclo operativo. Las variables relevantes y consideradas para el ejercicio son:

- *id:* Identificador único de cada orden.
- *created_at_date_local / created_at_time_local:* Fecha y hora local de generación de la orden.
- *bucket:* Intervalo de 15 minutos al cual se asigna cada orden, utilizado como unidad de análisis.

- *day_of_week_text:* Día de la semana de la creación de la orden (formato textual).
- *accepted_at_time_local / dropped_off_at_time_local:* Fecha y hora de aceptación y entrega de la orden por parte del repartidor.
- *service_time_min:* Tiempo (minutos) desde la aceptación hasta la entrega.
- *to_vendor_time:* Tiempo(minutos) hasta llegar al comercio.
- *at_vendor_time:* Tiempo (minutos) esperando la preparación del pedido.
- *to_customer_time:* Tiempo (minutos) hasta llegar al cliente.
- *at_customer_time:* Tiempo (minutos) en el punto de entrega.
- *interarrival_time:* Tiempo (segundos) entre la creación de la orden y la anterior

La base resultante está conformada por datos consistentes y completos, con baja proporción de valores nulos o erróneos. Esta calidad de origen permitió trabajar sobre el conjunto de datos completo, sin necesidad de realizar filtrados adicionales ni transformaciones avanzadas. La estructura detallada de las variables habilita un análisis riguroso del comportamiento operativo del sistema, tanto a nivel agregado como por franjas horarias y segmentos semanales, sentando las bases para el análisis exploratorio y la posterior simulación del modelo.

2.2 Análisis descriptivo exploratorio

Antes de avanzar con la formulación del modelo y los experimentos de simulación, resulta fundamental comprender en profundidad el comportamiento operativo observado en el sistema real. Esta sección presenta un análisis exploratorio de las variables identificadas en la sección anterior, con foco en la dinámica de creación y servicio de órdenes, sus patrones temporales y su variabilidad. El objetivo es identificar regularidades y puntos críticos que permitan sustentar conceptualmente las decisiones de modelado y, al mismo tiempo, validar empíricamente los supuestos adoptados en las etapas posteriores. A partir de este diagnóstico, se podrá evaluar con mayor precisión la adecuación de la heurística actual, así como el impacto potencial de las configuraciones optimizadas.

Con ese propósito, se calcularon estadísticas descriptivas para las principales variables del ciclo operativo: la tasa de arribo, el tiempo entre arribos (interarrival_time), el tiempo de servicio (service_time_min), el tiempo al local (to_vendor_time) , el tiempo en el local (at_vendor_time), el tiempo al domicilio del cliente (to_customer_time) y el tiempo en el domicilio del cliente (at_customer_time). La Tabla 3 del Apéndice A presenta un resumen global de estas variables, incluyendo medidas de tendencia central, dispersión y percentiles. Por su parte, la Tabla 4 del mismo apéndice muestra únicamente las medias de cada métrica desagregadas por intervalos de 15 minutos entre las 07:00 y las 02:00. Esta apertura horaria permite visibilizar de forma clara

las tendencias a lo largo del día, facilitando la identificación de comportamientos diferenciados según la franja horaria.

2.2.1 La tasa de arribo y el tiempo entre arribos

La tasa de arribo define la frecuencia con la cual las órdenes llegan dentro de un intervalo de tiempo. Dado que los turnos en el sistema se reservan cada 15 minutos, este parámetro se expresa como la cantidad promedio de órdenes por minuto. Es un indicador clave para evaluar qué tan demandado se encuentra el sistema en distintos momentos del día, y un factor determinante en la planificación de los recursos necesarios para responder eficientemente a las variaciones en la demanda. La tasa de arribo representada por la letra lambda, se expresa en órdenes promedio por minuto y se calcula dividiendo el número total de órdenes registradas en un intervalo de tiempo por la duración de dicho intervalo. En este caso, para intervalos de 15 minutos:

$$\lambda = \frac{Total\ de\ Órdenes}{Duración\ del\ intervalo} \tag{2}$$

Por otro lado, la regularidad o aleatoriedad entre los tiempos de llegada de las órdenes influye directamente en la carga de trabajo del sistema y, en consecuencia, en la capacidad para gestionar los tiempos de espera de los usuarios y mantener la calidad del servicio.

En este contexto, resulta crucial complementar el análisis de la tasa de arribo con la métrica del tiempo entre arribos, que mide el tiempo transcurrido entre la llegada de un pedido y el siguiente. Esta métrica proporciona una perspectiva granular sobre las dinámicas del sistema, permitiendo evaluar no solo el volumen promedio de llegadas, sino también la aleatoriedad e independencia entre eventos consecutivos. La métrica se expresa en unidades de tiempo (en segundos) entre dos llegadas consecutiva y se calcula tomando la diferencia de tiempo entre el momento en que ocurre una orden y la siguiente:

$$Interarrival\ Time = t_{(i+1)} - t_{(i)} \tag{3}$$

El tiempo entre arribos está intrínsecamente relacionado con la tasa de arribo: en sistemas donde los eventos ocurren de manera aleatoria e independiente, el tiempo entre arribos sigue una distribución exponencial. Esta relación matemática permite validar que la tasa de arribo, en términos agregados, sigue un modelo de Poisson. Dicho de otro modo, al comprobar que el tiempo entre arribos cumple con las propiedades de la distribución exponencial, se refuerza la validez del modelo de Poisson para describir la tasa de arribo en el sistema.

2.2.1.1 Comportamiento de la tasa de arribo

La tasa de arribo de pedidos presenta una dinámica fuertemente dependiente del día de la semana y la franja horaria. Esta variación refleja patrones de consumo arraigados en los hábitos diarios de los usuarios, particularmente asociados a las comidas principales (almuerzo y cena), que concentran la mayor parte del volumen de órdenes procesadas. En la Figura 1 se ilustra el patrón intradiario promedio de llegadas, expresado en órdenes por minuto, para cada día de la semana, restringido a la ventana operativa observada (07:00-02:00 del día siguiente).

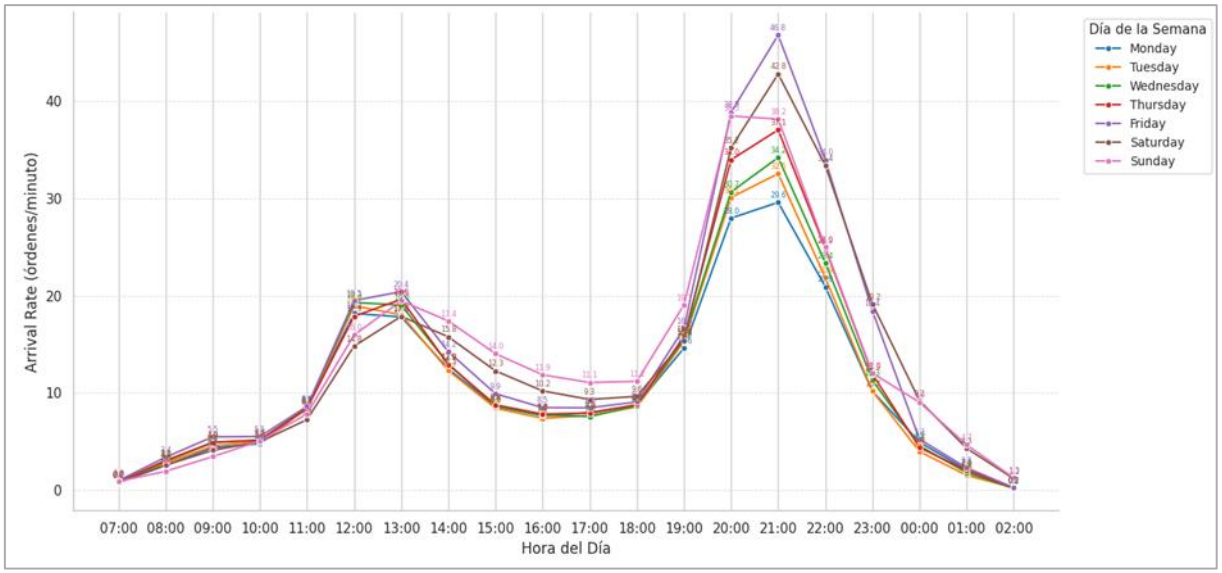

Figura 1 - Tasa de arribo promedio por hora, según día de la semana

Este gráfico permite identificar de forma clara la estructura de picos y valles que caracteriza a la operación (donde pico se refiere, según la terminología utilizada internamente en la empresa, a los momentos de mayor concentración de demanda), con diferencias evidentes entre días laborables y fines de semana:

Picos y valles de actividad

Se distinguen tres momentos principales de alta demanda:

- Crecimiento matinal (11:00-13:00): crecimiento progresivo de la demanda, con una transición suave hacia el pico de almuerzo.
- Pico de almuerzo (12:00-14:30): es el primer gran pico del día, con tasas que oscilan entre 18 y 22 órdenes por minuto en la mayoría de los días, alcanzando hasta 24 en domingo.
- Pico nocturno (20:00-22:00): representa el momento de mayor actividad operativa, particularmente los viernes (hasta 48.6 órdenes/min) y sábados (~43 órdenes/min), con valores notablemente más altos que en días laborables.

- El valle vespertino (15:00-18:30) se presenta como una caída marcada en la tasa de arribo, donde se estabiliza entre 8 y 12 órdenes por minuto. Este comportamiento se observa con mayor nitidez en días laborales, mientras que en fines de semana la caída es más atenuada y progresiva.

Diferencias entre días laborables y fines de semana

- Días laborables (lunes a jueves):
    - Exhiben una curva de demanda más estable y predecible, con valles bien definidos entre el almuerzo y la cena.
    - El pico nocturno es menos agresivo (≈30-35 órdenes/min), y el descenso posterior es más rápido, cerrando la operación con menores niveles de arribo después de las 23:00.
- Fines de semana (viernes, sábado y domingo):
    - Muestran una demanda más intensa y extendida: los picos son más altos y prolongados, especialmente hacia la noche.
    - Los viernes en particular concentran el mayor volumen de pedidos, y la actividad se mantiene elevada incluso después de la medianoche.
    - Sábados y domingos presentan curvas más "ensanchadas", con menor diferencia entre los picos y los valles, reflejando una demanda más continua y menos segmentada.

2.2.1.2 Distribución empírica de la tasa de arribo

La caracterización estadística de la tasa de arribo, medida en órdenes por minuto, permite complementar el análisis descriptivo anterior con una mirada más agregada sobre la dispersión y comportamiento de esta variable clave del sistema. En la Figura 2 se presenta el histograma de la tasa de arribo promedio, construido a partir de todos los buckets horarios (intervalos de 15 minutos) del conjunto de datos, considerando todos los días del período analizado.

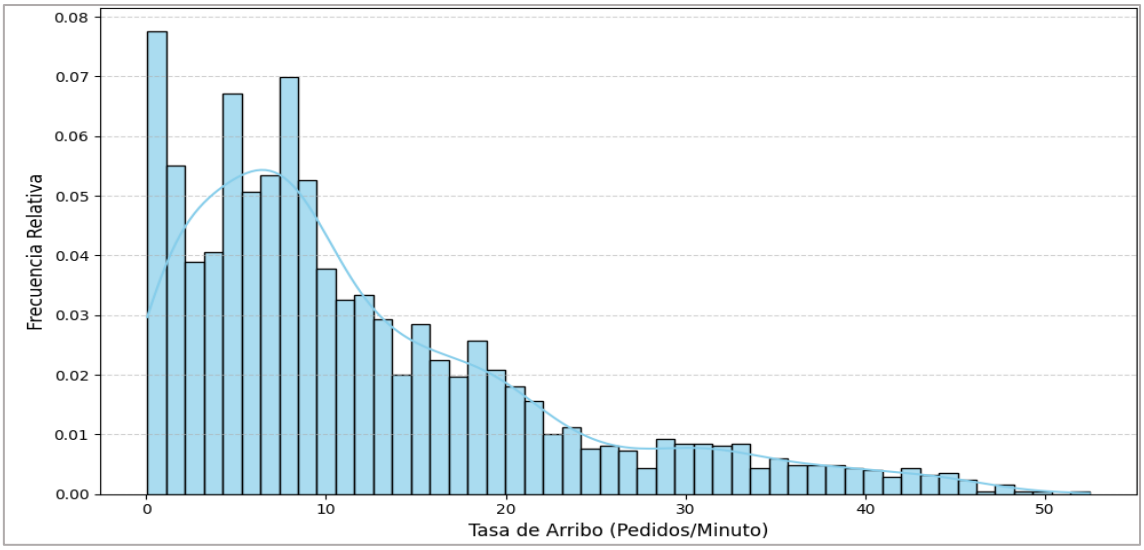

Figura 2 - Frecuencia relativa de la tasa de arribo (pedidos por minuto)

El gráfico revela una distribución con las siguientes características:

- Asimetría positiva marcada:

  La distribución presenta una cola larga hacia la derecha, con una concentración significativa de valores en los rangos bajos de tasa de arribo. Más del 60% de los buckets tienen una tasa menor a 12 órdenes/minuto, mientras que los valores extremos, por encima de 30, son poco frecuentes, pero altamente relevantes desde el punto de vista operativo. Esta asimetría sugiere que la mayor parte del día se concentra en momentos de baja o media actividad, con eventos más intensos agrupados en ventanas temporales específicas (principalmente los picos del almuerzo y la noche).

- Moda baja y comportamiento típico:

  La moda de la distribución se ubica entre 5 y 10 pedidos por minuto, lo que representa la tasa más común en el sistema. Este rango puede interpretarse como el régimen "normal" de operación, donde la infraestructura logística mantiene estabilidad sin requerir recursos adicionales.

- Eventos extremos e implicancias operativas:

  Los valores superiores a 25-30 pedidos/minuto, aunque minoritarios en frecuencia, constituyen picos operativos críticos que deben ser contemplados explícitamente en la planificación de recursos y dimensionamiento de la flota. Su ocurrencia, aunque esporádica, puede generar cuellos de botella si no se cuenta con buffers suficientes.

- Aporte de la curva de densidad KDE:

  La curva KDE (Kernel Density Estimation), superpuesta al histograma, refuerza la idea de que la tasa de arribo no sigue una distribución simétrica ni normal, invalidando

supuestos clásicos de modelado basados en Gaussianas. Este hallazgo motiva el enfoque adoptado en el presente trabajo, donde se utilizan simulaciones empíricas basadas en muestras reales por bucket horario, en lugar de asumir una distribución teórica para esta variable.

Este comportamiento de la tasa de arribo, altamente asimétrico y disperso, refuerza la necesidad de modelar la operación como un sistema estocástico con alta variabilidad intertemporal. La incorporación de esta variabilidad dentro del modelo de colas utilizado permitirá analizar con mayor fidelidad la presión ejercida sobre el sistema en distintos momentos del día, así como evaluar configuraciones más robustas en contextos de incertidumbre.

2.2.1.3 Distribución de Poisson

Como se mencionó previamente, en teoría de colas la frecuencia de llegadas se modela habitualmente bajo el supuesto de que los eventos de llegada son independientes entre sí y ocurren de manera aleatoria, lo que permite suponer una distribución de Poisson en el proceso. En un proceso de Poisson, la probabilidad de que ocurra un cierto número de llegadas en un intervalo de tiempo específico es independiente de lo que ocurre en otros intervalos. Esto significa que la llegada de un pedido no influye ni depende de las llegadas anteriores o futuras, lo que define un proceso aleatorio sin memoria.

Matemáticamente, la distribución de Poisson es una función de probabilidad discreta que describe la probabilidad de que ocurran exactamente k eventos en un intervalo fijo de tiempo o espacio, dado un promedio conocido de ocurrencias, λ (lambda). De esta forma, la probabilidad de k llegadas en un intervalo se expresa como:

$$P(K = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4}$$

En el contexto de la operación de entrega a domicilio, esta función permite modelar las llegadas de pedidos a la plataforma durante un intervalo de tiempo definido, como un minuto o un cuarto de hora. El parámetro λ representa el promedio de pedidos que se generan en ese intervalo, considerando la demanda histórica del sistema. Por ejemplo, si λ es 5, la función de Poisson calculará la probabilidad de observar 0, 1, 2, ..., hasta más de 5 pedidos en ese intervalo.

Volviendo al supuesto de independencia, este modelo se sustenta en la naturaleza descentralizada y autónoma de los usuarios que generan los pedidos. Cada usuario decide cuándo realizar su pedido, y estas decisiones no están sincronizadas ni coordinadas entre sí. Aunque los picos de demanda, como los de almuerzo y cena, puedan incrementar el valor de λ

en ciertos intervalos, dentro de esos picos las llegadas de pedidos siguen ocurriendo de manera suficientemente aleatoria. Esto refleja la interacción no sincronizada de múltiples usuarios con la plataforma, respaldando así el supuesto de independencia y la naturaleza sin memoria del proceso.

Por tanto, la función de probabilidad de Poisson no sólo cuantifica la probabilidad de diferentes niveles de carga en el sistema, sino que también traduce matemáticamente la variabilidad inherente a la operación de entrega a domicilio, ofreciendo una base sólida para modelar las llegadas como un proceso aleatorio en el tiempo.

2.2.1.4 Validación del modelo de Poisson para la tasa de arribo

Matemáticamente, si la tasa de arribo sigue una distribución de Poisson con parámetro λ (tasa promedio de llegadas por unidad de tiempo), entonces el tiempo entre arribos (Δt) sigue una distribución exponencial con parámetro λ (1/λ como escala de tiempo). Como se estableció previamente la distribución exponencial caracteriza procesos sin memoria, lo que significa que la probabilidad de que ocurra una llegada en el futuro no depende de eventos pasados, cumpliendo así con el supuesto de independencia. Validar que los tiempos entre llegadas sigan una distribución exponencial es equivalente a verificar que las llegadas sean independientes y aleatorias, lo que a su vez avala el uso de la distribución de Poisson para modelar el tasa de llegada.

Durante el análisis, se observó que, en algunos casos, la prueba de bondad de ajuste para la distribución de Poisson aplicada a la tasa de arribo no arrojaba resultados del todo satisfactorios. Sin embargo, se verificó que el tiempo entre llegadas sí seguía una distribución exponencial en la mayoría de los buckets analizados. Esto es relevante porque, aunque la tasa de arribo modelada como Poisson pueda fallar en pruebas estadísticas debido a restricciones como el tamaño de muestra o eventos extremos, el hecho de que el tiempo entre llegadas se ajuste bien a una distribución exponencial proporciona evidencia suficiente de que las llegadas son aleatorias e independientes.

Los pasos seguidos para validar el supuesto de ajuste fueron:

1. **Cálculo del Tiempo entre llegadas**

   Se utilizó la columna `interarrival_time` del conjunto de datos, que representa los tiempos entre llegadas consecutivas en segundos. Estos valores fueron calculados para cada bucket (intervalo de 15 minutos) y categorizados según el día de la semana y hora del día.

2. **Aplicación del Test de Kolmogorov-Smirnov (K-S)**

Para cada bucket, se estimó el parámetro λ como la inversa del promedio de los tiempo entre llegadass.El test K-S compara la distribución acumulada empírica de los datos con la distribución teórica exponencial definida por λ.

3. **Cálculo del Estadístico K-S y el p-valor**

   El estadístico K-S mide la máxima diferencia entre las distribuciones empírica y teórica. Un p-valor mayor a 0.05 indica que no se puede rechazar la hipótesis nula de que los datos siguen una distribución exponencial.

El test de Kolmogorov-Smirnov fue aplicado a todos los buckets del conjunto de datos. En total, el 92.5 % de los buckets analizados presentaron p-valores mayores a 0.05, indicando que el interarrival_time sigue una distribución exponencial en la mayoría de los casos. Los casos que no cumplieron con el supuesto de ajuste se detallan en la Tabla 5 del apéndice B. Si bien estos últimos representan apenas el 7.5 % del total, se analizan brevemente en dicho apéndice con el objetivo de ofrecer mayor transparencia metodológica y contextualizar posibles fuentes de desvío.

A continuación, a modo de ejemplo, se presentan los resultados obtenidos para los viernes en el bucket de las 00:00.

- *λ Calculado: 0.12*
- *Estadístico K-S: 0.0155*
- *p-valor: 0.8834*

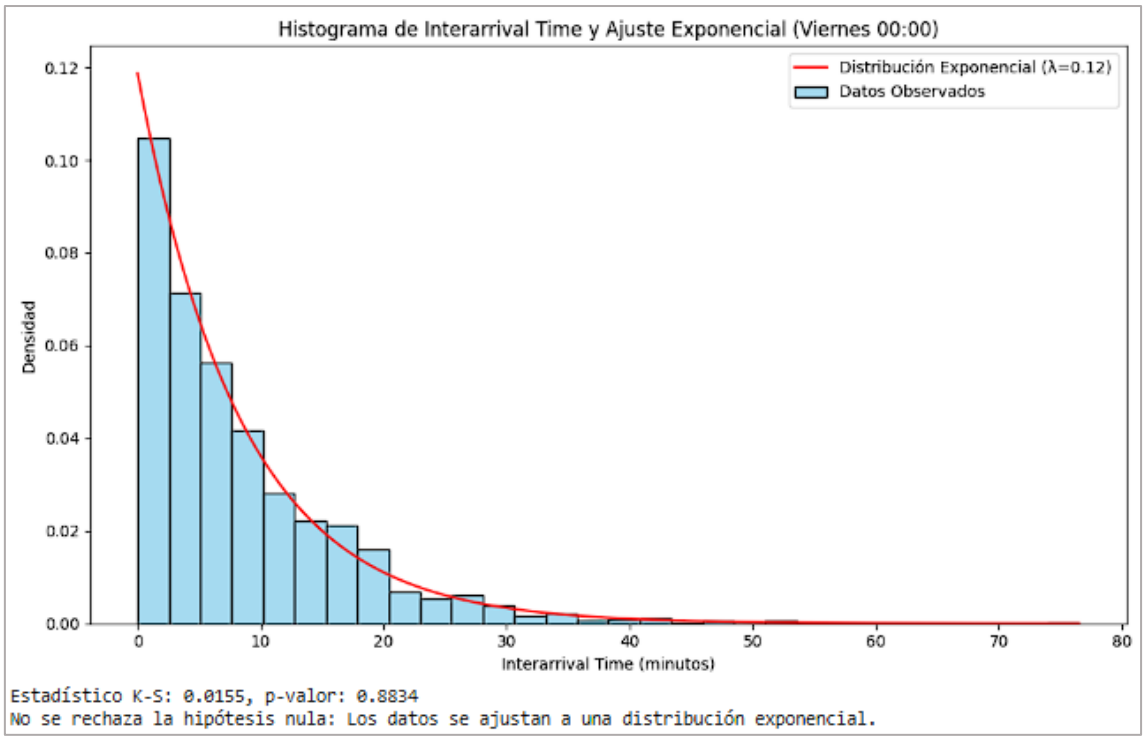

Figura 3 -Histograma con curva teórica (viernes 00:00)

La figura 3 muestra un ajuste satisfactorio, reforzando la validez del supuesto. En caso de que futuros análisis requieran mayor granularidad, se podría evaluar individualmente el ajuste para buckets específicos o considerar enfoques alternativos. Sin embargo, los resultados obtenidos son suficientes para validar el uso del modelo de Poisson en el sistema analizado.

2.2.2 El tiempo de servicio

El tiempo de servicio representa el intervalo necesario para que un repartidor complete un pedido desde el momento de aceptación hasta la entrega final al cliente. En el contexto del sistema de entrega a domicilio, este parámetro resume el tiempo efectivo que un servidor (repartidor) dedica a la atención de una orden, y por lo tanto, constituye una de las variables más relevantes para entender la capacidad operativa del sistema. Es un concepto central tanto desde el punto de vista logístico como desde el marco de la teoría de colas. A nivel operativo, impacta directamente en los tiempos de espera, la calidad del servicio y el dimensionamiento necesario de la flota. A nivel analítico, forma parte del núcleo del modelo de colas utilizado, y condiciona el comportamiento esperado del sistema frente a diferentes configuraciones de entrada.

El tiempo total de servicio se compone de distintas etapas, cada una asociada a momentos específicos del proceso de entrega y sujeta a fuentes de variabilidad propias. Para efectos de este análisis, se consideran cuatro componentes:

- Tiempo hasta el local (*to_vendor_time*): Tiempo desde la aceptación de la orden hasta la llegada del repartidor al restaurante. Refleja distancia recorrida y disponibilidad operativa de la flota.
- Tiempo en el local (*at_vendor_time*): Tiempo de espera en el local hasta que el pedido está listo para ser recogido. Está influido por la eficiencia del restaurante y la congestión operativa.
- Tiempo hasta el domicilio del cliente (*to_customer_time*): Tiempo de desplazamiento entre el restaurante y el domicilio del cliente. Depende de la distancia y las condiciones del tráfico urbano.
- Tiempo en el domicilio del cliente (*at_customer_time*): Tiempo dedicado a la interacción con el cliente en el punto de entrega. Puede incluir demoras por verificación del pedido o esperas asociadas al contacto final.

Este desglose permite capturar con mayor precisión las causas subyacentes de variabilidad dentro del tiempo de servicio, y facilita un análisis más granular en secciones posteriores.

En los modelos clásicos de teoría de colas, el tiempo de servicio suele modelarse mediante una distribución exponencial. Esta elección permite un tratamiento analítico sencillo del sistema, al asumir que los servicios tienen una duración promedio constante y siguen un patrón de ocurrencia aleatorio. No obstante, en un contexto operativo real como es el sistema de entrega a domicilio tratado en este trabajo, donde el tiempo de servicio resulta de la combinación de múltiples etapas heterogéneas, esta suposición puede no reflejar con precisión la distribución observada en los datos. Por esta razón, en las secciones siguientes se analizará empíricamente el comportamiento del tiempo de servicio, evaluando su ajuste a la distribución exponencial y considerando alternativas en caso de ser necesario.

2.2.2.1 Comportamiento tiempo de servicio

La Figura 4 presenta la evolución intradiaria del tiempo de servicio promedio para cada día de la semana, junto con el volumen promedio de órdenes por hora (barras grises). Este gráfico permite visualizar no solo las diferencias entre franjas horarias, sino también los contrastes entre días, resaltando la variabilidad estructural del sistema.

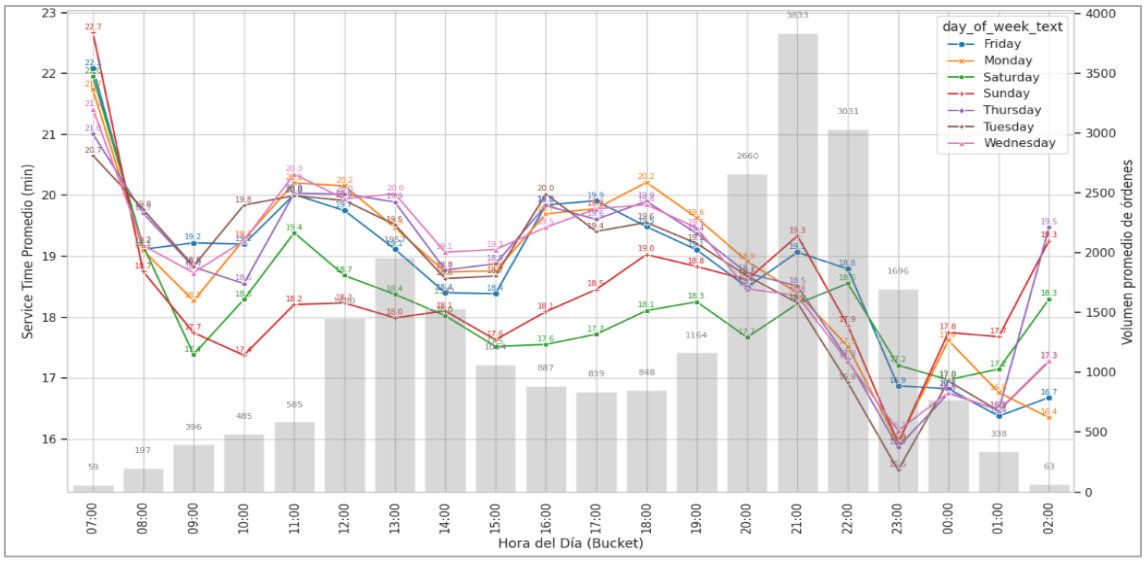

Figura 4 – Tiempo de servicio promedio por hora, según día de la semana

El análisis del tiempo de servicio promedio a lo largo del día revela una estructura intradiaria compleja, marcada por variaciones sistemáticas que reflejan las dinámicas operativas del sistema. A fin de capturar adecuadamente estos patrones, se segmentó la jornada en bloques horarios que agrupan comportamientos similares observados en el gráfico:

Patrones generales por franja horaria

- *07:00-10:00: Inicio de operaciones y descenso inicial*

  Durante las primeras horas de actividad, el tiempo de servicio parte de niveles elevados (≈21 minutos), reflejando los efectos de la activación operativa tras la inactividad nocturna y los bajos volúmenes de demanda (<50 órdenes por bucket). A medida que se estabiliza el flujo de pedidos, el tiempo de servicio disminuye progresivamente, alcanzando valores en torno a 18–18.5 minutos hacia las 10:00.
- *10:00-13:00: Ascenso hacia el pico de almuerzo*

  En este tramo, el tiempo de servicio vuelve a incrementarse de manera sostenida en paralelo al aumento de la demanda, que alcanza entre 300 y 400 órdenes por bucket. El tiempo de servicio se estabiliza entre 18.5 y 19.2 minutos, evidenciando un sistema operando bajo condiciones de saturación moderada.
- *13:00-15:00: Descenso post-pico de almuerzo*

  Posteriormente al pico de mediodía, se registra una disminución en el tiempo de servicio hacia valores de 18–18.2 minutos, en correspondencia con la retracción del volumen de órdenes. Esta caída sugiere una fase de descompresión operativa y recuperación de eficiencia.
- *15:00-18:00: Meseta de estabilidad operativa*

Durante la tarde, el tiempo de servicio se mantiene relativamente estable, oscilando entre 18 y 18.5 minutos, con un leve repunte hacia las 18:00. Esta franja refleja un equilibrio operativo, en el cual la presión de demanda es moderada y el sistema sostiene niveles de servicio constantes.

- *18:00-21:00: Pico de cena*

  El segundo gran pico de demanda diaria, caracterizado por volúmenes de hasta 600 órdenes por bucket, no genera un deterioro significativo en los tiempos de servicio, que se mantienen entre 18.5 y 19.5 minutos. Esta estabilidad frente a altos volúmenes destaca la resiliencia del sistema para absorber cargas intensas sin afectar los niveles de servicio.

- *21:00-23:00: Transición nocturna*

  Tras el pico vespertino, el tiempo de servicio experimenta una caída progresiva desde valores cercanos a 18.5 minutos hasta alcanzar mínimos de aproximadamente 15.5 minutos hacia las 23:00. Esta reducción acompaña la disminución abrupta de la demanda, indicando una mayor fluidez en los procesos de entrega.

- *23:00-02:00: Nocturna tardía y repunte final*

  Durante la madrugada, el tiempo de servicio se estabiliza inicialmente en torno a 16-16.5 minutos. Sin embargo, pasada la 01:00, se observa un leve repunte hasta valores cercanos a 17 minutos, posiblemente atribuible a mayores tiempos de espera generados por la escasa densidad de órdenes y repartidores activos.

Esta caracterización evidencia que el tiempo de servicio no presenta un comportamiento homogéneo a lo largo del día, sino que atraviesa fases diferenciadas, de aceleración, saturación, recuperación y dispersión, que deben ser consideradas en el modelado operativo.

Variabilidad intersemanal

Más allá de las tendencias intradiarias, se identifican diferencias relevantes entre días de la semana:

- Martes y miércoles presentan una mayor dispersión, especialmente en los tramos de baja demanda (mañana y madrugada), indicando una menor robustez en la planificación operativa.
- Jueves anticipa el comportamiento típico de fines de semana, con una tendencia hacia mayor estabilidad a partir de la tarde.

- Viernes y sábados exhiben notable estabilidad a lo largo de todas las franjas, incluso bajo condiciones de alta demanda, lo que sugiere una mayor madurez operativa o una mejor capacidad adaptativa en estos días.
- El domingo muestra un patrón más errático, con tiempos de servicio elevados tanto en el inicio como en el cierre de la jornada, posiblemente reflejando una demanda menos estructurada.

En términos generales, se evidencia una relación inversa entre el volumen de órdenes y el tiempo de servicio: cuando la demanda disminuye, los tiempos tienden a incrementarse, posiblemente debido a mayores tiempos de espera para iniciar las entregas. Sin embargo, esta relación no es estrictamente lineal. Un caso particular se observa en la franja 22:00–23:00, donde la caída en la demanda no se traduce en un aumento inmediato del tiempo de servicio, lo que sugiere la influencia de dinámicas operativas internas. Aunque a priori podría resultar contraintuitivo, una posible explicación radica en la menor inercia operativa fuera de los picos: con una flota menos activa y una menor presión sobre la operación, tanto repartidores como usuarios podrían mostrar comportamientos diferentes. Es posible que los pedidos sean de otra naturaleza, que la disponibilidad de los clientes para recibir los productos sea más inmediata, o que se activen tiempos de respuesta más cortos en los nodos operativos. Esta hipótesis será explorada con mayor detalle al descomponer empíricamente los distintos componentes del tiempo de servicio en las secciones siguientes.

2.2.2.2 Distribución empírica del tiempo de servicio

A fin de caracterizar empíricamente el comportamiento del tiempo de servicio, se construyeron histogramas de frecuencia relativa para cada día de la semana, representando la distribución de los tiempos de servicio observados en intervalos de 15 minutos. Cada histograma incluye, además, una estimación de densidad (KDE) suavizada para facilitar la visualización de la forma general de la distribución. La Figura 5 muestra la distribución por día, permitiendo identificar patrones comunes y diferencias sutiles entre jornadas laborales y fines de semana.

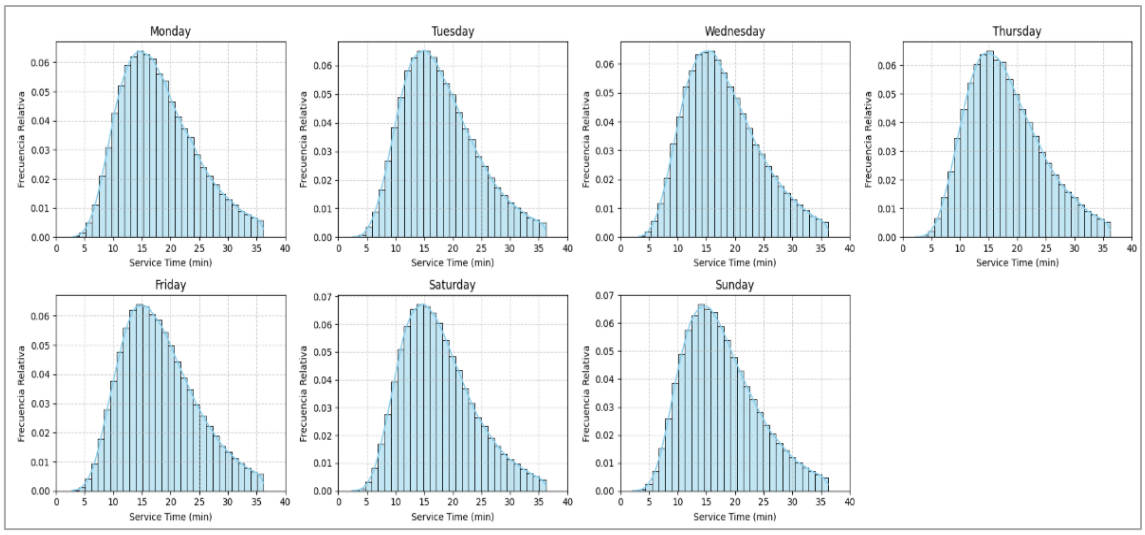

Figura 5 - Distribución empírica del tiempo de servicio por día de semana

La distribución del tiempo de servicio presenta patrones consistentes a lo largo de los distintos días de la semana, evidenciando una estructura asimétrica positiva (cola hacia la derecha). En todos los casos, se observa una alta concentración de tiempos de servicio en el rango de 10 a 20 minutos, seguida de una disminución progresiva en la frecuencia de valores más elevados.

Sin embargo, pueden destacarse algunas diferencias sutiles entre días:

- Fines de semana (sábado y domingo) exhiben un pico de frecuencia más marcado y concentrado alrededor de los 10–15 minutos, con una dispersión relativamente menor hacia tiempos elevados. Esta concentración sugiere una mayor homogeneidad en la operación y podría estar asociada a condiciones de demanda más estables o flujos de pedidos más predecibles.
- Días laborables (especialmente lunes y martes) muestran distribuciones más extendidas, con colas más largas hacia la derecha. Esto indica una mayor proporción de pedidos con tiempos de servicio prolongados, reflejando posiblemente una mayor variabilidad operativa, fluctuaciones en la disponibilidad de repartidores o ineficiencias asociadas a menores niveles de actividad.
- Viernes y jueves presentan un comportamiento intermedio: aunque siguen una estructura asimétrica, su dispersión es menor que la de lunes y martes, acercándose progresivamente al patrón observado en fines de semana.

Un aspecto relevante es la ausencia de multimodalidad en los histogramas. Cada día muestra un único pico dominante, sin evidencias de subpoblaciones diferenciadas (por ejemplo, no se observa la coexistencia de dos tipos de entregas claramente separadas en velocidad). Esto

refuerza la hipótesis de que el tiempo de servicio se explica por un único proceso estocástico principal, aunque con variabilidad inherente.

Finalmente, la asimetría consistente y la presencia de colas largas en todos los días avalan la necesidad de modelar el tiempo de servicio utilizando distribuciones flexibles que puedan capturar tanto la concentración principal como la existencia de valores extremos. Esta observación respalda la elección de distribuciones como Gamma o Lognormal para la representación empírica, en lugar de la distribución exponencial clásicamente asumida.

2.2.2.3 Validación del modelo de distribución del tiempo de servicio

Para evaluar de manera rigurosa la distribución empírica del tiempo de servicio, se realizó un test de bondad de ajuste basado en el estadístico de Kolmogorov-Smirnov (KS) comparando cinco distribuciones teóricamente relevantes: Gamma, Weibull, Lognormal, Exponencial y Normal.

La metodología seguida constó de los siguientes pasos:

1. El análisis se efectuó de manera diferenciada por hora del día (bucket horario), a fin de capturar las variaciones intradiarias observadas en secciones previas.
2. Para cada bucket, se filtraron los datos extremos aplicando un recorte entre los percentiles 1% y 99%, asegurando una representación más robusta de la variabilidad típica.
3. Sobre el conjunto filtrado, se ajustaron las distribuciones y se evaluaron mediante el test KS, utilizando el p-valor asociado como criterio de selección de la mejor distribución para cada hora.

Para mayor detalle la tabla 6 en el Apéndice C resume la mejor distribución encontrada para cada bucket. A modo de ejemplo, la Figura 6 ilustra el ajuste realizado para la hora 00:00, mostrando las distribuciones evaluadas sobre los datos observados. En este ejemplo, la distribución Gamma obtuvo el mejor desempeño (p-valor = 0.1111), superando a las demás alternativas.

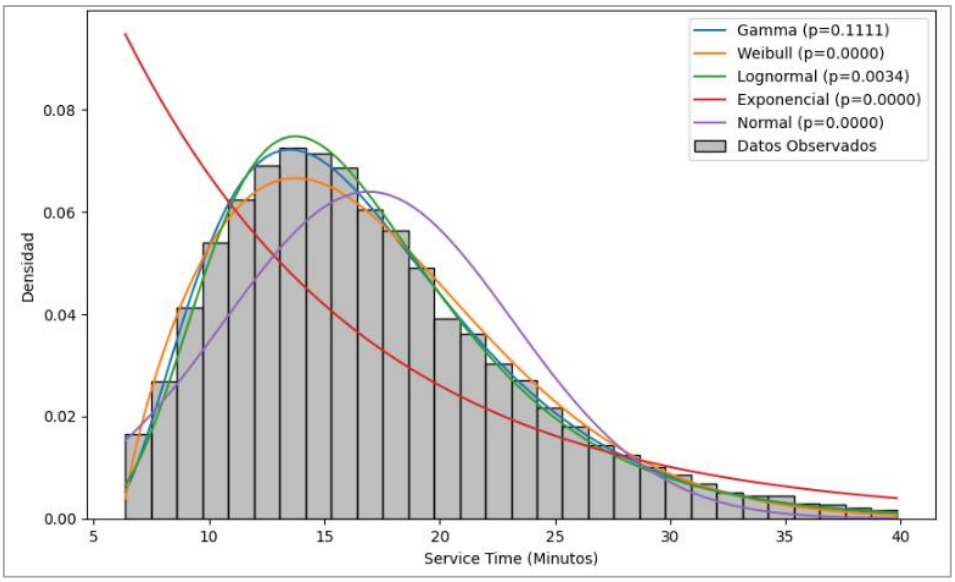

Figura 6 - Comparación de ajuste del tiempo de servicio con distintas distribuciones teóricas

El resumen de la validación por hora indica que:

- En 21 de 24 horas analizadas, la distribución Gamma fue la que obtuvo el mejor ajuste según el test de KS.
- En las franjas horarias de 07:00 a 08:00, el ajuste favoreció a una distribución Lognormal, aunque con un p-valor moderado.
- A pesar de pequeñas variaciones en los niveles de significancia (p-valor), en general la distribución Gamma capturó adecuadamente el patrón empírico del Tiempo de servicio.

Si bien el ajuste global a una distribución Gamma es en general satisfactorio, los resultados muestran que la adherencia no es perfecta en todos los intervalos horarios, particularmente en aquellas horas de menor volumen de órdenes o mayor dispersión de tiempos. Dado este hallazgo, y considerando además la descomposición multicomponente del tiempo de servicio (desplazamiento, espera, entrega), se plantea como opción complementaria para el modelado simular el tiempo de servicio de forma estocástica a partir de muestras empíricas históricas en lugar de basarse exclusivamente en una distribución teórica ajustada. Esta estrategia híbrida permite capturar tanto la tendencia central observada como la dispersión real, fortaleciendo la robustez del modelo de simulación desarrollado en esta tesis.

2.2.3 Los componentes del tiempo de servicio

Si bien el análisis intradiario y la distribución general del tiempo de servicio permiten caracterizar su comportamiento global, resulta necesario profundizar aún más en la estructura interna de este tiempo. Dado que el tiempo de servicio en sistemas de entrega a domicilio de última milla es el resultado de la combinación de varias etapas operativas (desplazamiento al comercio,

espera en el local, desplazamiento al cliente y entrega final), su variabilidad podría surgir de dinámicas diferenciadas en cada uno de estos componentes.

Con el objetivo de entender mejor los factores que explican las fluctuaciones observadas, en esta sección se analizará de manera desagregada cada componente del tiempo de servicio. Esto permitirá identificar patrones particulares, evaluar si ciertos subprocesos son más críticos en la variabilidad total, y aportar evidencia adicional para validar la elección de un enfoque de modelado estocástico en la representación del sistema.

2.2.4 Tiempo hasta el local

2.2.4.1 Análisis de patrones y tendencias

La Figura 7 muestra la evolución intradiaria del tiempo promedio de desplazamiento hacia el comercio (to_vendor_time) para cada día de la semana. Sobre el mismo gráfico se incorpora el volumen promedio de órdenes por hora (barras grises), lo cual permite contextualizar la dinámica operativa a lo largo del día. Esta representación facilita la identificación de patrones de comportamiento tanto intradiarios como entre distintos días, aspectos que resultan clave para la caracterización estocástica de este componente del tiempo de servicio.

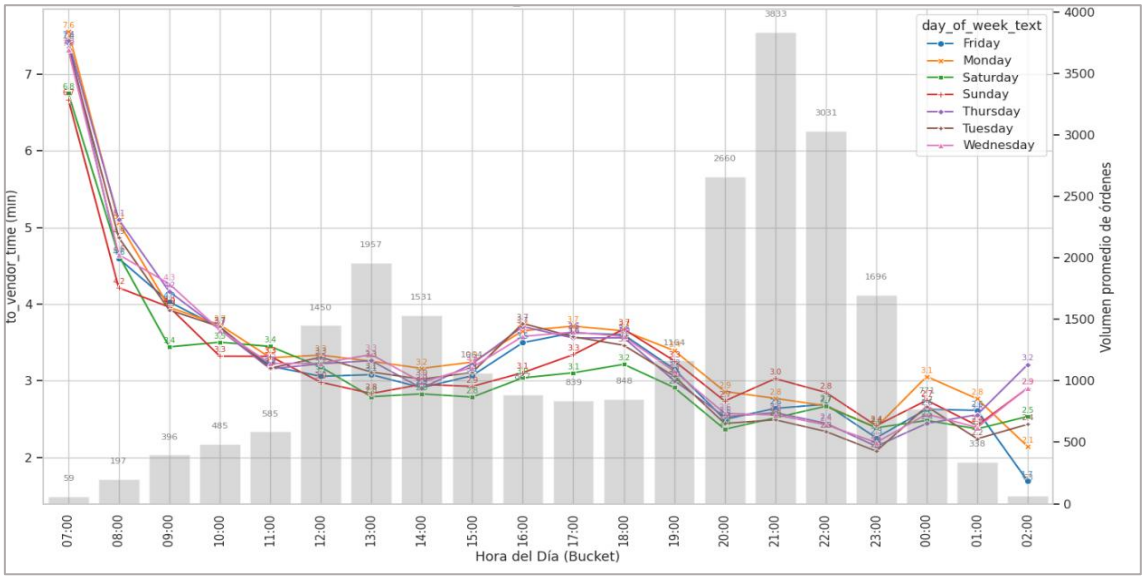

Figura 7 – Tiempo al loca (to_vendor_time)l promedio por hora, según día de la semana

El análisis revela patrones de variabilidad intradiaria y entre días de la semana que reflejan las dinámicas operativas del sistema de entrega a domicilio. Los principales hallazgos son:

- *Pico inicial elevado (07:00)*

  El tiempo promedio de traslado al comercio alcanza entre 6 y 7 minutos, coincidiendo con un bajo volumen de órdenes. Esta situación sugiere una fase de reactivación del

sistema tras la inactividad nocturna, caracterizada por menores niveles de flota disponible y mayores tiempos de asignación.

- *Descenso progresivo y estabilización (08:00–14:00)*

  A medida que se incrementa la densidad operativa, el tiempo al local desciende y se estabiliza en torno a 3 minutos, incluso en presencia de un volumen creciente de pedidos. Este comportamiento indica una etapa de equilibrio entre oferta y demanda dentro de la red de repartidores.

- *Repunte vespertino (16:00–18:00)*

  Se registra un aumento moderado del tiempo al local (hasta aproximadamente 3.5 minutos), coincidente con el inicio del pico de la cena. Esto sugiere una mayor presión sobre la operación, posiblemente relacionada con la congestión urbana o con una utilización más intensiva de la flota.

- *Descenso nocturno (18:00–23:00)*

  A medida que progresa la operación nocturna (incluyendo el pico y el posterior descenso de demanda), el tiempo de traslado vuelve a descender progresivamente, alcanzando mínimos cercanos a los 2.5 minutos hacia el final del día.

- *Leve repunte en la madrugada (00:00–02:00)*

  A pesar de la baja demanda, se observa un ligero incremento del tiempo al local, que podría explicarse por una menor disponibilidad relativa de repartidores activos en esas franjas horarias.

- Diferencias entre días de la semana:
  - Sábados y domingos: Exhiben los menores tiempos de traslado en casi todas las franjas horarias, sugiriendo una mayor eficiencia o menores restricciones operativas durante el fin de semana.
  - Lunes y martes: Presentan tiempos más elevados, especialmente en las primeras horas de la mañana y durante la tarde, reflejando una menor fluidez operativa respecto al resto de los días.
  - Jueves y viernes: Muestran comportamientos intermedios, con mayor estabilidad en el cierre de jornada.

Estos patrones evidencian que el tiempo hasta el local no presenta un comportamiento estacionario, sino que responde a condiciones operativas dinámicas y cambiantes.

2.2.4.2 Análisis de dispersión

La Figura 8 presenta como se comporta empíricamente la dispersión del tiempo tiempo al local por hora del día utilizando diagramas de caja. El objetivo es explorar no solo el comportamiento

promedio ya observado, sino también la variabilidad interna y la presencia de eventos extremos a lo largo de la jornada.

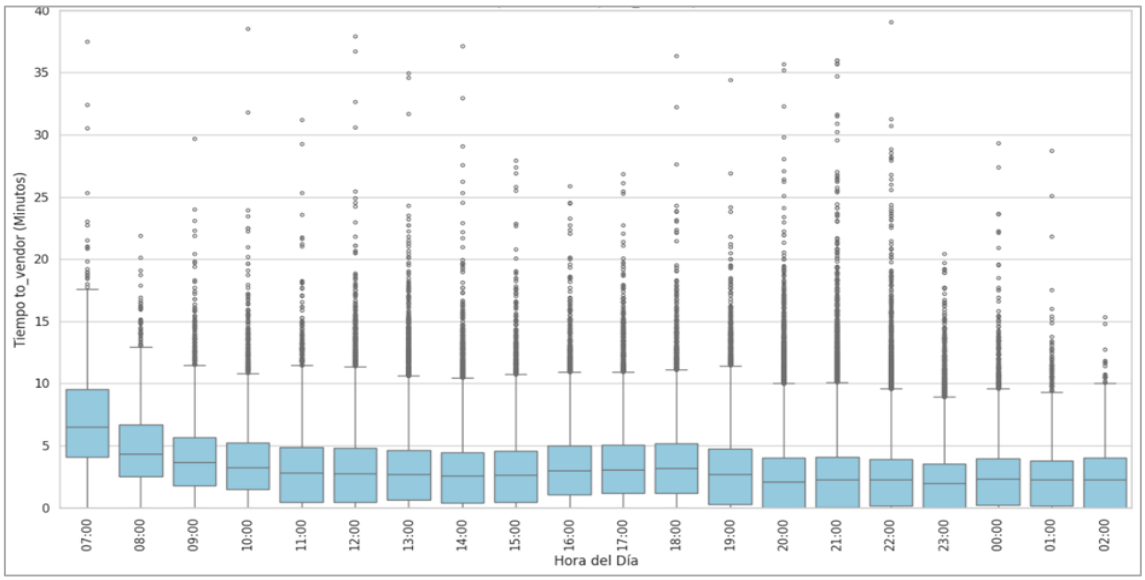

Figura 8 - Variabilidad del tiempo al local a lo largo del día (boxplot por hora)

De la lectura del gráfico se recopilan los siguientes puntos:

- *Mayor dispersión en las primeras horas operativas (07:00–09:00):*
  La mediana es más elevada (~5 minutos), pero también se observa una amplitud intercuartílica mayor (mayor separación entre Q1 y Q3). Esto indica una operación más heterogénea en el arranque del día, posiblemente vinculada a la puesta en marcha de los procesos operativos y la baja densidad inicial de pedidos y repartidores.
- *Reducción progresiva de la dispersión hacia el mediodía (10:00–14:00):*
  La mediana baja a valores cercanos a 3–4 minutos, y las cajas se vuelven más compactas, lo que sugiere una mayor homogeneización operativa en los momentos de mayor flujo de pedidos, donde tanto la disponibilidad de repartidores como la dinámica de los comercios son más estables.
- *Leve repunte en la dispersión en el turno vespertino (16:00–20:00):*
  Se evidencia un pequeño ensanchamiento en la altura de las cajas y un aumento en el rango de valores atípicos. Este fenómeno podría estar relacionado con picos de tráfico, cambios de turnos de repartidores o variaciones en el ritmo operativo de los comercios.
- *Concentración y estabilidad relativa en horario nocturno (21:00–00:00):*
  Las cajas se reducen en altura nuevamente, indicando una menor dispersión. El sistema parece operar de manera más eficiente y homogénea durante las horas de cena y primeras horas de la noche.

- *Incremento en la dispersión durante la madrugada (00:00–02:00):*
  Aunque la mediana permanece baja (~3 minutos), se observa una mayor dispersión y un número importante de valores típicos. Esto podría reflejar la escasa densidad de pedidos y repartidores en esos horarios, donde el azar en la asignación de órdenes genera mayor variabilidad.
- *Presencia de valores típicos en todos los horarios:*
  En todas las franjas horarias se identifican órdenes cuyo tiempo hasta el local es excepcionalmente alto (superior a 20–25 minutos). Esto es completamente esperable en un contexto de entrega a domicilio y pueden deberse a incidentes como cancelaciones de órdenes, problemas de localización de comercios, inconvenientes en el tráfico entre otros.

El análisis evidencia un patrón intradiario relativamente estable durante las horas de mayor actividad, acompañado de niveles moderados de dispersión, en especial en franjas de baja demanda. Si bien esta variabilidad no parece extrema, la presencia de valores atípicos y fluctuaciones puntuales sugiere que este componente podría contribuir de manera no despreciable a la variabilidad total del tiempo de servicio.

Estos resultados refuerzan la necesidad de modelar el tiempo al local como un proceso estocástico y justifican avanzar en el análisis de los siguientes componentes, en particular el tiempo en el local (at_vendor_time), con el objetivo de comprender de manera más completa las fuentes de variabilidad en el sistema.

2.2.5 Tiempo en el local

2.2.5.1 Análisis de patrones y tendencias

La Figura 9 presenta la evolución intradiaria del tiempo promedio de espera en el restaurante o comercio (at_vendor_time) para cada día de la semana, junto con el volumen promedio de órdenes por bucket horario (barras grises). Este gráfico permite observar cómo varía el tiempo de espera en los comercios a lo largo del día y su interacción con los patrones de demanda.

.

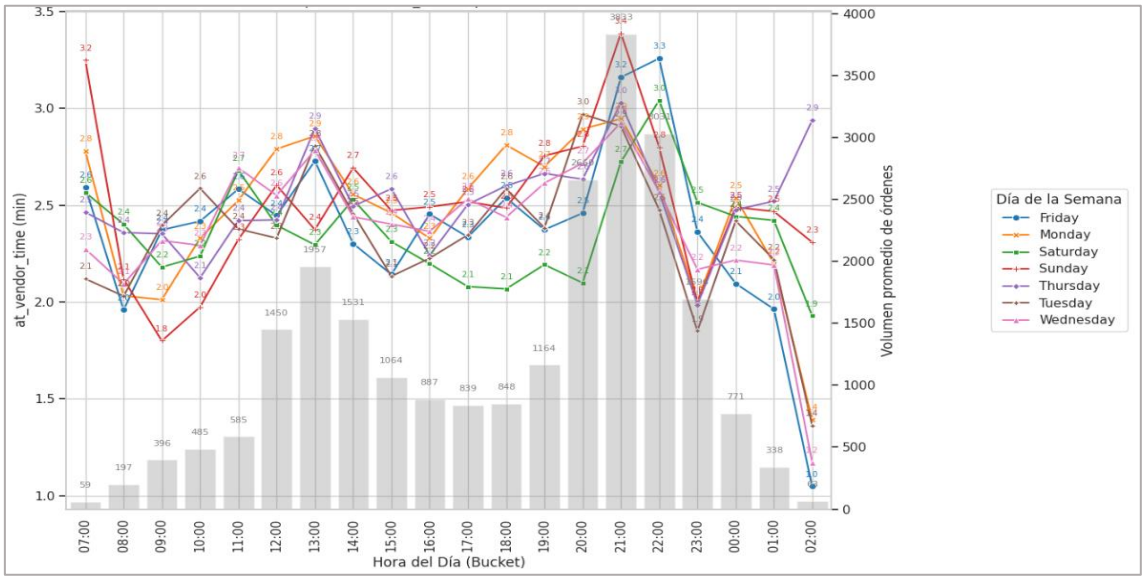

Figura 9 - 2.2.5 Tiempo en el local promedio por hora, según día de la semana

Principales patrones observados:

- *Estabilidad relativa durante la jornada diurna:*

  Entre las 10:00 y las 18:00 horas, el tiempo en el local se mantiene en valores relativamente constantes (≈2.1–2.8 minutos), con baja dispersión entre días. Esta estabilidad sugiere un funcionamiento regular de los restaurantes y comercios en las horas de mayor actividad operativa.

- *Incremento nocturno asociado a picos de demanda:*

  A partir de las 20:00 horas, coincidiendo con el pico de órdenes, el tiempo de espera tiende a incrementarse, alcanzando valores de hasta 3.3 minutos en jornadas de alta demanda como viernes y domingos. Este comportamiento refleja la presión que la concentración de pedidos ejerce sobre la capacidad de procesamiento de los restaurantes y comercios.

- *Descenso hacia la madrugada:*

  Tras el cierre del pico nocturno, el tiempo en el local disminuye de forma progresiva entre las 23:00 y las 02:00, en línea con la reducción del volumen de órdenes y una menor saturación de los comercios.

- Variabilidad entre días de la semana:
  - Viernes y domingos presentan incrementos más marcados en el horario nocturno.
  - Los sábados exhiben un comportamiento más estable y tiempos de espera más bajos durante toda la jornada.

- Lunes y martes muestran leves elevaciones en las franjas de alta demanda.

Estas observaciones permiten identificar variaciones sistemáticas en el comportamiento del tiempo en el local, que deberán ser consideradas en el modelado posterior del sistema.

2.2.5.2 Análisis de dispersión

La Figura 10 muestra la distribución empírica del tiempo en el local a lo largo del día mediante diagramas de caja, permitiendo analizar la variabilidad y la presencia de valores atípicos en cada franja horaria.

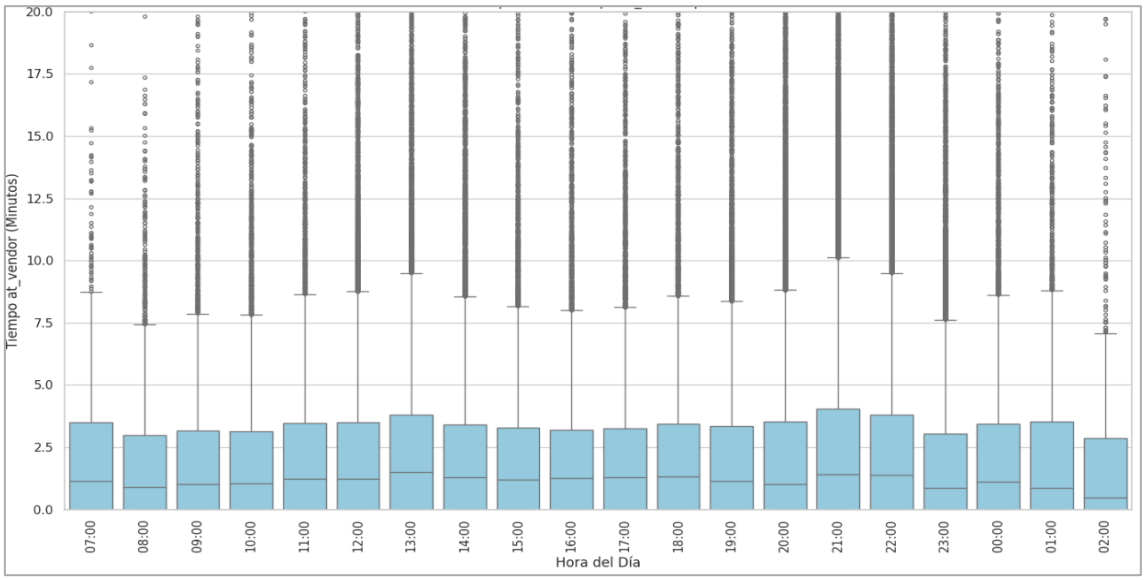

Figura 10 - Variabilidad del tiempo en el local a lo largo del día (boxplot por hora)

Las observaciones más relevantes son:

- *Dispersión moderada y relativamente homogénea*

  El rango intercuartílico (IQR) se mantiene relativamente constante entre horarios, oscilando entre 1 y 2 minutos, lo que sugiere una dispersión moderada en el núcleo de la distribución.

- *Presencia significativa de valores típicos*

  En todas las franjas horarias se observan valores atípicos que superan ampliamente los valores del percentil 75. Esto indica la existencia de eventos menos frecuentes, como demoras excepcionales en los locales, que pueden impactar de manera relevante en algunos pedidos individuales.

- *Asimetría en la distribución*

  En la mayoría de los horarios, la mediana es inferior al promedio observado en los gráficos anteriores. Esta diferencia evidencia una asimetría hacia la derecha, donde

existen casos esporádicos de tiempos de espera considerablemente mayores que la media operativa.

- *Incremento de la dispersión en picos de demanda*

  Especialmente en torno a las 20:00 y 21:00, coincidiendo con los máximos de volumen de órdenes, se observa un leve ensanchamiento del IQR y una mayor densidad de valores típicos, lo que sugiere que la presión operativa sobre los locales introduce una mayor variabilidad en el proceso de despacho.

Este patrón de comportamiento destaca que, aunque el tiempo en el local suele ser relativamente estable para la mayoría de los pedidos, existen escenarios específicos donde puede desviarse significativamente, representando una fuente no despreciable de incertidumbre en el ciclo total de servicio.

2.2.6 Tiempo hasta el domicilio del cliente

2.2.6.1 Análisis de patrones y tendencias

La Figura 11 presenta el tiempo promedio de desplazamiento hasta la ubicación del cliente (to_customer_time) a lo largo del día, segmentado por día de la semana. El gráfico incluye también el volumen promedio de órdenes por hora (barras grises), permitiendo analizar el comportamiento operativo en relación con la evolución de la demanda.

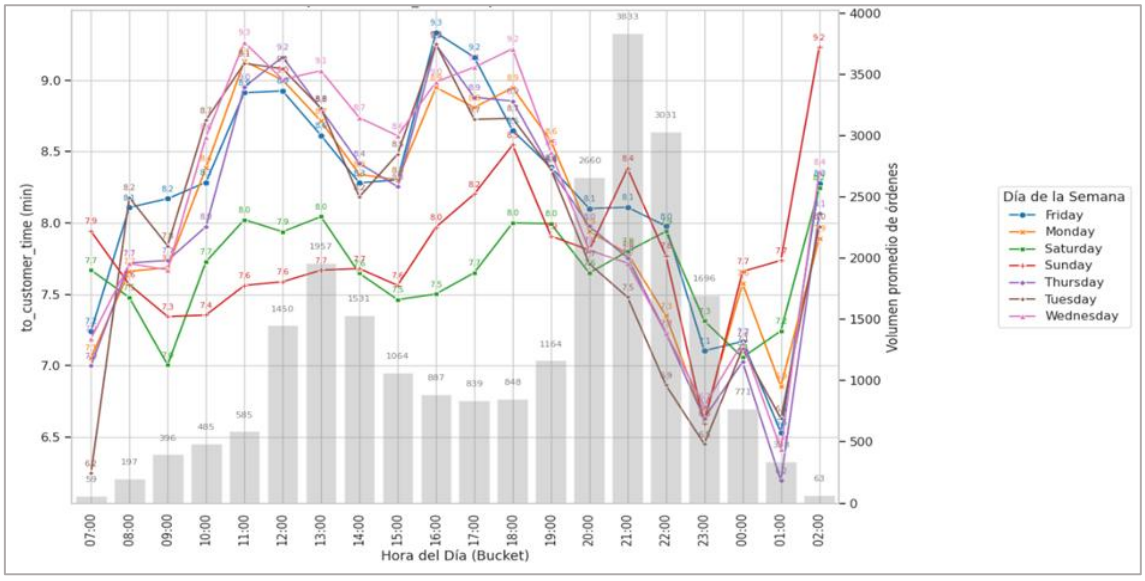

Figura 11 – Tiempo hasta el domicilio del cliente promedio por hora, según día de la semana

El análisis revela un patrón intradiario caracterizado por mesetas, picos moderados y descensos nocturnos, en línea con la evolución general del volumen de órdenes.

- *Franja vespertina (16:00–18:00):*

Durante este tramo se observa un aumento del tiempo, alcanzando valores entre 8.5 y 9.3 minutos, a pesar de tratarse de un periodo de menor volumen de órdenes. Este comportamiento podría estar asociado a una menor densidad de repartidores, lo que incrementa las distancias promedio, así como a una mayor dispersión en la naturaleza de los pedidos.

- *Meseta diurna tardía (11:00–15:00):*

  En las horas centrales del día, se mantiene relativamente estable entre 8.3 y 8.8 minutos. Esta estabilidad sugiere un equilibrio operativo entre la oferta de repartidores y la concentración de la demanda vinculada al almuerzo.

- *Descenso nocturno (20:00–23:00):*

  Conforme disminuye el volumen de órdenes tras el pico de cena, el tiempo de desplazamiento hasta el cliente desciende progresivamente hasta alcanzar valores mínimos cercanos a 6.5–7.0 minutos, reflejando rutas más directas y menor congestión operativa.

- *Repunte de madrugada (00:00–02:00):*

  En la madrugada se observa un nuevo incremento, alcanzando entre 7.5 y 8.5 minutos, en un contexto de muy baja densidad de órdenes. La escasa disponibilidad de repartidores y la dispersión geográfica explican este comportamiento.

- *Diferencias entre días de la semana:*

  Aunque el patrón general se mantiene, miércoles y jueves presentan tiempos más elevados durante la tarde, mientras que el sábado sostiene niveles más bajos y homogéneos. El domingo, en cambio, muestra un repunte nocturno más pronunciado.

En conjunto, los resultados reflejan que el tiempo hasta el domicilio del cliente exhibe una variabilidad significativa en función del horario y del día de la semana, lo que sugiere la necesidad de incorporar esta variabilidad de forma explícita en la modelización del sistema de entregas.

2.2.6.2 Análisis de dispersión

La Figura 12 muestra la dispersión del tiempo hasta el cliente por hora del día, permitiendo identificar tanto la tendencia central como la variabilidad de este componente a lo largo de la jornada.

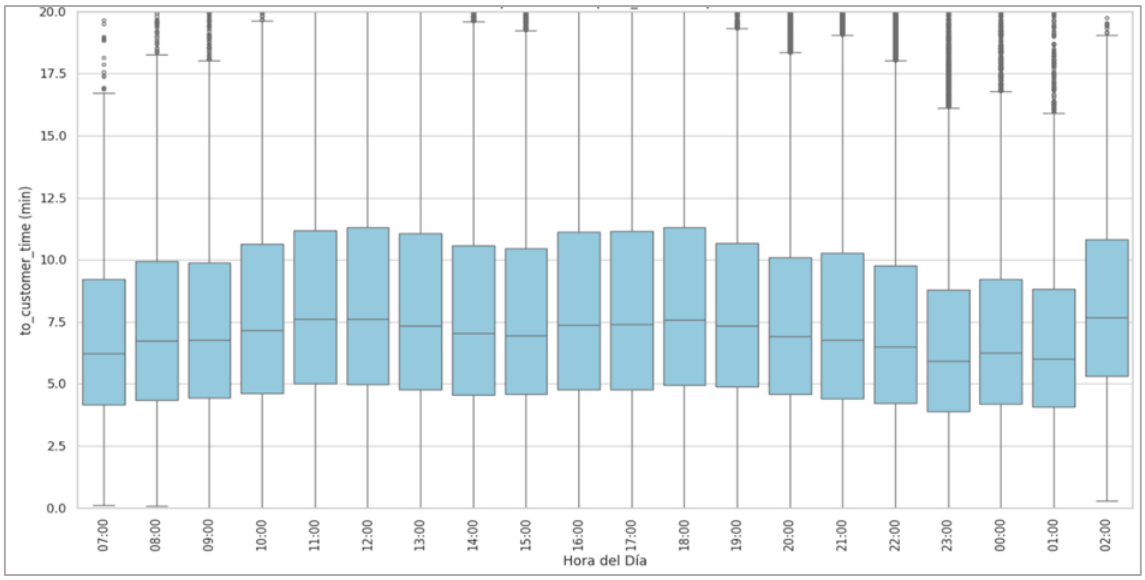

Figura 12 - Variabilidad del tiempo hasta el domicilio del cliente a lo largo del día (boxplot por hora)

Los principales hallazgos que surgen del análisis son los siguientes:

- *Incremento progresivo durante la mañana:*

  La mediana del tiempo al cliente aumenta desde aproximadamente 6 minutos a las 07:00 hasta valores cercanos a 8–8.5 minutos entre las 11:00 y 13:00. Esta tendencia refleja probablemente el impacto combinado del incremento de tráfico urbano y la reorganización de rutas tras el primer pico operativo.
- *Dispersión acentuada en horas de baja demanda:*

  Durante las franjas de 07:00–09:00 y 16:00–18:00, donde el volumen de órdenes desciende, se observa un ensanchamiento del rango intercuartílico. Esta mayor dispersión sugiere una menor densidad de repartidores activos y mayor heterogeneidad en la configuración de las rutas de entrega.
- *Estabilización nocturna:*

  A partir de las 23:00, la mediana desciende progresivamente, alcanzando mínimos cercanos a 5 minutos hacia las 02:00. Esto refleja un entorno operativo con menor congestión vehicular y rutas más directas, a pesar de la presencia de algunos valores atípicos.
- *Valores típicos persistentes en toda la jornada:*

  Se detectan tiempos de entrega superiores a 20 minutos a lo largo de todas las horas analizadas, aunque con mayor frecuencia relativa en los períodos de menor actividad. Estos valores extremos podrían asociarse a rutas particularmente largas o a situaciones operativas excepcionales.
- *Compresión en picos de alta actividad:*

Durante los horarios de máxima demanda (12:00–14:00 y 20:00–22:00), el rango intercuartílico se reduce de manera significativa, indicando un comportamiento operativo más predecible cuando la flota está plenamente desplegada.

Del análisis se evidencia una dinámica intradiaria donde tanto la mediana como la variabilidad fluctúan en función de los patrones de demanda y la densidad operativa. Estas observaciones refuerzan la necesidad de contemplar no solo la media sino también la dispersión en la modelización del proceso, para capturar adecuadamente los efectos de los diferentes escenarios horarios sobre la performance del sistema.

2.2.7 Tiempo en el domicilio del cliente

2.2.7.1 Análisis de patrones y tendencias

La Figura 13 presenta el tiempo promedio de espera en destino (at_customer_time) a lo largo del día, segmentado por día de la semana. Al igual que en los componentes anteriores, se incluye también el volumen promedio de órdenes por bucket horario (barras grises) para brindar contexto sobre la dinámica de la demanda.

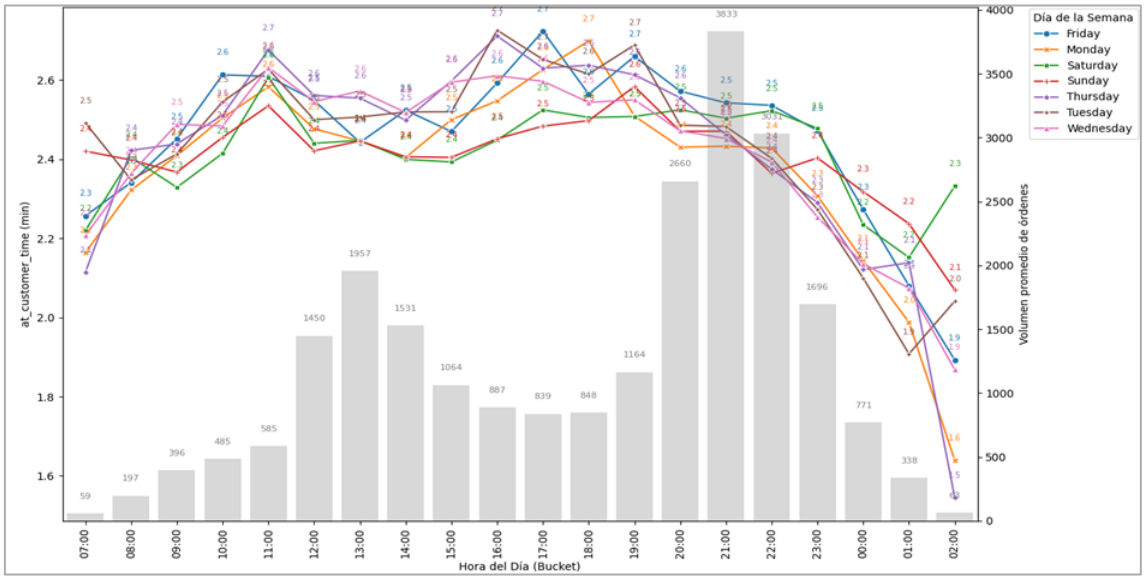

Figura 13 – Tiempo en el domicilio del cliente promedio por hora, según día de la semana

Patrones observados:

- *Picos matinales previos al almuerzo (10:00–11:00):*

  Se registra un aumento hasta valores cercanos a 2.7 minutos. Esta demora podría estar asociada a una mayor actividad urbana y laboral, donde los clientes, aún inmersos en sus rutinas matutinas, demoran más en acudir a recibir el pedido.
- *Segundo ascenso vespertino (16:00–19:00):*

Luego de una relativa estabilidad durante el mediodía, se observa un nuevo incremento en el tiempo de espera hacia el final de la tarde. Aunque el volumen de órdenes no alcanza su máximo, la menor disponibilidad inmediata de los clientes en horarios laborales podría explicar este patrón.

- *Descensos en horarios de consumo principal (12:00–13:00 y 20:00–21:00):*
  En las franjas típicas de almuerzo y cena, el tiempo disminuye levemente. Es probable que durante estos momentos los clientes estén más atentos y predispuestos a recibir sus pedidos de manera ágil.
- *Reducción progresiva en horario nocturno (22:00–02:00):*
  Luego de la cena, se evidencia una disminución continua de la demora en el destino, alcanzando valores mínimos hacia la madrugada, en un contexto de menor volumen de órdenes y mayor disponibilidad de los clientes para recibir sus entregas.
- *Consistencia semanal:*
  Aunque se observan ligeras diferencias entre días de semana y fines de semana, la estructura general del patrón se mantiene, lo que sugiere una dinámica relativamente estable en el comportamiento de los clientes.

El análisis muestra que el tiempo en el cual el repartidor permanece en el domicilio del cliente no depende únicamente del volumen operativo, sino que está fuertemente condicionado por la disponibilidad y el comportamiento de los clientes en distintos momentos del día. Esta evidencia refuerza la importancia de modelar este componente considerando su variabilidad intradiaria y su vínculo con los patrones de consumo.

2.2.7.2 Análisis de dispersión

La Figura 14 presenta la distribución empírica del tiempo de espera en el punto de entrega, desagregada por hora del día. Se utilizó un diagrama de caja para representar la mediana, el rango intercuartílico y la presencia de valores atípicos, facilitando así la detección de patrones de variabilidad intradiaria.

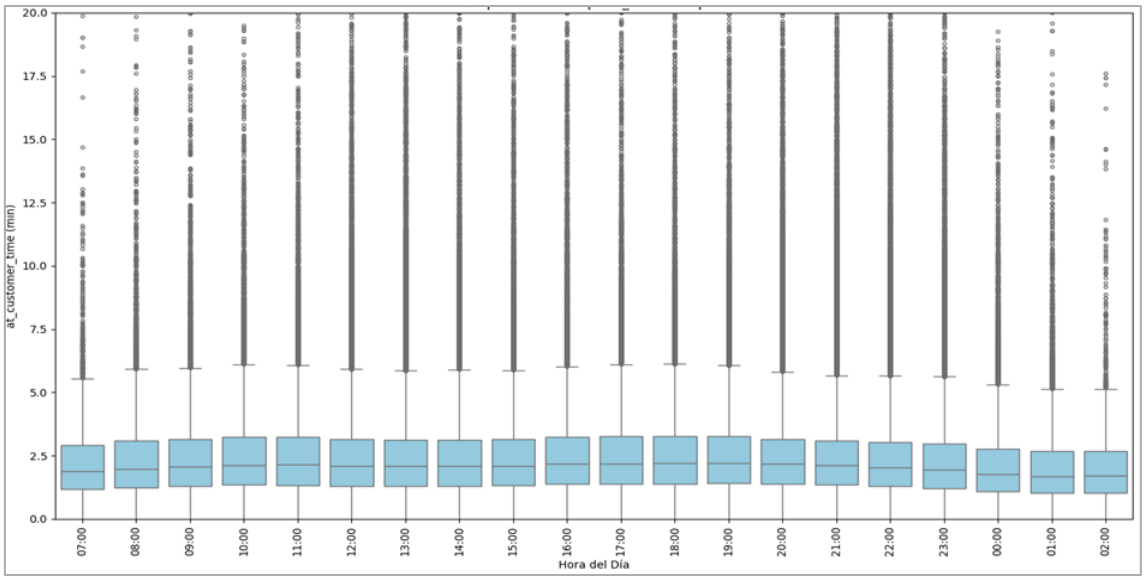

Figura 14 - Variabilidad del – Tiempo en el domicilio del cliente a lo largo del día (boxplot por hora)

Los principales hallazgos que se desprenden del análisis son:

- *Mediana y rango intercuartílico*

  En la mayoría de las horas pico (12:00–14:00 y 20:00–22:00), la mediana se ubica entre 1 y 2 minutos, con un rango intercuartílico relativamente acotado (≈1.3–2.7 minutos). Esto refleja una mayor homogeneidad en los tiempos de entrega, posiblemente debido a una mayor predisposición de los clientes a recibir su pedido en horarios de comida. Por el contrario, durante las franjas intermedias de menor demanda (especialmente entre 16:00 y 19:00), tanto la mediana como el rango intercuartílico se incrementan ligeramente. Esta tendencia podría explicarse por una menor disponibilidad inmediata de los clientes, que en esos horarios suelen estar trabajando o realizando otras actividades, demorando la recepción del pedido.
- Patrón en madrugada y primeras horas del día

  En las primeras horas de la mañana (07:00–09:00) y durante la madrugada (00:00–02:00), se observa un ensanchamiento de las distribuciones y la presencia de múltiples valores atípicos. Esto podría deberse a factores como demoras en que los clientes respondan a la llegada del pedido o entregas en zonas de baja densidad de actividad.
- Valores típicos consistentes a lo largo del día

  En todas las franjas horarias se detectan valores extremos superiores a los 15 minutos. Aunque representan una minoría de los casos, estos eventos contribuyen significativamente a la dispersión total del tiempo de espera en destino, y suponen situaciones particulares que deben ser contempladas al modelar el proceso completo.

El análisis de dispersión pone en evidencia que, si bien el tiempo de espera en destino suele ser breve y relativamente consistente en horarios de almuerzo y cena, existen variaciones no despreciables en franjas laborales y en horarios marginales. Además, la presencia sistemática de valores típicos refuerza la necesidad de modelar explícitamente la variabilidad en la etapa de entrega final.

3 Metodología

3.1 Teoría de Colas

Tal como se introdujo en la primera sección, el marco teórico adoptado en este trabajo es el de la teoría de colas, una rama de la investigación operativa que estudia sistemas donde diferentes entidades (clientes, tareas, órdenes, etc.) esperan en fila para ser atendidas por uno o más servidores. Los contenidos desarrollados en esta sección, incluyendo la definición de variables clave, tipos de modelos y métricas de desempeño, se basan principalmente en la obra de Gross, Shortle, Thompson y Harris (2008), uno de los textos de referencia más reconocidos en la materia.

Un modelo de colas puede entenderse en términos generales como un sistema compuesto por tres elementos principales: (1) proceso de llegada (input), (2) estructura de cola y (3) mecanismo de atención (servidores). La dinámica y desempeño del sistema dependen fundamentalmente de la interacción entre estas partes, así como de las características particulares del sistema estudiado.

3.1.1 Definición y variables clave

Para definir formalmente un modelo de colas, es necesario caracterizar ciertas variables esenciales:

- **Tasa de arribo (λ):** es el número promedio de entidades (clientes, pedidos, etc.) que llegan al sistema por unidad de tiempo. Generalmente, se asume que los arribos siguen una distribución de probabilidad específica (habitualmente Poisson), aunque esto dependerá del sistema analizado.
- **Tasa de servicio (μ):** representa el número promedio de entidades atendidas por un servidor por unidad de tiempo. El tiempo de servicio puede seguir diferentes distribuciones de probabilidad (exponencial, Gamma, normal, etc.). La elección de la distribución influye considerablemente en el comportamiento y desempeño del modelo
- **Número de servidores (c):** indica cuántas unidades de atención operan simultáneamente en el sistema. Un mayor número de servidores típicamente reduce el tiempo de espera, pero incrementa los costos operativos asociados.
- **Disciplina de la cola:** define la regla bajo la cual se atienden los clientes en espera. La más común es FIFO (first-in, first-out), es decir, el primer cliente en llegar es el primero en ser atendido. Existen otras disciplinas como LIFO (last-in, first-out), o prioridades según características particulares del cliente o pedido.

Además, para evaluar el desempeño del sistema se utilizan generalmente algunas métricas clave, tales como:

- **Tiempo promedio en la cola (Wq):** tiempo promedio que una entidad pasa esperando a ser atendida.
- **Tiempo promedio en el sistema (W):** suma del tiempo promedio en espera más el tiempo promedio de atención.
- **Número promedio de entidades en cola (Lq)**
- **Número promedio de entidades en el sistema (L)**
- **Utilización del servidor (ρ o UTR):** proporción del tiempo en que los servidores están ocupados.

3.1.2 Tipos de modelos relevantes

Existen diferentes categorías de modelos de colas según la distribución del tiempo de llegada y el tiempo de servicio. Los más comunes se identifican mediante la notación de Kendall, que utiliza tres símbolos separados por barras (A/B/c), donde:

- A: distribución del tiempo de llegada (generalmente M para Poisson o exponencial).
- B: distribución del tiempo de servicio (generalmente M para exponencial, G para distribución general).
- c: número de servidores.

Los modelos más utilizados incluyen:

- Modelo M/M/1: un servidor único con tiempos de llegada y servicio exponenciales.
- Modelo M/M/c: múltiples servidores, tiempos de llegada y servicio exponenciales.
- Modelo M/G/1: un servidor único, tiempo de llegada exponencial y servicio con distribución general.
- Modelo M/G/c: múltiples servidores, tiempo de llegada exponencial y servicio con distribución general.

En este trabajo, se adopta el modelo M/G/c, dado que las observaciones empíricas del sistema de entrega a domicilio evidencian que el tiempo de servicio no sigue una distribución exponencial, sino que se ajusta mejor a una distribución Gamma. Esta elección del modelo permite capturar mejor la variabilidad observada en los tiempos de servicio reales, manteniendo un equilibrio razonable entre complejidad y precisión.

Este marco teórico proporciona las bases necesarias para modelar la operación de un sistema de entrega a domicilio bajo un enfoque de teoría de colas, permitiendo así analizar el impacto de diferentes asignaciones de recursos (número de repartidores) sobre la calidad del servicio (tiempos de espera y utilización del sistema).

3.1.3 El modelo de colas aplicado al sistema de entrega a domicilio

Dado el marco conceptual adoptado en este trabajo para modelar el funcionamiento del sistema estudiado, se propone representar la operación de un servicio de entrega a domicilio bajo demanda como un sistema de colas discreto en el tiempo, donde las entidades corresponden a las órdenes generadas por los usuarios de la plataforma que arriban al sistema de despacho, y los repartidores representan los servidores que las atienden luego de ser asignados y aceptarlas

Este enfoque permite estudiar el comportamiento del sistema bajo distintas configuraciones de flota, capturando tanto métricas de rendimiento (como el tiempo de espera) como métricas de eficiencia (como la tasa de utilización del recurso). Dado que la asignación de un repartidor a una orden no ocurre de forma instantánea, y que los tiempos de servicio presentan alta variabilidad, se opta por un modelo que incorpora esas fuentes de aleatoriedad a través de una simulación basada en eventos discretos.

El proceso operativo se estructura en tres momentos principales: (i) creación de la orden, (ii) asignación del repartidor, y (iii) finalización de la entrega. A partir de esta segmentación, se define el tiempo de espera como la diferencia entre los momentos (ii) y (i), el tiempo de servicio como el lapso desde la asignación hasta la entrega, y el tiempo de ciclo como el total desde la creación hasta la entrega.

El sistema presenta una capacidad finita de servidores disponibles en cada bucket de tiempo (intervalo de 15 minutos), y se modela bajo un esquema FIFO (first-in-first-out) para la asignación de órdenes. Asimismo, se contempla la posibilidad de apilado de órdenes, es decir, que un repartidor entregue múltiples pedidos en un mismo viaje, aunque en esta versión del modelo su efecto ha sido excluido deliberadamente para simplificar la estructura de la simulación inicial.

Cabe destacar que este sistema no opera bajo un régimen estacionario clásico, ya que tanto la llegada de órdenes como la disponibilidad de repartidores varían de forma significativa a lo largo del día. Por ello, el modelo se implementa con una granularidad temporal elevada (por bucket) y se estudia el comportamiento del sistema de forma desagregada para cada intervalo.

3.1.3.1 Ciclo de vida de una orden

El sistema de entrega a domicilio bajo análisis puede ser entendido como un proceso dinámico en el cual las órdenes atraviesan distintas etapas desde su creación hasta su entrega al cliente final. Esta secuencia de eventos configura lo que, en el marco de la teoría de colas, se denomina el ciclo de vida de una orden, y permite identificar con precisión los momentos en los cuales se definen métricas operativas clave.

Como fue mencionado, en un abordaje simplificado, se pueden distinguir tres momentos fundamentales en este ciclo: (i) la creación de la orden, que marca el inicio del proceso y representa el ingreso al sistema; (ii) la aceptación de la orden por parte de un repartidor, que equivale al inicio del servicio y determina el fin del tiempo de espera en cola; y (iii) la entrega efectiva de la orden, que constituye el evento de egreso y cierre del ciclo. Esta secuencia se representa en la Figura 15, que sintetiza las principales transiciones del sistema.

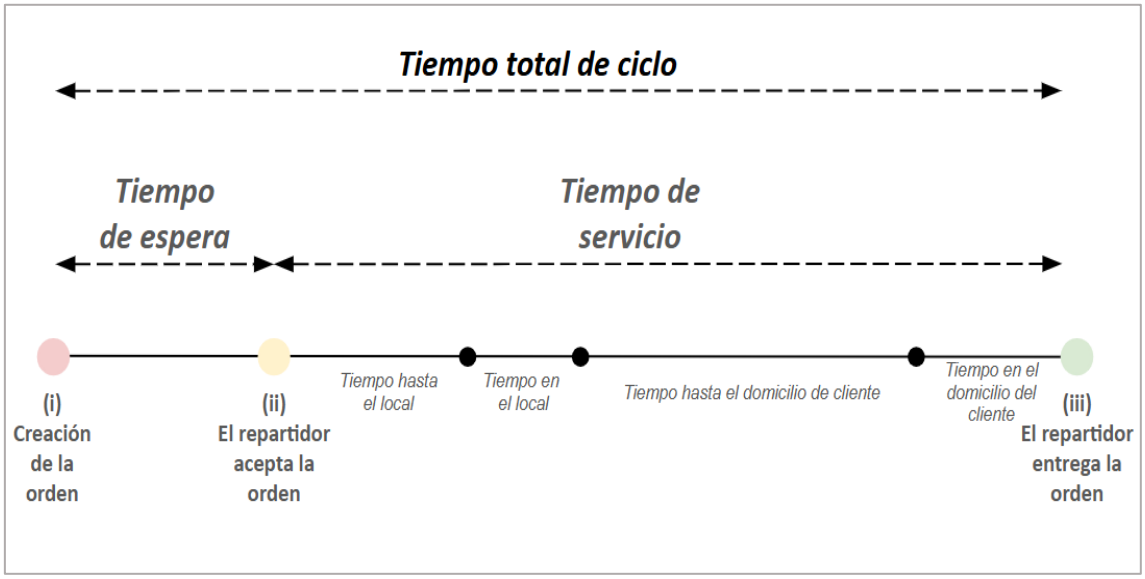

Figura 15 Ciclo de vida de una orden: representación de eventos y métricas temporales

A partir de esta conceptualización es posible definir las siguientes métricas de desempeño asociadas a cada etapa:

- **Tiempo de espera:** intervalo transcurrido entre la creación de la orden y su aceptación por parte de un repartidor. En términos de la teoría de colas, representa el tiempo de permanencia en la cola.
- **Tiempo de servicio:** intervalo entre la aceptación de la orden y su entrega al cliente. Incluye tanto los desplazamientos necesarios para completar la orden (hacia el local comercial y luego hacia el domicilio del cliente) y como los tiempos de espera correspondientes

- **Tiempo de ciclo:** tiempo total desde la creación hasta la entrega. Resulta de la suma del tiempo de espera y el tiempo de servicio.
- **Rendimiento:** cantidad de órdenes completadas por unidad de tiempo. Se calcula en función de las salidas del sistema durante un período determinado.
- **Utilización del sistema (ρ):** razón entre el tiempo total de trabajo activo de los servidores (repartidores) y el tiempo total disponible. Se utiliza como proxy de la eficiencia operativa.

Este esquema permite modelar el proceso de asignación y entrega como un sistema de colas de múltiples servidores, en el cual las órdenes ingresan de forma estocástica, esperan su turno para ser asignadas, y luego son despachadas por un servidor hacia su destino final. Las simplificaciones adoptadas para operacionalizar este modelo se detallan en la sección siguiente.

3.1.3.2 Supuestos y simplificaciones del modelo

Para modelar el sistema de asignación y atención de órdenes en un entorno de entrega a domicilio mediante teoría de colas, fue necesario adoptar ciertos supuestos que permitieran reducir la complejidad del sistema real y facilitar su análisis. A continuación, se enumeran y justifican las principales simplificaciones incorporadas:

- **Repartidores homogéneos en términos de desempeño:**

  Se asume que todos los servidores presentan un comportamiento homogéneo en relación con su capacidad para atender órdenes, es decir, comparten una misma distribución de tiempo de servicio. Esta simplificación evita modelar variabilidad entre individuos —como diferencias de velocidad, experiencia o familiaridad con la zona— que, si bien relevantes en la práctica, complejizan significativamente la estructura del modelo. Bajo esta hipótesis, todos los servidores son estadísticamente equivalentes, lo cual permite utilizar una única función de distribución para representar el tiempo de servicio en todo el sistema.
- **Asignación inmediata bajo lógica FIFO (first-in, first-out):**

  El modelo asume que, una vez creada, cada orden es asignada de manera inmediata al repartidor disponible más cercano, siguiendo una lógica de tipo FIFO. Esta simplificación omite ciertos mecanismos clave del algoritmo real de asignación, como el tiempo de retención que se aplica a cada orden con el objetivo de sincronizar el momento de asignación con el tiempo necesario de preparación en cocina. En la operación real, el sistema busca minimizar el tiempo de espera total evitando que el repartidor llegue antes de que el pedido esté listo. Sin embargo, modelar esta lógica requeriría incorporar estimaciones de tiempo de preparación y ventanas dinámicas de asignación, lo que

excede el alcance del presente análisis. Por ello, se adopta esta versión simplificada, que facilita la construcción del modelo sin necesidad de simular decisiones algorítmicas complejas ni comportamientos coordinados entre múltiples actores.

- **Distribución Gamma para los tiempos de servicio:**

  Los tiempos de servicio se modelan como variables aleatorias que siguen una distribución Gamma, de parámetros ajustados empíricamente a partir de datos históricos. Esta elección responde a la flexibilidad de dicha distribución para capturar la asimetría y dispersión observadas en los datos reales, características que no pueden ser adecuadamente representadas por una distribución exponencial. Asimismo, la elección de una distribución continua con soporte positivo resulta coherente con la naturaleza de los tiempos de atención.

- **Capacidad infinita en cola:**

  Se considera que la cola de espera para la asignación de órdenes no tiene un límite superior de capacidad, es decir, que todas las órdenes que arriban pueden ingresar al sistema sin ser rechazadas. Esta suposición permite modelar el sistema como una cola de tipo M/G/s con capacidad ilimitada, lo cual es razonable para una plataforma digital en la que las órdenes no son descartadas inmediatamente ante una saturación operativa, sino que experimentan una demora mayor antes de ser asignadas.

- **Apilado de órdenes limitado o simplificado:**

  La asignación de múltiples órdenes simultáneas a un mismo repartidor es un fenómeno presente en operaciones reales. No obstante, su incorporación completa implicaría un modelo sustancialmente más complejo, dado que los servidores dejarían de atender a una única orden por vez. En este análisis se adopta una versión simplificada (no más de dos órdenes por repartidor), representada de manera indirecta y con restricciones predefinidas. Esta aproximación permite reflejar parcialmente los efectos del apilado de órdenes sin perder la trazabilidad del modelo base.

3.1.4 Diseño Experimental

A partir del modelo teórico previamente definido, esta sección describe el diseño del experimento implementado para evaluar distintas configuraciones de flota en un sistema de entrega a domicilio simulado. El objetivo es comparar el desempeño de dos formas de determinar la dotación de repartidores por intervalo de tiempo: una configuración basada en el enfoque actualmente utilizado en la planificación operativa (calculada heurísticamente a partir de una tasa objetivo de utilización), y otra obtenida mediante simulaciones sucesivas que permiten identificar la cantidad de repartidores que minimiza el tiempo promedio de espera de

las órdenes. Para ello, se simula la operación de la plataforma durante una jornada completa de viernes, reproduciendo las condiciones de llegada de órdenes y tiempos de servicio observados en datos reales. El número de repartidores disponibles en cada intervalo de 15 minutos (bucket) se establece de acuerdo con la configuración bajo análisis, y se evalúa el rendimiento del sistema a través del tiempo de espera, la utilización de recursos y el rendimiento obtenido (órdenes completadas en cada bucket).

A continuación, se presentan las variables involucradas en el ejercicio, distinguiendo su rol dentro del diseño experimental.

3.1.4.1 Variables y parámetros del experimento

Variables independientes

La principal variable de decisión es el número de repartidores asignados por bucket. Esta cantidad se define de dos formas en el experimento: una primera configuración basada en un cálculo heurístico determinado por el cociente entre órdenes estimadas y una tasa de utilización objetivo (UTR), y una segunda configuración optimizada utilizando simulaciones para minimizar el tiempo de espera efectivo.

Variables dependientes

Son aquellas métricas utilizadas para evaluar el desempeño del sistema bajo cada configuración. Incluyen:

- Tiempo de espera promedio: tiempo de espera promedio de las órdenes en cada bucket.
- Rendimiento: cantidad de órdenes completadas por unidad de tiempo.
- UTR (tasa de utilización): métrica operativa interna definida como el cociente entre la cantidad de órdenes completadas y el total de horas de trabajo de los repartidores conectados. Esta métrica se emplea también para calcular la flota heurística.

Variables de control

Son parámetros que se mantienen constantes a lo largo del experimento, incluyen:

- Tasa de arribo (tiempo entre llegadas): extraída empíricamente de los datos reales para el mismo día y franja horaria.
- Distribución del tiempo de servicio: se utiliza una distribución Gamma ajustada a partir de datos históricos para representar el tiempo de servicio.

Parámetros temporales del experimento

El análisis se enfoca en una jornada operativa representativa correspondiente al viernes, cubriendo el intervalo de 07:00 a 02:00 del día siguiente. La simulación se estructura en buckets de 15 minutos, resultando en 76 bloques temporales en total.

3.1.4.2 Fuentes y procesamiento de datos

El experimento se nutre de datos operativos históricos extraídos de la plataforma, correspondientes a un período de observación de trece semanas. La información utilizada ya fue descripta en detalle en la sección de Datos; en esta instancia se presentan únicamente los tratamientos específicos realizados para estructurar los insumos requeridos por el modelo.

Para caracterizar la demanda del sistema, se trabajó con una tabla consolidada de órdenes (tesis_base6.csv), desde la cual se calculan las tasas de arribo (tiempo entre llegadas) y se extraen muestras empíricas para componer los tiempos de servicio. Estos tiempos incluyen el desplazamiento hacia el comercio (to_vendor_time), la espera en el local (at_vendor_time), el trayecto al cliente (to_customer_time) y la espera al momento de la entrega (at_customer_time). El cálculo de estos valores se restringe a los días viernes y al rango horario comprendido entre las 07:00 y las 02:00 del día siguiente.

La llegada de órdenes se modela de forma empírica, utilizando la secuencia real de tiempos entre eventos registrada en los datos históricos, lo que permite capturar la variabilidad observada sin asumir una distribución teórica particular.

Adicionalmente, se utiliza una tabla auxiliar (bucketed_friday.csv) que consolida las órdenes promedio por bucket de 15 minutos, agrupadas por franja horaria, a partir de la cual se calcula la cantidad de repartidores requeridos en la configuración heurística. Esta cantidad se obtiene dividiendo la demanda estimada por una tasa de utilización esperada (UTR), previamente definida en el módulo de configuración.

Finalmente, se emplea el archivo utr_base1.csv como fuente de contraste externo para validar el comportamiento real del sistema, ya que contiene información agregada de órdenes efectivamente realizadas, UTR observados y dotación programada en fechas históricas.

Todos los valores fueron alineados en torno a una misma unidad temporal (buckets de 15 minutos) para asegurar la comparabilidad entre simulaciones, métricas empíricas y configuraciones previstas.

3.1.5 Implementación del Modelo

Con el objetivo de evaluar el desempeño del sistema bajo distintas configuraciones de dotación, se desarrolló una simulación estocástica que replica el funcionamiento de una operación real de

entrega a domicilio durante una jornada representativa. El modelo fue implementado en Python utilizando la librería SimPy, y se estructuró en módulos específicos para garantizar la claridad, la reutilización de componentes y la trazabilidad del experimento.

La lógica general de la simulación consiste en definir el número de repartidores conectados por bucket, generar órdenes de manera estocástica, simular su asignación y entrega, y registrar las métricas de desempeño clave. Para ello, se implementaron tres componentes principales: (i) el cálculo heurístico inicial, (ii) el proceso de optimización secuencial y (iii) el entorno simulado utilizando SimPy. A continuación, se detalla cada uno de estos bloques.

3.1.5.1 Cálculo Inicial Heurístico

Como punto de partida, se estima el número inicial de repartidores necesarios para cada bucket (intervalo de 15 minutos) mediante una fórmula simple, basada en el cociente entre el promedio estimado de órdenes y una tasa de utilización objetivo (UTR) definida por hora. La relación utilizada es la siguiente:

Donde el factor 0.25 representa la duración del bucket (15 minutos expresados en horas). Esta heurística se encuentra implementada en el módulo configuration_optimization.py, específicamente en la función calculate_heuristic, que toma como entrada el archivo bucketed_friday.csv, el cual contiene los promedios históricos por bucket.

Esta estrategia, aunque simple, refleja el procedimiento actualmente utilizado por equipos de planificación operativa y sirve como referencia base para comparar mejoras potenciales.

3.1.5.2 Proceso de Optimización Secuencial

El segundo bloque del modelo corresponde al proceso de optimización. A partir de la configuración heurística inicial, se realiza una búsqueda local en torno a cada bucket horario para identificar el número óptimo de repartidores que minimice el tiempo de espera promedio de las órdenes, penalizando al mismo tiempo desviaciones respecto del UTR objetivo.

Este procedimiento se ejecuta de forma secuencial, optimizando bucket por bucket en orden cronológico. Se asume que las configuraciones futuras pueden construirse parcialmente sobre la base de decisiones ya adoptadas en buckets anteriores. La lógica se encuentra implementada en el módulo configuration_optimization.py, en la función optimize_configuration.

Los elementos más relevantes de este proceso son los siguientes:

- **Rango de exploración adaptativo**

  Para cada bucket se define un conjunto discreto de configuraciones candidatas, centradas en el valor heurístico. La amplitud del rango de exploración se adapta según

la carga relativa (cantidad de repartidores) del bucket, calculada como su percentil dentro del conjunto de valores horarios del día. Esta lógica permite interpolar entre una exploración más amplia en buckets de baja carga y una más acotada en aquellos de alta carga.

A modo de ejemplo, en uno de los escenarios evaluados se utilizó un esquema que interpolaba entre un rango del ±25 % respecto del valor heurístico para los buckets de menor carga y un ±8 % para aquellos con mayor carga. Para ilustrar el funcionamiento de este enfoque, consideremos un ejemplo simple con diez buckets y sus correspondientes valores heurísticos: [50, 75, 80, 85, 90, 95, 100, 110, 120, 130]. Supongamos que queremos calcular el rango de búsqueda adaptativo para el bucket con valor heurístico igual a 90. En este caso, cinco de los diez valores son menores o iguales a 90, por lo que su percentil es 50 %. Aplicando una interpolación lineal entre los extremos del rango de exploración (±25 % para los buckets de menor carga y ±8 % para los de mayor carga), se asigna al bucket un rango intermedio proporcional a su percentil. Este mecanismo permite ajustar dinámicamente la amplitud de búsqueda según el nivel relativo de carga, ampliando la exploración en horarios menos críticos y restringiéndola cuando la demanda es alta.

$$relative_p = 0{,}25 - (0{,}25 - 0{,}08) \times 0{,}5 = 0{,}165 \quad (4)$$

Este valor implica que se considerará una amplitud del ±16,5 % en torno al valor heurístico. Traducido a número de repartidores, el intervalo de exploración será:

$$H \pm \left[H \times relative_p\right] = 90 \pm 14 \quad (5)$$

es decir, se evaluarán configuraciones entre 76 y 104 repartidores. Si se generan once candidatos equiespaciados dentro de ese rango, los valores evaluados serán aproximadamente [76, 79, 82, 85, 88, 91, 94, 97, 100, 103, 104].

- **Penalización por desviación del UTR**

  Además del tiempo de espera promedio obtenido en la simulación, el modelo incorpora una penalización proporcional a la diferencia entre el UTR simulado y el UTR objetivo configurado para ese bucket. Esta penalización busca desalentar configuraciones que, si bien podrían reducir el tiempo de espera, se alejan de los niveles de utilización de flota deseados.

  El parámetro que regula la magnitud de esta penalización es un hiperparámetro escalar, denotado como k, que pondera la importancia relativa del desvío en relación con el tiempo de espera. En uno de los escenarios experimentales se utilizó un valor de k=100,

lo que implica que por cada 0,01 de diferencia entre el UTR simulado y el objetivo se suman 1 unidad de penalización al tiempo de espera promedio. Formalmente:

$$penalización = k \times \left|UTR_{simulado} - UTR_{objetivo}\right| \quad (6)$$

Por ejemplo, si para un bucket determinado el UTR objetivo es de 0,80 y una configuración candidata arroja un UTR simulado de 0,75, la penalización calculada es:
Esto implica que el tiempo de espera promedio de esa configuración será incrementado en 5 unidades antes de ser comparado con otras opciones. De esta manera, el modelo favorece aquellas soluciones que no sólo sean eficientes en términos de tiempos de espera, sino también consistentes con el nivel de utilización previsto por el plan operativo.

$$penalización = 100 \times |2 - 2{,}05| = 5 \quad (7)$$

- Número de repeticiones.

  Dado que el sistema simulado es estocástico, es decir, que presenta variabilidad aleatoria en los arribos y los tiempos de servicio, cada evaluación de una configuración puede arrojar resultados distintos aun si se repite con los mismos parámetros. Por ello, cada candidato es evaluado mediante múltiples simulaciones independientes, y se utiliza el promedio de las métricas obtenidas para realizar la comparación entre alternativas.

  Este número de repeticiones es también un hiperparámetro del modelo. En el experimento de jornada completa se estableció en diez repeticiones por configuración candidata. Esto significa que, para cada cantidad de repartidores considerada dentro del rango de exploración de un bucket, se ejecutan diez simulaciones independientes y se promedian tanto el tiempo de espera como la penalización correspondiente.

  Por ejemplo, si una configuración candidata arroja tiempos de espera promedio de 12, 11, 13, 10, ..., en diez réplicas, y las penalizaciones por UTR en esas simulaciones son respectivamente 4, 6, 5, 5, ..., se calcula el promedio de cada componente (por ejemplo, $\overline{WT}$=11,5 y $\overline{penalización}$=5) y se define un “tiempo de espera efectivo” como:

$$score_{candidato} = \overline{WT} + \overline{penalización} = 11{,}5 + 5 = 16{,}5 \quad (8)$$

  Este valor se utiliza para comparar con el score de otras configuraciones del mismo bucket y seleccionar la más conveniente.

- **Selección del mejor candidato**

  Se elige el número de repartidores que minimiza el tiempo de espera efectivo, definido como la suma del tiempo de espera promedio y la penalización por desviación del UTR.

Este enfoque permite balancear el objetivo de reducir demoras con la necesidad de mantener una utilización eficiente del recurso. En aquellos buckets donde la simulación presenta alta variabilidad, varios candidatos tienden a obtener resultados similares, y el modelo suele conservar el valor heurístico por no existir diferencias significativas entre opciones.

- Estrategia de fallback.

  Durante el proceso secuencial, al evaluar un candidato en un determinado bucket, es posible que los buckets posteriores aún no hayan sido optimizados. En estos casos, se utiliza como configuración de referencia para los buckets futuros el valor heurístico original, garantizando así que cada simulación se realice con una configuración completa y operativa.

3.1.5.3 Simulación en SimPy

La simulación del sistema fue implementada utilizando la librería SimPy, lo que permitió modelar eventos discretos con entidades independientes y lógica personalizada para la creación, asignación y entrega de órdenes. El entorno simulado refleja con fidelidad las condiciones operativas reales.

Los componentes principales del entorno son:

- Generador de órdenes: las órdenes se crean siguiendo una distribución empírica basada en datos reales. La tasa de arribo (tiempo entre llegadas) es variable por bucket y se calcula a partir del promedio de órdenes reales por intervalo.
- Entidades: cada orden es representada como un proceso que entra al sistema, espera a ser asignada, es despachada por un repartidor y finalmente entregada. Los repartidores son modelados como recursos con capacidad múltiple.
- Lógica de servicio: una vez asignada, cada orden atraviesa las siguientes etapas: to_vendor, at_vendor, to_customer y at_customer. Los tiempos correspondientes se extraen de muestras empíricas almacenadas en arrays y se suman para obtener el tiempo de servicio total. La asignación se resuelve bajo una lógica FIFO
- Tamaño de la simulación: el experimento abarca 76 buckets (de 07:00 a 02:00), cubriendo 68400 segundos simulados. Para cada configuración evaluada, se simulan órdenes, se registran métricas por bucket, y se consolida información sobre tiempo de espera, rendimiento y utilización.

3.1.5.4 Estructura del código

Módulo	Función principal
data_extraction.py	Extrae y transforma los datos históricos, filtrando por día y hora, y calculando métricas agregadas por bucket.
presimulation.py	Preprocesa los datos históricos, filtra los pedidos del horario objetivo y construye muestras empíricas.
config.py	Define parámetros globales del ejercicio: hora de inicio y finalización, configuración inicial del UTR y los límites correspondientes, número de réplicas
simulation.py	Contiene la lógica de SimPy: generación de órdenes, manejo de recursos
configuration_optimization.py	Calcula configuración heurística y realiza la optimización secuencial
visualization.py	Reúne todas las funciones de visualización utilizadas para generar gráficos del modelo y sus resultados.
main.py	Ejecuta la simulación y orquesta el flujo general del experimento

Tabla 1 - Estructura y función de los módulos del código

4 Resultados

4.1 Objetivo del análisis experimental

El propósito de esta sección es presentar y analizar los resultados obtenidos a partir de la simulación del sistema de entrega a domicilio modelado en los capítulos anteriores. A través de un enfoque comparativo, se busca evaluar el impacto de distintas configuraciones de planificación de flota sobre el desempeño operativo del sistema. En particular, se contrastan las dos estrategias de dimensionamiento: la heurística, basada en métodos simples actualmente utilizados en la práctica, y la optimizada, obtenida mediante simulaciones que exploran configuraciones alternativas con el objetivo de minimizar el tiempo de espera promedio.

El análisis se enmarca en un entorno simulado que reproduce las condiciones de un viernes típico, considerando las tasas de arribo observadas y los tiempos de servicio ajustados empíricamente. La evaluación se realiza a lo largo de toda la jornada operativa (07:00 a 02:00), segmentada en bloques de 15 minutos, y contempla métricas clave como el tiempo de espera, el rendimiento, y la utilización del sistema.

Dado que el objetivo no es replicar con exactitud el comportamiento observado en la operación real, sino evaluar el rendimiento relativo de ambas estrategias bajo un mismo conjunto de condiciones controladas, no se realiza una comparación directa con datos históricos completos. En cambio, se examina la consistencia, eficiencia y robustez de cada enfoque en diferentes escenarios experimentales, particularmente ante condiciones más exigentes o de mayor dispersión. Esto permite valorar el potencial del modelo propuesto como herramienta de apoyo a la planificación.

4.2 Escenarios analizados

El análisis experimental se construyó a partir de la ejecución de múltiples simulaciones bajo distintas configuraciones de hiperparámetros. El objetivo principal es evaluar cómo se comporta el modelo frente a variaciones en los criterios de optimización y en los márgenes permitidos para la exploración de soluciones. En particular, se busca observar de qué forma estas decisiones impactan tanto en la solución heurística como en la optimizada, y cómo se reflejan en las métricas de performance del sistema.

Cada escenario se define por un conjunto específico de hiperparámetros, incluyendo:

- El porcentaje de exploración adaptativa aplicado sobre el valor heurístico de flota (esto es, cuán amplio es el rango de candidatos simulados por bucket).

- La cantidad de puntos evaluados en la malla de candidatos (es decir, el tamaño del conjunto discreto de posibles repartidores).
- El coeficiente de penalización aplicado por desvío respecto al UTR objetivo (parámetro k), que introduce un trade-off entre eficiencia y estabilidad.

Estas configuraciones permiten analizar cómo se ajusta el número óptimo de repartidores al flexibilizar o endurecer los criterios del proceso de optimización. Además, permiten evaluar la sensibilidad del sistema a pequeñas variaciones en la dotación, particularmente en aquellos buckets donde la relación entre demanda y capacidad es más crítica. Los tres escenarios definidos se resumen en la Tabla 2, donde se detallan sus principales parámetros y objetivos de análisis.

Escenario	Rango de exploración	Configuraciones candidatas	Penalización UTR (k)	Simulaciones por candidato	Observaciones
Escenario 1 – Configuración base	±25 % a ±8 % (según percentil de carga)	11	100	30	Rango adaptativo amplio y penalización moderada. Usado como referencia base.
Escenario 2 – Exploración extendida	−40 % a +40 %	8	100	30	Explora mayor flexibilidad en la dotación, sin aumentar la penalización.
Escenario 3 – Mayor robustez estadística	−30 % a +5 %	11	200	50	Penalización alta al UTR y mayor robustez por cantidad de simulaciones.

Tabla 2 – Configuraciones de escenarios simulados

4.3 Análisis de los resultados obtenidos

Nota: Para facilitar la interpretación de los gráficos, en el Apéndice D se incluye una tabla de referencia que relaciona cada índice de bucket con su horario correspondiente.

4.3.1 Flota sugerida

4.3.1.1 Escenario 1

La Figura 16 muestra que la optimización del tamaño de flota introduce ajustes puntuales sobre la flota heurística sólo en momentos de mayor demanda. En las primeras horas (buckets 0 a 15, 07:00 a 11:00) la dotación se mantiene igual al escenario base. Durante el subpico matinal (buckets 16 a 30), el modelo incrementa la dotación en 5 a 15 repartidores, especialmente en aquellos intervalos con mayor tolerancia operativa. Entre los buckets 31 a 50, con demanda estable, convergen ambas curvas. En el pico vespertino (51 a 75). reaparecen incrementos controlados. En las horas tardías (desde el bucket 70 en adelante), la estrategia retoma un enfoque más conservador. En síntesis, el Escenario 1 actúa de forma reactiva y contenida, ajustándose siempre dentro del rango permitido por el UTR.

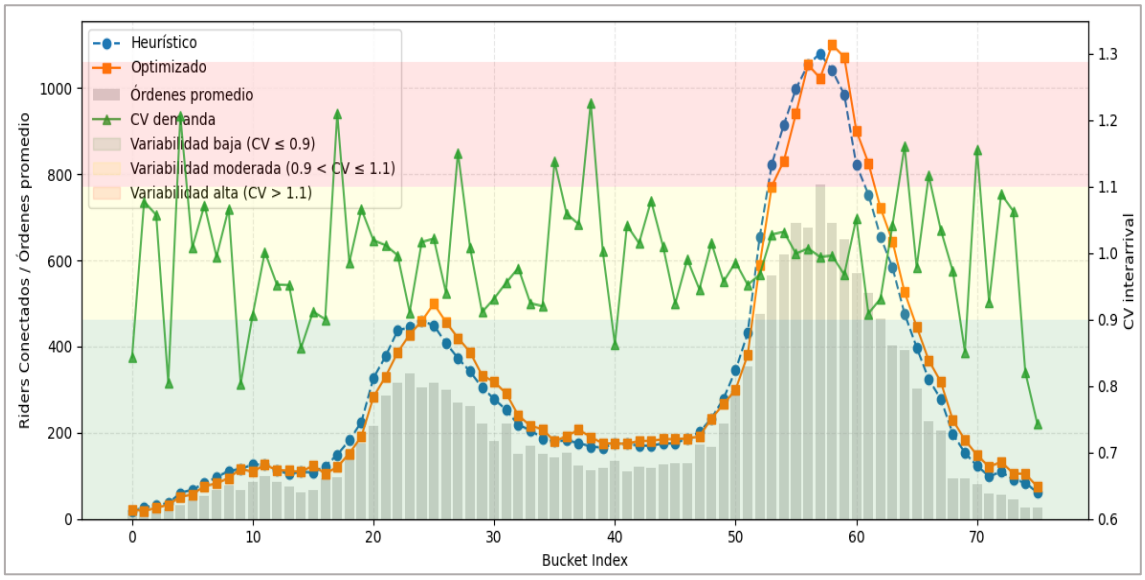

Figura 16 – Flota estimada por método heurístico y optimizado frente a variabilidad de la demanda (Escenario 1)

4.3.1.1 Escenario 2

Con un rango de búsqueda de −40 % a +40 % y k = 100, este escenario exhibe el mayor contraste. En las horas iniciales (0 a 15) se observa una separación de las curvas que representan la planificación heurística y óptima. Este patrón se repite en ambos picos: durante la fase ascendente de la demanda, la curva optimizada se sitúa por debajo de la heurística; al alcanzar el pico, se invierte la relación y la optimizada queda por encima durante el descenso, sosteniendo la flota para gestionar las órdenes en espera. La fase central también evidencia desviaciones, y la reducción de dotación al final del día es más lenta que en el Escenario 1. Esta inercia posterior al pico refleja la lógica secuencial con la que opera el optimizador, evalúa cada bucket suponiendo que los siguientes seguirán la heurística y compensa los periodos más estresados.

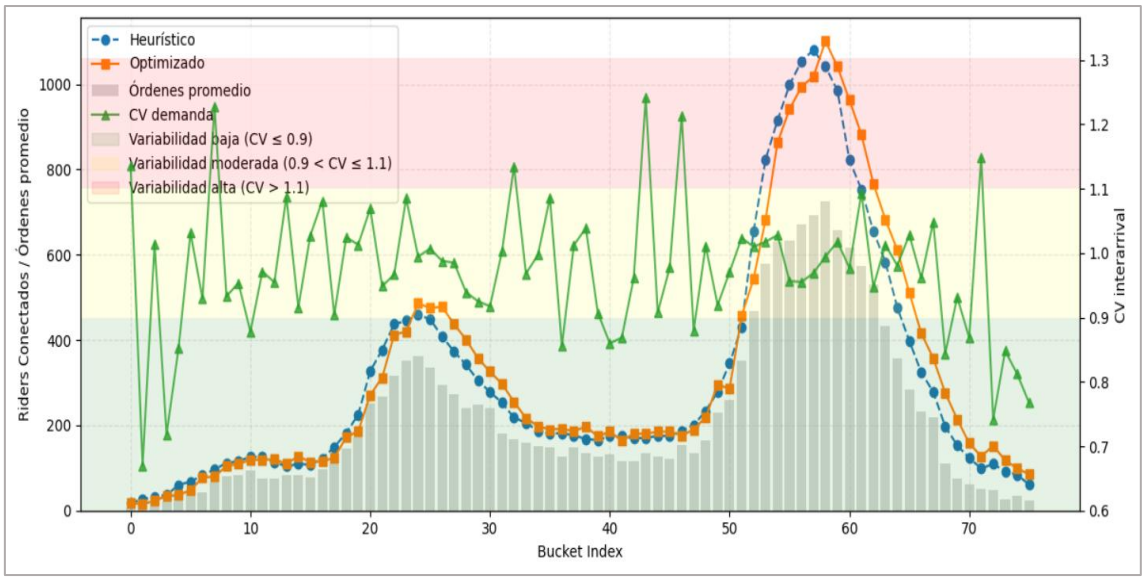

Figura 17 – Flota estimada por método heurístico y optimizado frente a variabilidad de la demanda (Escenario 2)

4.3.1.1 Escenario 3

Con k = 200 y un rango limitado (−30 % a +5 %), el modelo es mucho más controlado. En las horas iniciales reduce la flota, y en ambos picos repite el mismo patrón inverso: dotación optimizada inferior mientras la carga sube, y ligeramente superior en el descenso. El margen de ajuste, sin embargo, es pequeño. Fuera de los tramos de pico, la curva coincide casi por completo con la heurística. Incluso con estas restricciones, el modelo sostiene algo más de flota tras los picos para absorber el arrastre operativo.

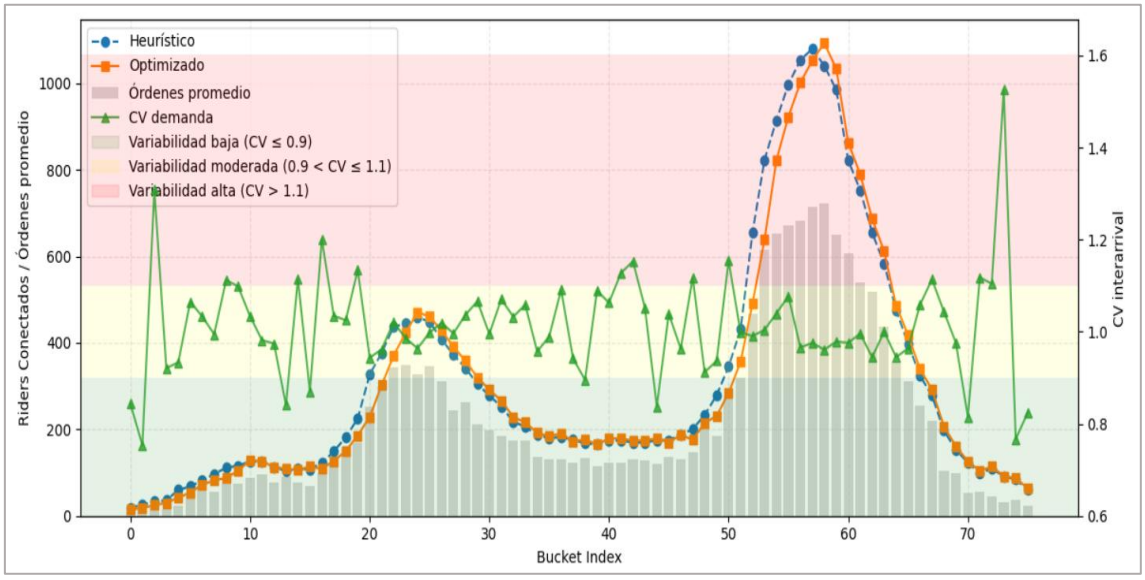

Figura 18 Flota estimada por método heurístico y optimizado frente a variabilidad de la demanda (Escenario 3)

4.3.1.2 Comparativa entre los tres escenarios

Los tres escenarios ilustran un gradiente claro de flexibilidad operativa:

- El escenario 1 introduce cambios puntuales solamente cuando la tolerancia de UTR lo permite, de modo que la flota se mantiene muy cercana a la configuración heurística durante la mayor parte del día.
- El escenario 2 despliega la respuesta más expansiva: recorta dotación en el ascenso a cada pico y la incrementa inmediatamente después, sosteniéndose por encima de la heurística durante la fase donde se gestiona la mayor parte de los pedidos acumulados.
- Finalmente, el escenario 3 se adhiere casi por completo al escenario inicial, salvo ajustes mínimos en los tramos de mayor estrés.

Un rasgo común atraviesa los tres casos: tras cada pico de demanda el modelo conserva una dotación superior a la heurística. El efecto resorte se intensifica a medida que aumenta el rango de exploración disponible. La explicación radica en la naturaleza secuencial de la optimización: al estimar cada bucket suponiendo que los siguientes permanecerán en modo heurístico, el modelo prefiere sobreflotar los intervalos inmediatamente posteriores al pico para garantizar la descarga de la cola acumulada.

Esta misma dinámica aclara por qué, en los segmentos donde las curvas convergen (horas centrales y primeras rampas de cada pico), no se observan mejoras sustanciales en el tiempo de espera. Allí la penalización asociada al UTR, grande o pequeña según el escenario, pesa más que los segundos de espera potencialmente ahorrables. En ese equilibrio, la relación eficiencia-recursos/UTR prevalece como criterio dominante, llevando al optimizador a validar la dotación heurística.

En síntesis, la comparación muestra que la libertad de exploración amplía los márgenes para reducir espera, pero sólo resulta efectiva allí donde la demanda deja margen para sobre o subdimensionar sin violar el UTR. Cuando la variabilidad intrabucket es alta y la penalización de eficiencia se vuelve determinante, las tres estrategias convergen y las oportunidades de mejora en tiempo de espera se diluyen.

4.3.1.3 Influencia del rango de tolerancia de UTR

Un aspecto relevante que emerge del análisis es el efecto modulador del rango de tolerancia del UTR sobre la configuración final sugerida por el modelo. En buckets donde la tolerancia del UTR es reducida (±1-2 %), la optimización secuencial termina replicando casi de forma exacta la configuración heurística original. En contraste, cuando la tolerancia es más amplia (±5-9 %), el modelo encuentra espacio para seleccionar configuraciones con ajustes de repartidores más significativos, permitiendo mayores desviaciones en relación con la heurística.

4.3.2 Tiempo de espera

4.3.2.1 Escenario 1

La figura 19 evidencia una alta dispersión entre estrategias. En los buckets 20 a 25 la optimización alcanza picos de hasta 285 s, muy por encima de la curva heurística, lo que refleja la sensibilidad del modelo ante shocks de demanda inesperados. A partir del bucket 40, coincidiendo con los tramos de mayor carga, la curva optimizada desciende y se iguala (o mejora levemente) frente a la heurística, prueba de que la lógica de búsqueda local es más efectiva bajo condiciones de estrés sostenido.

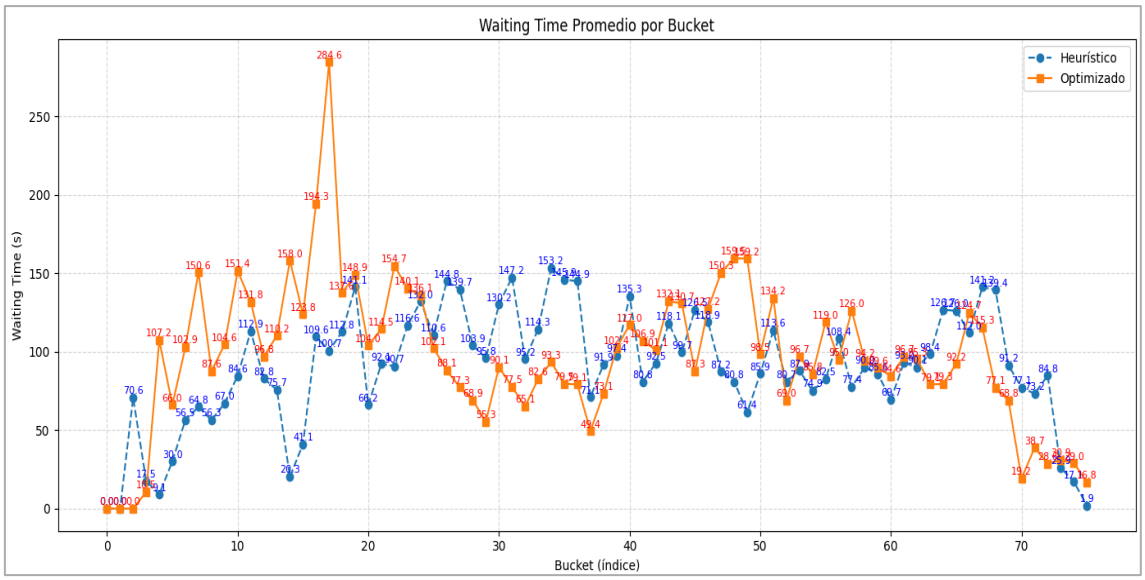

Figura 19 - Comparación del tiempo de espera promedio por bucket: heurístico vs. Optimizado (Escenario 1)

4.3.2.2 Escenario 2

Con mayor libertad de acción, la optimización reduce en promedio el tiempo de espera respecto a la heurística, pero a costa de extremos más pronunciados. En los primeros cinco buckets aparecen dos picos aislados (≈240 s y ≈180 s) que superan a la curva del escenario base y se asocian a la sobre-dotación inicial que mostró la flota. No obstante, desde el bucket 10 hasta el 60 el modelo consigue mantener la espera sistemáticamente por debajo o igual a la heurística, con descensos de hasta 30 s en los intervalos 45 a 55. El patrón “por debajo en ascenso, por encima en descenso” observado en la flota se traduce aquí en un recorte de espera durante la rampa y un ligero aumento post pico, aunque el saldo neto es favorable.

Figura 20 Comparación del tiempo de espera promedio por bucket: heurístico vs. Optimizado (Escenario 2)

4.3.2.3 Escenario 3

La penalización por el desvío respecto de los UTRs configurados inicialmente restringe en gran medida la posibilidad de seleccionar valores que reduzcan significativamente el tiempo de espera. En la mayoría de los buckets, las diferencias frente a los valores obtenidos con la simulación heurística no superan los ±15 segundos. No obstante, el patrón inverso por dotación se refleja en la espera: durante el ascenso de la demanda (cuando se asigna menos flota que en la heurística) aparecen picos de espera más altos, mientras que, en el tramo de descenso, al sostener una flota levemente superior, la optimización logra recortes marginales. El resultado final es un perfil menos volátil, con una curva de espera más estable, aunque con un impacto limitado sobre el promedio diario.

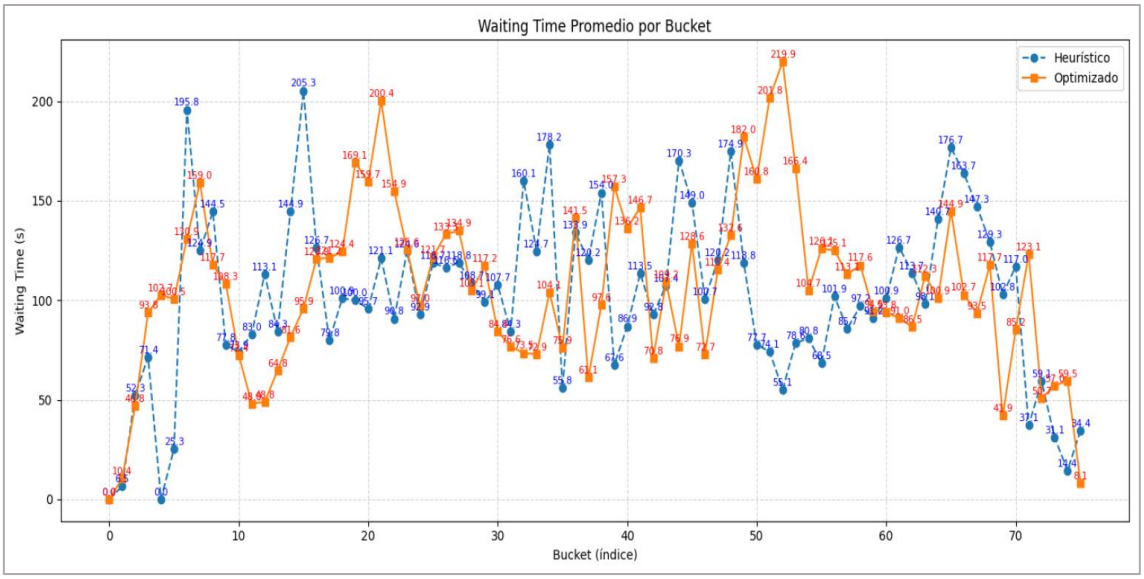

Figura 21 - Comparación del tiempo de espera promedio por bucket: heurístico vs. Optimizado (Escenario 3)

4.3.2.4 Comparación entre escenarios

La comparación confirma que el rango de exploración es el factor que más influye en la capacidad de reducir espera. Con libertad amplia (Escenario 2) el modelo consigue rebajar los tiempos en la mayor parte del día, aunque estos se ven algo empañados con la volatilidad al principio del día. Con rango acotado y penalización alta (Escenario 3) las curvas se estabilizan, priorizando consistencia sobre mejoras. El Escenario 1 queda en un punto intermedio: presenta dispersión elevada en la mañana, pero recorta espera en los tramos críticos de la tarde.

En síntesis, las oportunidades reales de mejora en tiempo de espera emergen cuando el modelo puede ajustar flota sin enfrentar penalizaciones excesivas, pero dichas mejoras se distribuyen de manera desigual a lo largo del día y se concentran en los intervalos de mayor estrés operativo.

4.3.3 Utilización del sistema (UTR)

4.3.3.1 Escenario 1

La curva optimizada se aproxima más al objetivo configurado que la heurística, especialmente entre los buckets 20 y 55, donde logra atenuar los desvíos más pronunciados. En los tramos inicial y final de la jornada, ambos enfoques se apartan del valor objetivo, afectados por el bajo volumen de órdenes y la elevada variabilidad relativa.

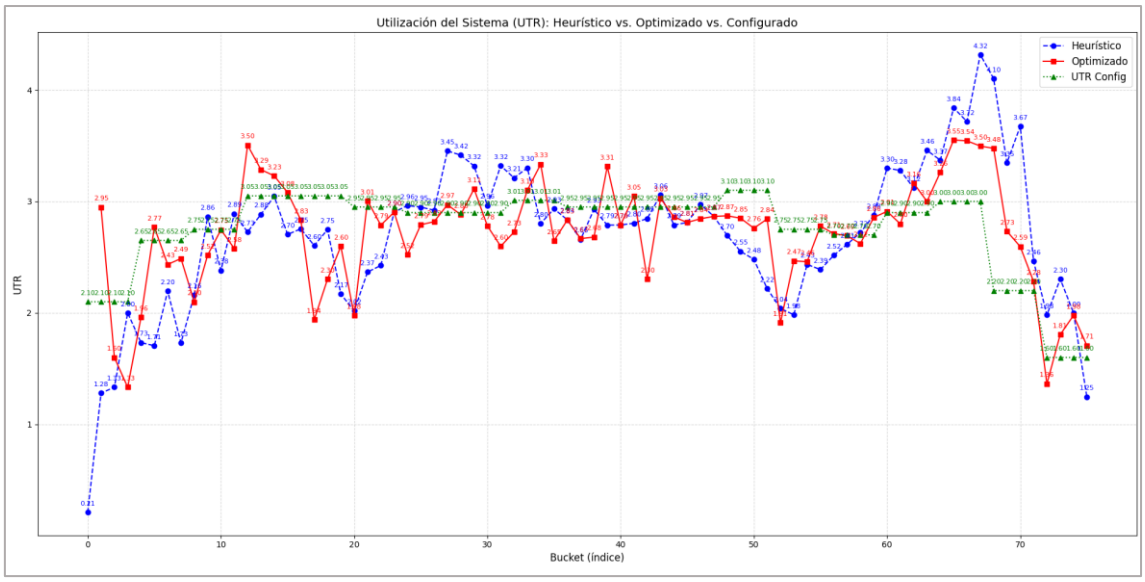

Figura 22-Comparación de la Utilización del Sistema (UTR) por bucket: heurístico, optimizado y configurado (Escenario 1)

4.3.3.2 Escenario 2

Con un mayor margen de exploración, la optimización sigue la meta de UTR con una dispersión comparable a la del Escenario 1, aunque se observan leves sobreutilizaciones durante el descenso de cada pico. En los tramos centrales (buckets 25 a 50), la curva roja oscila dentro de ±0.15 puntos del objetivo, corrigiendo parcialmente las fluctuaciones abruptas presentes en la curva heurística.

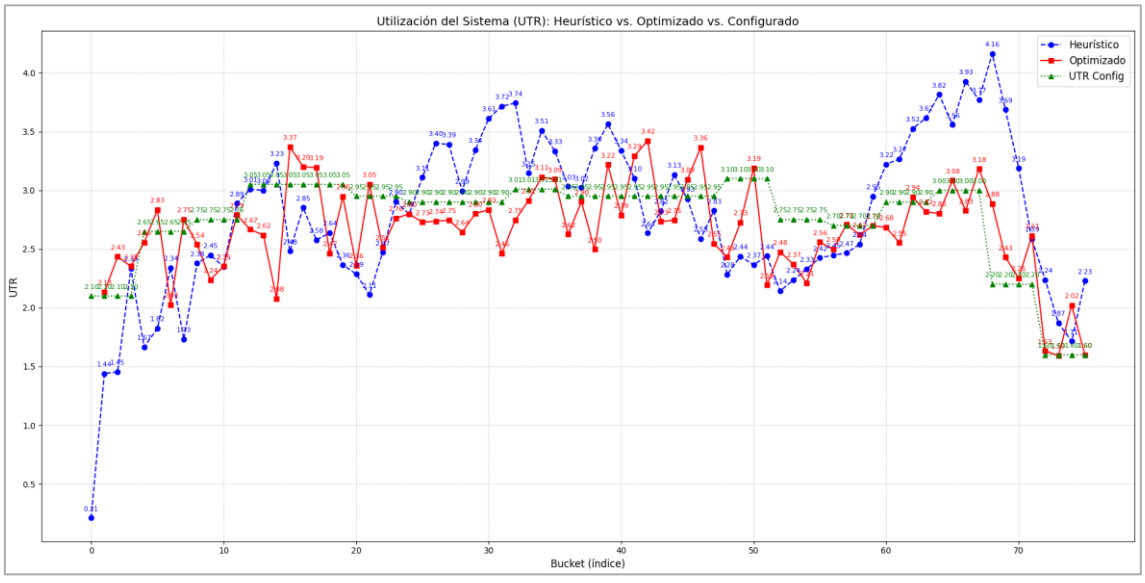

Figura 23 Comparación de la Utilización del Sistema (UTR) por bucket: heurístico, optimizado y configurado (Escenario 2)

4.3.3.3 Escenario 3

La penalización elevada (k = 200) obliga al modelo optimizado a mantenerse muy próximo al objetivo configurado durante casi toda la jornada. La dispersión se reduce al mínimo, aunque se toleran pequeñas subutilizaciones durante las rampas ascendentes y leves sobreutilizaciones tras los picos. Esta estrategia evita extremos, pero no corrige los desvíos observados al inicio y al final del día, cuando la demanda es más baja.

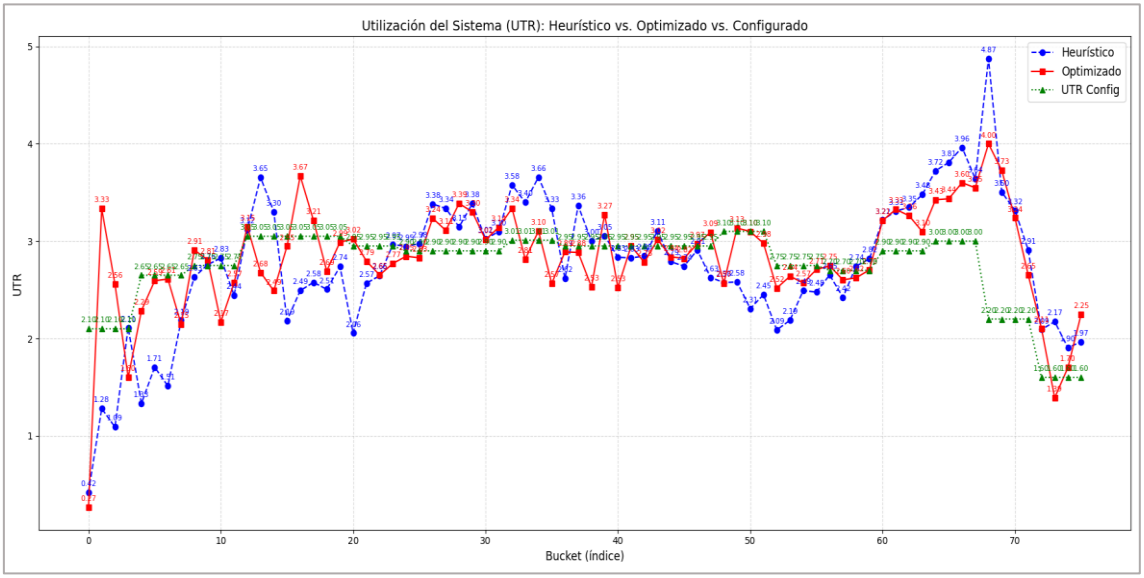

Figura 24 Comparación de la Utilización del Sistema (UTR) por bucket: heurístico, optimizado y configurado (Escenario 3)

4.3.3.1 Comparación entre escenarios

Los tres escenarios muestran distintos grados de alineación con el UTR objetivo. El Escenario 1 reduce desvíos en los tramos de mayor estabilidad operativa, pero mantiene variabilidad en los extremos del día. El Escenario 2, con mayor libertad, suaviza oscilaciones, aunque presenta sobreutilizaciones puntuales tras los picos. En contraste, el Escenario 3 prioriza el cumplimiento estricto del objetivo, sacrificando flexibilidad. En conjunto, se observa que mayor penalización implica menor dispersión, pero también menor capacidad de adaptación dinámica.

4.3.4 Rendimiento

4.3.4.1 Escenario 1

Al analizar el rendimiento por bucket (órdenes completadas en ese bucket), se observa que ambas configuraciones presentan resultados prácticamente idénticos a lo largo del día. La optimización no genera cambios sustanciales en la tasa de procesamiento de órdenes frente a la heurística, con diferencias mínimas y sin tendencias claras o significativas.

Figura 25-Comparación del rendimiento por bucket: heurístico vs. Optimizado (Escenario 1)

4.3.4.2 Escenario 2

En este escenario, al igual que en el primero, el rendimiento no muestra variaciones importantes respecto a la configuración heurística. Ambas curvas coinciden estrechamente durante toda la jornada operativa, lo que sugiere que ambas flotas propuestas son capaces de atender exitosamente la demanda generada. Las pequeñas diferencias observadas carecen de una dirección clara o relevancia operativa, lo que indica que la mayor flexibilidad del modelo no se traduce necesariamente en una mejora en esta métrica.

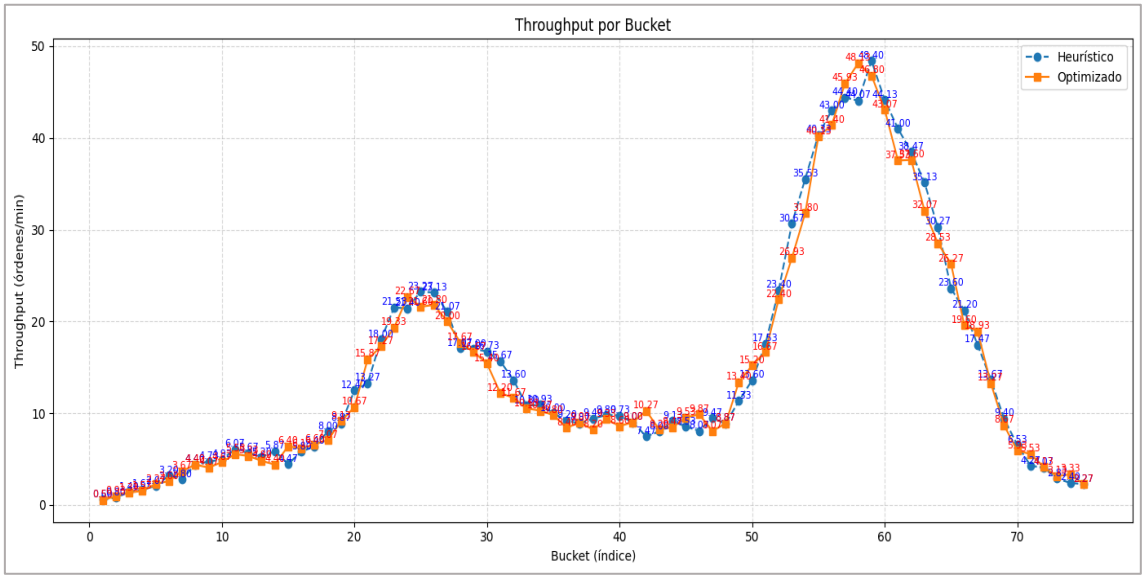

Figura 26 Comparación del rendimiento por bucket: heurístico vs. Optimizado (Escenario 2)

4.3.4.3 Escenario 3

El patrón observado en los escenarios anteriores se repite en este caso: el rendimiento permanece prácticamente idéntico entre la configuración heurística y la optimizada. La fuerte restricción sobre las desviaciones del UTR no parece afectar significativamente la capacidad del sistema para procesar la demanda generada. Este resultado refuerza la idea de que, bajo condiciones normales y con la demanda completamente satisfecha, la optimización propuesta tiene un impacto limitado sobre esta métrica específica.

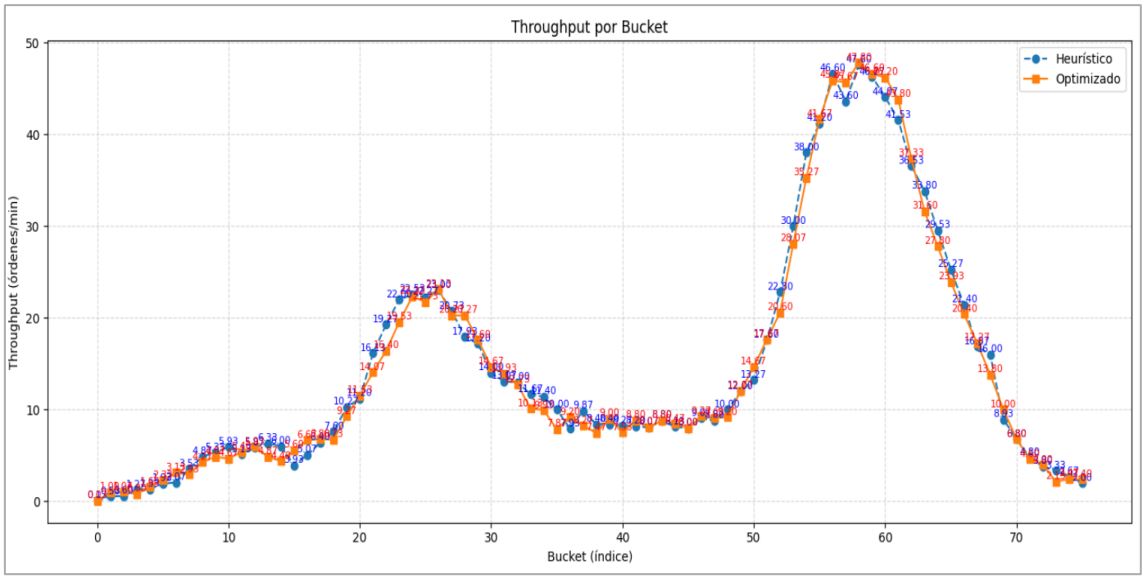

Figura 27 Comparación del rendimiento por bucket: heurístico vs. Optimizado (Escenario 3)

4.3.4.4 Comparación entre escenarios

En resumen, los resultados evidencian que el rendimiento no constituye una métrica diferenciadora relevante entre los escenarios evaluados. La razón principal radica en que la demanda simulada fue atendida exitosamente por todas las configuraciones analizadas, independientemente del rango exploratorio o la penalización aplicada. Esto implica que las ventajas o limitaciones observadas en otros aspectos, como tiempo de espera y estabilidad del UTR, no influyen significativamente sobre la capacidad general del sistema para procesar órdenes. En otras palabras, mientras la demanda sea totalmente satisfecha, la optimización influirá en la calidad del servicio y en la eficiencia operativa, pero no en la cantidad de órdenes procesadas por unidad de tiempo.

5 Conclusiones

Este trabajo se enfocó en estimar con mayor precisión el tamaño necesario de la flota intradiaria en un sistema urbano de entrega a domicilio, abordando directamente la variabilidad operativa del sistema. A diferencia de la lógica heurística tradicional, basada en promedios históricos y tasas objetivo, el enfoque propuesto incorpora explícitamente la incertidumbre en la demanda y los tiempos de servicio mediante simulaciones discretas.

Desde la etapa exploratoria se identificaron dos fuentes estructurales de variabilidad que condicionan fuertemente el desempeño del sistema: (1) la variabilidad en la demanda a lo largo del día y (2) la naturaleza diversa de las órdenes, manifestada en la distribución y dispersión de los tiempos de las distintas etapas que conforman el tiempo de servicio total. Este doble componente introduce desafíos que no pueden ser adecuadamente capturados por métodos deterministas o reglas fijas de planificación.

A través de simulaciones construidas con muestras empíricas de datos históricos, se evaluó el desempeño del sistema bajo distintas configuraciones de flota, midiendo su impacto sobre métricas clave como el tiempo de espera (tiempo de espera), la utilización de recursos (UTR) y el rendimiento. La lógica del optimizador explora múltiples alternativas de dotación para cada intervalo del día, seleccionando aquella que minimiza el tiempo de espera sin desviarse significativamente del UTR configurado.

Los experimentos realizados sobre una jornada típica de viernes permitieron evaluar tres escenarios con distintos niveles de flexibilidad en la exploración y penalizaciones asociadas al UTR. Si bien todos lograron atender exitosamente la demanda generada, resultando en niveles similares de rendimiento, se observaron diferencias significativas en la gestión de la flota y la experiencia del cliente. En particular, el modelo optimizado mostró una mejor alineación con los valores objetivo del UTR, especialmente en tramos de alta exigencia operativa, logrando reducciones sustantivas en el tiempo de espera durante los momentos críticos.

Cabe destacar que el modelo, al haber sido deliberadamente simplificado para facilitar su implementación y análisis, puede no haber absorbido completamente el efecto de la variabilidad en los tiempos de servicio. Sin embargo, los resultados obtenidos demuestran que incluso con un enfoque básico, la simulación permite una aproximación más sensible a las dinámicas reales del sistema. La capacidad de testear escenarios diversos y ajustar configuraciones según contexto supera ampliamente las limitaciones de los enfoques heurísticos, especialmente en entornos operativos caracterizados por alta incertidumbre.

En términos metodológicos, se concluye que los hiperparámetros del modelo, particularmente el rango de exploración y la penalización por desvío del UTR, desempeñan un rol central en el equilibrio entre eficiencia y calidad del servicio. Configuraciones con mayor flexibilidad muestran una mejor adaptación a los picos de demanda, aunque con mayor dispersión en los niveles de utilización, mientras que alternativas más conservadoras privilegian la estabilidad, pero con menos capacidad de ajuste dinámico.

En síntesis, la potencia del modelo no radica en ofrecer una única solución óptima, sino en brindar una estructura flexible capaz de adaptarse a diferentes condiciones operativas y prioridades estratégicas.

5.1 Mejoras y oportunidades

A partir de los resultados obtenidos y las limitaciones identificadas en el modelo actual, emergen diversas oportunidades para futuros desarrollos. En primer lugar, resulta prometedor profundizar en el ajuste de los hiperparámetros a nivel intradiario. Una segmentación más fina permitiría definir rangos específicos de variación del UTR, siendo más flexibles durante periodos con alta variabilidad operativa y más estrictos en franjas estables, optimizando así el balance entre eficiencia y calidad de servicio.

Asimismo, podría explorarse la inclusión de mecanismos adaptativos en la penalización del UTR, permitiendo que su peso varíe dinámicamente en función de la cantidad de órdenes pendientes por asignar o de la tasa reciente de incumplimiento del servicio. Esto posibilitaría una respuesta más ágil ante situaciones imprevistas o cambios abruptos en la demanda.

Otra línea de mejora consiste en expandir el modelo a la simulación de escenarios de estrés operativo, como eventos climáticos extremos o picos de demanda no previstos. Analizar el comportamiento del sistema en condiciones no estacionarias facilitaría la validación de la robustez del modelo y su aplicabilidad en contextos reales de alta incertidumbre.

También se destaca la necesidad de seguir profundizando en el desarrollo del modelo de simulación que representa la operación, con el objetivo de incorporar progresivamente las complejidades reales del sistema. Si bien el enfoque actual fue deliberadamente simplificado para facilitar su implementación inicial, futuras versiones podrían mejorar su realismo incorporando dinámicas operativas más representativas. Entre ellas, se incluye la posibilidad de modelar la asignación múltiple (dos o más órdenes) a un mismo repartidor, la proporción de órdenes que son rechazadas y su consecuente reasignación, así como las lógicas de priorización

y la retención momentánea de la orden que aplicada por el algoritmo de despacho antes de asignarla, posibilitando la sincronización con el tiempo de preparación del pedido en el restaurante. En este trabajo se asumió una lógica simplificada de asignación FIFO, pero la incorporación de estos elementos permitiría capturar con mayor fidelidad el impacto de la variabilidad operativa y aumentar el poder explicativo y predictivo de la simulación como herramienta de planificación.

Por otro lado, resultaría valioso avanzar en la incorporación explícita del impacto económico asociado a cada configuración de flota. Evaluar los costos operativos marginales derivados de la sobreutilización o subutilización de recursos permitiría traducir las decisiones del modelo en indicadores financieros concretos, fortaleciendo así su utilidad como herramienta de apoyo a la toma de decisiones estratégicas. No obstante, esta línea de trabajo implica un mayor nivel de complejidad metodológica y de disponibilidad de datos, por lo que se recomienda como una instancia posterior dentro del camino evolutivo del modelo.

Finalmente, la metodología desarrollada presenta un alto potencial de aplicación práctica en los entornos operativos actuales, dado que su implementación no requiere desarrollar nuevas herramientas, sino modificar la lógica de cálculo que alimenta las curvas de dotación horaria. En particular, el sistema de planificación interna que actualmente utilizan los equipos locales podría incorporar esta lógica de simulación de manera modular, sumando como nuevos insumos algunos de los parámetros clave (como el rango de exploración o el peso asignado a cada objetivo operativo, ya sea tiempo de espera o tasa de servicio). Esta adaptación permitiría aprovechar la infraestructura ya disponible, pero con una capacidad mucho mayor para representar contextos de alta variabilidad.

Además de su valor analítico, disponer de un motor de simulación simple pero flexible habilita a los equipos locales a explorar escenarios alternativos con bajo riesgo, testear configuraciones de flota ante picos de demanda o variaciones en los patrones de servicio, y tomar decisiones mejor informadas sin necesidad de escalar cada caso a desarrollos centralizados. Esta posibilidad de experimentar en un entorno controlado es especialmente valiosa en una operación como la entrega a domicilio, donde los desafíos cambian no solo entre zonas, sino también a lo largo del día. Incluso podrían definirse umbrales críticos de cobertura a partir de simulaciones, para generar alertas que activen medidas tácticas como incentivos en horarios sensibles. En ese sentido, la metodología no solo mejora la precisión de la planificación, sino que también contribuye a una gestión más contextual, proactiva y descentralizada.

6 Referencias

Agatz, N., Campbell, A., Fleischmann, M., & Savelsbergh, M. (2013). Optimization for dynamic ride-sharing: A review. European Journal of Operational Research, 223(2), 295–303. https://doi.org/10.1016/j.ejor.2012.05.028

Banco Interamericano de Desarrollo (BID). (2023). Nuevas modalidades laborales en la economía digital: Un estudio empírico del trabajo de reparto en Argentina. https://publications.iadb.org/publications/spanish/document/nuevas-modalidades-laborales-en-la-economia-digital.pdf

Banco Interamericano de Desarrollo (BID (2022). Digital platforms in Latin America: A study of their impact on the economy and employment https://publications.iadb.org/en/digital-platforms-latin-america-study-their-impact-economy-and-employment

Buldeo Rai, H., Verlinde, S., Merckx, J., & Macharis, C. (2022). *Crowdsourced entrega a domicilio: A review of platforms and academic literature*. Transportation Research Part C: Emerging Technologies, 132, 103374.

Crainic, T. G., Ricciardi, N., & Storchi, G. (2009). *Models for evaluating and planning city logistics systems*. Transportation Science, 43(4), 432–454. https://doi.org/10.1287/trsc.1090.0279

Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). Fundamentals of queueing theory (4th ed.). Wiley.

McKinsey & Company. (2021, noviembre). Ordering in: The rapid evolution of food entrega a domicilio. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ordering-in-the-rapid-evolution-of-food-entrega a domicilio

Srihita, S., Krishnamoorthy, S., & Thomas, G. (2019). *The economics of food entrega a domicilio*. FactorDaily. https://archive.factordaily.com/the-economics-of-food-entrega a domicilio/

Zhang, R., & Pavone, M. (2014). Control of robotic mobility-on-demand systems: A queueing-theoretical perspective**.** *The International Journal of Robotics Research*, 35(1–3), 186–203. Disponible en: https://arxiv.org/abs/1404.439

Apéndice A - Estadísticas descriptivas de los tiempos operativos

Este apéndice presenta un resumen estadístico de las variables temporales claves utilizadas en el ejercicio. La Tabla 3 muestra los principales estadísticos descriptivos (media, desvío estándar, percentiles y valores extremos) para cada uno de los componentes del ciclo de vida de la orden:

	mean	std	min	25%	50%	75%	max
interarrival_time	5.9	130.19	0.0	0.83	2.21	5.17	18236.35
service_time_min	18.38	7.48	2.15	13.1	17.08	22.2	156.52
to_vendor_time	2.85	2.67	0.0	0.28	2.47	4.32	39.1
at_vendor_time	2.59	3.92	0.0	0.0	1.23	3.52	101.88
to_customer_time	7.9	4.98	0.0	4.45	6.83	10.2	106.25
at_customer_time	2.45	1.81	0.0	1.32	2.08	3.1	55.15

Tabla 3 Estadísticos descriptivos generales para los componentes de tiempo

La Tabla 4 detalla el comportamiento promedio total por bucket de cada uno de los componentes del tiempo de servicio durante la jornada operativa. Los valores están expresados en minutos, exceptuando el tiempo entre llegadas que está en segundos. Se utilizó una escala de color para resaltar visualmente los tramos con mayor duración relativa en cada componente.

Bucket	***Interarrival mean***	***Service mean***	***to_vendor mean***	***at_vendor mean***	***to_customer mean***	***at_customer mean***
07:00	1819,75	21,62	7	2,62	7,23	2,23
07:15	75,67	21,11	5,76	2,74	7,83	2,28
07:30	65,6	20,12	5,38	2,38	7,78	2,26
07:45	57,05	19,4	5	2,21	7,58	2,36
08:00	30,68	19,18	4,7	2,12	7,72	2,34
08:15	23,27	18,85	4,47	2,2	7,54	2,39
08:30	20,32	18,84	4,35	2,16	7,73	2,38
08:45	17,7	19	4,13	2,15	8,05	2,45
09:00	14,92	18,39	3,94	2,22	7,61	2,38
09:15	13,36	18,57	3,83	2,26	7,82	2,45
09:30	12,6	18,63	3,75	2,26	7,9	2,5
09:45	12,44	18,43	3,59	2,13	7,87	2,54
10:00	12,09	18,76	3,54	2,3	8,07	2,48
10:15	11,78	18,72	3,41	2,2	8,12	2,55
10:30	11,9	18,96	3,31	2,35	8,2	2,59
10:45	11,44	19,05	3,19	2,39	8,31	2,58
11:00	10,04	19,57	3,19	2,49	8,6	2,58
11:15	8,24	19,7	3,22	2,59	8,6	2,59
11:30	6,9	19,59	3,14	2,59	8,58	2,5
11:45	5,53	19,47	3,19	2,5	8,52	2,5
12:00	4,05	19,47	3,13	2,5	8,62	2,47
12:15	3,43	19,37	3,1	2,57	8,59	2,45
12:30	3,13	19,18	3,12	2,56	8,5	2,46

12:45	3,04	19,16	3,07	2,69	8,46	2,46
13:00	3,01	19,1	3,03	2,71	8,44	2,44
13:15	3	18,92	2,98	2,73	8,25	2,45
13:30	3,23	18,67	2,89	2,59	8,2	2,47
13:45	3,52	18,57	2,9	2,52	8,13	2,47
14:00	3,81	18,39	2,89	2,48	8,04	2,44
14:15	4,14	18,24	2,85	2,44	7,97	2,45
14:30	4,45	18,26	2,89	2,38	7,98	2,49
14:45	4,83	18,15	2,87	2,34	7,95	2,5
15:00	5,5	18,18	2,97	2,33	7,95	2,45
15:15	5,77	18,32	3,02	2,32	8,1	2,45
15:30	6,21	18,52	3,15	2,31	8,21	2,49
15:45	6,39	18,8	3,25	2,39	8,39	2,52
16:00	6,59	18,81	3,32	2,34	8,45	2,53
16:15	6,92	18,74	3,36	2,29	8,39	2,57
16:30	7,01	18,88	3,33	2,34	8,49	2,58
16:45	7	18,95	3,39	2,35	8,48	2,59
17:00	7,03	18,98	3,39	2,42	8,47	2,56
17:15	7,06	19,13	3,48	2,45	8,56	2,54
17:30	6,93	19,74	3,58	2,59	8,88	2,59
17:45	7,06	19,29	3,47	2,51	8,57	2,58
18:00	6,91	19,29	3,45	2,5	8,6	2,54
18:15	6,69	19,27	3,39	2,49	8,55	2,59
18:30	6,43	19,44	3,28	2,63	8,56	2,57
18:45	6,04	19,23	3,15	2,56	8,4	2,58
19:00	5,03	19,02	3,09	2,52	8,22	2,56
19:15	4,3	18,82	2,86	2,49	8,12	2,6
19:30	3,53	18,64	2,72	2,52	8,01	2,54
19:45	2,82	18,45	2,64	2,48	7,9	2,55
20:00	2,2	18,42	2,53	2,64	7,8	2,48
20:15	1,84	18,42	2,49	2,73	7,8	2,47
20:30	1,67	18,53	2,52	2,9	7,82	2,45
20:45	1,56	18,55	2,54	2,97	7,82	2,45
21:00	1,53	18,53	2,6	3,03	7,79	2,44
21:15	1,53	18,53	2,62	3,09	7,79	2,42
21:30	1,63	18,31	2,61	3	7,74	2,41
21:45	1,77	17,91	2,57	2,82	7,58	2,4
22:00	1,93	17,81	2,56	2,81	7,48	2,4
22:15	2,13	17,4	2,51	2,62	7,32	2,39
22:30	2,46	17,04	2,42	2,45	7,17	2,39
22:45	2,84	16,62	2,34	2,29	7	2,32
23:00	3,45	16,37	2,27	2,2	6,83	2,35
23:15	4,08	16,26	2,27	2,05	6,76	2,4
23:30	5,15	16,63	2,36	2,15	6,91	2,48
23:45	6,15	16,77	2,57	2,31	7,19	2,28
00:00	7,74	17,2	2,61	2,43	7,27	2,2

00:15	9,43	17,04	2,69	2,27	7,22	2,17
00:30	11,71	17,12	2,69	2,15	7,37	2,1
00:45	14,31	16,68	2,57	2,09	6,86	2,13
01:00	17,28	17,05	2,47	2,39	6,97	2,1
01:15	20,02	16,88	2,47	2,24	6,91	2,12
01:30	25,03	16,84	2,57	2,11	6,97	2,08
01:45	32,1	16,73	2,47	2,14	6,75	2,13

Tabla 4- Promedios intradiarios por bucket para cada componente del tiempo de servicio

Apéndice B - Análisis de casos atípicos en el ajuste exponencial del tiempo entre llegadas

A pesar de que la gran mayoría de los buckets horarios presentan un buen ajuste a la distribución exponencial, se identificaron algunos casos que no cumplen con este supuesto estadístico según la prueba de Kolmogorov-Smirnov (p-valor < 0.05). En total, se detectaron X buckets no válidos (representando el Y % del total), cuyos detalles se presentan en la Tabla 5.

En estos casos, el desvío puede atribuirse a distintos factores, como:

- Tamaños de muestra muy grandes (e.g., más de 9000 observaciones en el bucket Friday 21:00), lo cual aumenta la sensibilidad del test y puede llevar al rechazo por pequeñas diferencias.
- Tamaños de muestra muy reducidos (e.g., menos de 100 observaciones), lo que puede provocar inestabilidad en la estimación de parámetros.
- Fenómenos extremos de comportamiento no aleatorio (picos de demanda, campañas promocionales, etc.) que alteran la independencia entre llegadas.

Por ejemplo, el bucket Friday 21:00, a pesar de contar con más de 9000 observaciones, obtuvo un p-valor de 0.0013, lo que sugiere una ligera pero estadísticamente significativa desviación respecto al modelo teórico. Sin embargo, el bajo valor del estadístico K-S (0.0199) sugiere que la diferencia práctica es muy pequeña, lo cual refuerza la idea de que el rechazo se debe más a la sensibilidad estadística que a un verdadero mal ajuste.

En definitiva, si bien estos casos merecen ser considerados, no comprometen la validez general del enfoque. Pueden ser tenidos en cuenta como valores típicos metodológicos o analizarse de manera específica si el modelo requiere una alta granularidad en ese rango horario.

Día	Bucket	λ estimado	Estadístico K-S	p-valor	N muestras	Ajuste válido
Friday	00:15	0.0939	0.0539	0.0034	1093	No
Friday	07:00	0.0006	0.7409	0.0	132	No
Friday	16:00	0.1496	0.0418	0.0043	1745	No
Friday	21:00	0.7906	0.0199	0.0013	9251	No
Monday	02:15	0.0052	0.2499	0.0014	56	No
Monday	07:00	0.0005	0.7549	0.0	98	No
Monday	13:15	0.314	0.0264	0.0117	3673	No
Monday	14:00	0.2326	0.0268	0.0398	2706	No
Monday	14:15	0.2149	0.0271	0.0469	2534	No
Monday	17:30	0.135	0.0364	0.0294	1585	No
Saturday	07:00	0.0006	0.7392	0.0	125	No

Saturday	08:15	0.0407	0.0616	0.0482	485	No
Saturday	20:30	0.642	0.0162	0.0392	7510	No
Saturday	21:30	0.73	0.0163	0.0211	8536	No
Saturday	22:45	0.4781	0.0192	0.032	5587	No
Sunday	07:00	0.0007	0.7649	0.0	147	No
Sunday	10:30	0.0856	0.0478	0.0201	1001	No
Sunday	11:30	0.1413	0.0346	0.0368	1658	No
Sunday	14:30	0.2821	0.0255	0.0267	3298	No
Sunday	20:30	0.684	0.0156	0.04	7998	No
Thursday	07:00	0.0006	0.7533	0.0	124	No
Thursday	10:15	0.0916	0.0502	0.0091	1064	No
Thursday	11:45	0.1875	0.0291	0.0481	2196	No
Thursday	12:15	0.2989	0.0296	0.0042	3495	No
Thursday	12:30	0.3267	0.0403	0.0	3770	No
Thursday	12:45	0.3236	0.0535	0.0	3679	No
Thursday	13:00	0.3201	0.0769	0.0	3908	No
Thursday	13:15	0.3813	0.0709	0.0	4454	No
Thursday	14:30	0.2151	0.0616	0.0	2530	No
Thursday	14:45	0.1843	0.0304	0.0373	2136	No
Thursday	15:00	0.1512	0.0425	0.0033	1767	No
Thursday	15:30	0.1429	0.0361	0.0203	1749	No
Thursday	23:45	0.147	0.0359	0.0245	1696	No
Tuesday	02:30	0.0043	0.2147	0.0467	39	No
Tuesday	07:00	0.0005	0.7208	0.0	105	No
Tuesday	17:30	0.1342	0.0515	0.0005	1562	No
Wednesday	00:00	0.1024	0.0544	0.0013	1230	No
Wednesday	07:00	0.0004	0.7184	0.0	96	No
Wednesday	11:45	0.1925	0.0316	0.0217	2256	No
Wednesday	16:30	0.1301	0.0404	0.0137	1518	No
Wednesday	20:00	0.4084	0.028	0.0011	4781	No
Wednesday	20:15	0.4982	0.0199	0.0193	5822	No

Tabla 5 - Buckets que no cumplen con el supuesto de distribución exponencial (p-valor < 0.05)

Apéndice C - Validación empírica de la distribución del tiempo de servicio

Para validar el modelo de distribución del service_time_min, se evaluó su ajuste a distintas distribuciones teóricas (Gamma, Lognormal, Weibull, Exponencial y Normal) utilizando el test de Kolmogorov-Smirnov por cada hora del día. Los resultados se resumen en la Tabla D.1 del Apéndice D, donde se indica la distribución con mejor ajuste, su p-valor y si dicho ajuste puede considerarse válido ($p > 0.05$). En 6 de las 20 horas analizadas, el p-valor fue superior al umbral de significancia, lo cual indica un ajuste estadísticamente aceptable. La distribución Gamma resultó ser la mejor opción en la mayoría de los casos, incluso en aquellas horas donde el test arrojó un p-valor bajo, por lo que se optó por modelar el tiempo de servicio utilizando esta distribución en la simulación base.

Hora	Mejor Distribución	p-valor	Ajuste válido (p > 0.05)
00:00	Gamma	0.1111	Sí
01:00	Gamma	0.4096	Sí
02:00	Gamma	0.8649	Sí
07:00	Lognormal	0.5982	Sí
08:00	Gamma	0.1244	Sí
09:00	Gamma	0.0169	No
10:00	Gamma	0.0371	No
11:00	Gamma	0.0279	No
12:00	Gamma	0.0001	No
13:00	Gamma	0.0001	No
14:00	Gamma	0.0001	No
15:00	Gamma	0.0738	Sí
16:00	Gamma	0.0287	No
17:00	Gamma	0.0030	No
18:00	Gamma	0.0086	No
19:00	Gamma	0.0000	No
20:00	Gamma	0.0000	No
21:00	Gamma	0.0000	No

22:00	Gamma	0.0023	No
23:00	Gamma	0.0354	No

Tabla 6 - Evaluación del ajuste estadístico por hora del día (Tiempo de servicio)

Apéndice D - Referencia horaria de los buckets

Con el fin de facilitar la interpretación de los gráficos presentados en la Sección 4 (Resultados), este apéndice provee una tabla de equivalencias entre los índices de buckets utilizados en el eje X y su correspondiente horario. La jornada simulada se extiende desde las 07:00 hasta la 02:00 del día siguiente, con intervalos de 15 minutos por bucket. Esta tabla permite al lector identificar con precisión a qué momento del día corresponde cada punto de la serie.

Bucket Index	Hora	Bucket Index	Hora	Bucket Index	Hora	Bucket Index	Hora
0	07:00	**19**	11:45	**38**	16:30	**57**	21:15
1	07:15	**20**	12:00	**39**	16:45	**58**	21:30
2	07:30	**21**	12:15	**40**	17:00	**59**	21:45
3	07:45	**22**	12:30	**41**	17:15	**60**	22:00
4	08:00	**23**	12:45	**42**	17:30	**61**	22:15
5	08:15	**24**	13:00	**43**	17:45	**62**	22:30
6	08:30	**25**	13:15	**44**	18:00	**63**	22:45
7	08:45	**26**	13:30	**45**	18:15	**64**	23:00
8	09:00	**27**	13:45	**46**	18:30	**65**	23:15
9	09:15	**28**	14:00	**47**	18:45	**66**	23:30
10	09:30	**29**	14:15	**48**	19:00	**67**	23:45
11	09:45	**30**	14:30	**49**	19:15	**68**	00:00
12	10:00	**31**	14:45	**50**	19:30	**69**	00:15
13	10:15	**32**	15:00	**51**	19:45	**70**	00:30
14	10:30	**33**	15:15	**52**	20:00	**71**	00:45
15	10:45	**34**	15:30	**53**	20:15	**72**	01:00
16	11:00	**35**	15:45	**54**	20:30	**73**	01:15
17	11:15	**36**	16:00	**55**	20:45	**74**	01:30
18	11:30	**37**	16:15	**56**	21:00	**75**	01:45

Tabla 7 - Equivalencia entre índice de bucket y horario simulado