Aplicación de modelos de colas y simulación para evaluar la eficiencia del servicio de helados en McDonald's Chapinero

Application of Queueing Models and Simulation to Evaluate the Efficiency of the Ice Cream Service at McDonald's Chapinero

Laura Sofía Guerrero Guavita[1], Gabriela Zipaquirá Antolínez[2] & Francy Rocío Castellanos Oviedo[3]

Fecha de recepción: 12-08-2025 | Fecha de aprobación: 07/10/2025

Resumen

McDonald's es una cadena de comida rápida reconocida mundialmente por la eficiencia operativa y estandarización en sus procesos. En Chapinero, Bogotá, se estableció un punto de venta especializado en helados, aprovechando su ubicación estratégica en una zona comercial y estudiantil con alto tránsito peatonal. Este punto, enfocado en productos como McFlurrys, conos y sundaes, generó inicialmente largas filas y tiempos de espera elevados, lo cual motivó la presente investigación.

El estudio tuvo como propósito analizar y optimizar el desempeño operativo del punto de venta mediante modelos de investigación de operaciones e inteligencia artificial. Se recopiló información real sobre tiempos de llegada y atención al cliente, analizando estadísticamente estos datos para identificar la distribución que mejor los representaba, encontrándose la distribución Gamma como la más apropiada. A partir de estos resultados, se realizaron simulaciones computacionales en Python y FlexSim para evaluar diferentes indicadores críticos como tiempo promedio de espera, tasa de servicio y nivel de saturación del sistema. Estas simulaciones permitieron visualizar claramente los procesos internos, identificar cuellos de botella y generar alternativas operativas viables.

1 Estudiante Facultad de Ingeniería Industrial, Universidad Santo Tomás, Bogotá, Colombia. Grupo de Investigación en Procesos Organizacionales – GIPO, Semillero SIGEO. Contacto: laura.guerreo@usantotomas.edu.co, celular: 3124964843.

2 Estudiante Facultad de Ingeniería Industrial, Universidad Santo Tomás, Bogotá, Colombia. Grupo de Investigación en Procesos Organizacionales – GIPO, Semillero SIGEO. Contacto: zipaquiragabriela@usantotomas.edu.co

3 Ingeniera Industrial, Magíster en Ingeniería – Ingeniería Administrativa. Docente Facultad de Ingeniería Industrial Universidad Santo Tomás, Bogotá, Colombia. Grupo de Investigación en Procesos Organizacionales – GIPO, Semillero SIGEO. Contacto: francycastellanos@usta.edu.co

Finalmente, el análisis permitió proponer mejoras concretas relacionadas con la capacidad instalada, la asignación efectiva de personal y la distribución óptima de tareas en función de la demanda real, logrando así una significativa reducción en los tiempos de espera y una mejora sustancial en la eficiencia del servicio ofrecido a los clientes.

Palabras Clave

Teoría de colas, Simulación, Tiempo de espera, Eficiencia Operativa, Optimización

Abstract

McDonald's is a globally recognized fast-food chain known for operational efficiency and standardized processes. In Chapinero, Bogotá, a specialized ice cream outlet was established, strategically located in a busy commercial and student area with high pedestrian traffic. This outlet, focused on products such as McFlurrys, cones, and sundaes, initially experienced long queues and extensive wait times, prompting this research.

The study aimed to analyze and optimize the operational performance of the outlet using advanced industrial engineering techniques and artificial intelligence. Real data on customer arrival and service times were collected and statistically analyzed to identify the distribution that best represented these times, determining the Gamma distribution as the most suitable. Subsequently, computational simulations were performed using Python and FlexSim to evaluate critical indicators such as average waiting time, service rate, and system saturation levels. These simulations provided clear visualization of internal processes, identified bottlenecks, and facilitated the exploration of alternative operational scenarios.

Finally, the analysis led to concrete improvement proposals related to installed capacity, efficient personnel allocation, and optimal task distribution based on actual demand. These changes significantly reduced wait times and substantially enhanced service efficiency and customer experience.

Keywords

Queueing theory, Simulation, Waiting time, Operational efficiency, Optimization

Abreviaciones:

- SAS: *Statistical Analysis System*. Software utilizado para el análisis estadístico y ajuste de distribuciones.
- M/M/1: Modelo de cola con llegadas Poisson (M), tiempos de servicio exponenciales (M) y un solo servidor (1).

- M/M/c: Modelo de cola con llegadas Poisson (M), tiempos de servicio exponenciales (M) y múltiples servidores (c).
- M/G/1: Modelo de cola con llegadas Poisson (M), tiempos de servicio con distribución general (G) y un solo servidor (1).
- M/G/k: Modelo de cola con llegadas Poisson (M), tiempos de servicio con distribución general (G) y múltiples servidores (k).
- KPI: *Key Performance Indicator*. Indicadores clave de rendimiento usados para evaluar la eficiencia del sistema.
- AD: *Anderson-Darling*. Prueba estadística para evaluar el ajuste de distribuciones.
- K-S: *Kolmogorov-Smirnov*. Prueba de bondad de ajuste utilizada para comparar distribuciones.
- W^2: *Cramér-von Mises*. Prueba estadística para evaluar la bondad de ajuste de una distribución.

Introducción

El análisis de la capacidad operativa y la eficiencia desempeña un papel crucial en el sector de servicios, particularmente en aquellos contextos donde la demanda es variable y la atención oportuna es un factor determinante en la satisfacción del cliente. En este ámbito, se destacan herramientas fundamentales como la teoría de colas, la simulación de eventos discretos y el análisis de distribuciones de probabilidad, que permiten modelar, analizar y optimizar procesos de servicio, como es el caso del sector de comidas rápidas.

La industria de comida rápida enfrenta desafíos crecientes para satisfacer las expectativas de los clientes en términos de rapidez y calidad del servicio. Los largos tiempos de espera representan un reto clave para los gerentes, quienes buscan mejorar la percepción del cliente y garantizar un servicio eficiente. Como señala Villarreal et al. (2021),

> tener una larga espera para acceder a múltiples servicios se cataloga como uno de los principales retos de los gerentes, directores y aquellos que toman decisiones para mejorar esa percepción de los usuarios y de esta manera responder a las quejas que a las empresas les permita generar un mejoramiento continuo en el servicio que brindan a sus clientes.

En el contexto de la pandemia de COVID-19, las cadenas de comida rápida han adoptado tecnologías como quioscos de autoservicio, pedidos móviles y aplicaciones de entrega a domicilio para reducir la presión sobre los puntos de venta físicos, similar a la virtualización observada en sectores como las agencias de viajes (Villarreal et al., 2021). Estas innovaciones reflejan la necesidad de optimizar recursos limitados frente a una demanda variable, un problema central en el punto de helados de McDonald's Chapinero, donde las largas filas y tiempos de espera prolongados motivaron esta investigación.

La teoría de colas ha sido ampliamente utilizada en Ingeniería Industrial para describir sistemas en los que se presentan llegadas aleatorias de entidades que requieren atención limitada por uno o varios servidores. Al comprender métricas como el tiempo promedio en cola, la tasa de servicio y la longitud media del sistema, es posible identificar cuellos de botella y proponer mejoras operativas (Gross & Harris, 2008). En este contexto, la distribución Gamma ha demostrado ser útil para modelar tiempos de servicio variables, especialmente cuando el proceso presenta asimetrías o desviaciones significativas respecto a la media (Walpole et al., 2012).

La teoría de colas se ha convertido en un método analítico clave para abordar problemas complejos y dinámicos, permitiendo analizar características específicas como el número medio de entidades en espera y el tiempo medio de permanencia en el sistema según (Camelo et al., 2010). Existen diversos modelos desde el punto de vista de la Investigación de operaciones; éstos se diferencian por características cómo: la distribución que siguen los tiempos entre llegadas y los tiempos de servicio, la población, la disciplina de la cola, la capacidad total del sistema y la cantidad de servidores. Particularmente los del tipo M/M/1, son efectivos para analizar tasas de servicio que se distribuyen exponencialmente relacionadas normalmente con sistemas de atención al cliente (Narváez-Gómez et al., 2018)

Además, la teoría de colas permite cuantificar el equilibrio entre la eficacia en la prestación de servicios y la eficiencia operativa, identificando claramente los parámetros clave para evaluar y optimizar los sistemas de servicio (Singer et al., 2008). De igual manera, la gestión eficiente de las líneas de espera reduce significativamente los tiempos de espera, incrementando la satisfacción del cliente y disminuyendo pérdidas económicas asociadas a la mala calidad del servicio (Villarreal et al., 2021).

La simulación de sistemas de transporte público mediante modelos de redes de colas abiertas permite evaluar el desempeño integral del sistema, incluyendo la estimación precisa de tiempos de espera y longitud de filas (Ortiz & Serrano, 2006). Finalmente, la correcta aplicación de la teoría de colas y simulación en instituciones financieras ha demostrado ser efectiva para mejorar procesos críticos, como los tiempos de atención al cliente (Gómez, 2008).

La simulación computacional es otra herramienta clave para la evaluación de sistemas de servicio. Según Law & Kelton (2014), la simulación permite replicar el comportamiento de procesos reales mediante la construcción de modelos virtuales, facilitando la experimentación sin alterar el entorno físico. La combinación de simulación en lenguajes de programación como Python y entornos gráficos como FlexSim permite tanto el análisis cuantitativo como la visualización tridimensional de flujos operativos.

La simulación dinámica de procesos permite representar sistemas complejos a través de modelos comprensibles, que apoyan decisiones estratégicas y tácticas (Roark et al., 2019). En este mismo

sentido, la simulación de eventos discretos ayuda a determinar requerimientos cuantitativos esenciales, considerando la aleatoriedad de los procesos y sus interacciones (Pérez & Riaño, 2007).

Además, la incorporación de técnicas de análisis estadístico apoyadas en software como SAS Studio ha permitido mejorar la calidad del ajuste de datos recolectados en campo. Mediante pruebas de bondad de ajuste como Kolmogorov-Smirnov, Cramér-von Mises y Anderson-Darling, es posible validar la representatividad de una distribución teórica frente a datos empíricos, fortaleciendo la credibilidad del modelo (SAS Institute Inc., 2024).

Llorente et al. (2001), también resaltan el valor de la simulación en sistemas de atención, subrayando cómo aumentar recursos disponibles no siempre implica mejoras significativas en tiempos de espera, destacando así la importancia de modelos precisos para decisiones operativas eficientes.

Diversos estudios aplicados al sector de alimentos han demostrado que la simulación de sistemas de atención en puntos de venta permite identificar con claridad los momentos críticos del día, optimizar la asignación de personal y mejorar la experiencia del cliente (Toneguzzi, 2022; Corral González, 2024). En el caso particular de McDonald's, su modelo operativo altamente estandarizado constituye un entorno ideal para aplicar herramientas de ingeniería industrial con fines de mejora continua.

McDonald’s es una de las cadenas de comida rápida más reconocidas a nivel mundial, destacada por su eficiencia operativa y estandarización de procesos (Corral, 2024). La sede ubicada en Chapinero, Bogotá, fue seleccionada como caso de estudio por su ubicación estratégica en una zona comercial y estudiantil con alto flujo peatonal. Para atender esta demanda, se implementó un punto de venta exclusivo para helados hacia la calle, con productos como McFlurrys, conos y sundaes (Toneguzzi, 2022). Sin embargo, esta estrategia generó largas filas y tiempos de espera considerables, lo cual motivó esta investigación. El objetivo fue caracterizar el rendimiento del servicio mediante herramientas de inteligencia artificial aplicadas a la ingeniería industrial. Se utilizó el software SAS para analizar estadísticamente los tiempos entre llegadas y atención al cliente (SAS Institute, 2025). Posteriormente, se desarrolló una simulación en Python (Google Colab) para estimar indicadores como tiempo promedio de espera, tasa de servicio y saturación del sistema. Finalmente, se integró una simulación en FlexSim para visualizar el proceso, detectar cuellos de botella y evaluar escenarios operativos alternativos.

Metodología

La presente investigación se desarrolló bajo un enfoque cuantitativo, utilizando técnicas de análisis estadístico y simulación computacional, con el fin de evaluar y proponer mejoras en el rendimiento operativo del punto de venta exclusivo de postres en McDonald’s Chapinero, Bogotá. La simulación, como herramienta clave en la investigación de operaciones, permitió modelar la complejidad del

sistema de atención al cliente, siguiendo los principios descritos por Banks et al. (2010), quienes destacan su capacidad para estudiar interacciones internas y experimentar con cambios operativos. De acuerdo con Law (2019), el diseño de modelos de simulación requiere una estructura sistemática que integre la recolección de datos, el análisis estadístico y la experimentación controlada, aspectos que guiaron el desarrollo del presente estudio. En esta misma línea, Portilla et al. (2010) subrayan que el diseño metodológico debe considerar la variabilidad del entorno operativo para garantizar que los resultados de la simulación sean representativos y útiles para la toma de decisiones.

El trabajo se desarrolló en tres fases principales: recolección de datos, ajuste estadístico y simulación computacional, descritas a continuación.

Recolección y análisis de datos

La fase inicial consistió en una observación directa del proceso de atención, recolectando datos sobre los tiempos entre llegadas y los tiempos de servicio (en caja y máquina de postres). La recolección se llevó a cabo durante el mes de marzo de 2025, con la participación de dos observadores capacitados que registraron manualmente los tiempos utilizando cronómetros digitales y formatos pre estructurados diseñados para el estudio. Dichos formatos incluían columnas para registrar: hora exacta de llegada del cliente, inicio y fin de atención en caja, inicio y fin de preparación en máquina de postres, y hora de salida del sistema.

Las jornadas de observación incluyeron tres días laborales, con el objetivo de capturar tanto el comportamiento regular del sistema como las fluctuaciones de demanda en días de alta afluencia. Los horarios monitoreados abarcaron desde las 11:00 a.m. hasta las 4:00 p.m., priorizando las horas pico (12:00 p.m. a 2:00 p.m.), identificadas previamente como las más críticas por la gerencia del local. Según Raj (2025), incorporar picos horarios en modelos de colas permite representar de manera más realista la variabilidad de la demanda y proponer intervenciones temporales (*staggered start times*) que mejoran la precisión y aplicabilidad del análisis. Esta evidencia respalda la importancia de incluir franjas horarias intercaladas y contextos de alta demanda en los estudios de líneas de espera.

Para garantizar la calidad de los datos, se adoptó un protocolo de doble registro, en el cual ambos observadores anotaron de manera independiente los mismos eventos y posteriormente los registros fueron validados mediante comparación cruzada, corrigiendo discrepancias mayores a más o menos 2 segundos. Además, los observadores fueron capacitados para minimizar sesgos, evitando interacciones con el personal y los clientes durante las mediciones, garantizando que los datos reflejaran el funcionamiento natural del sistema.

Este proceso permitió capturar promedios, variabilidad y patrones temporales en los datos, elementos indispensables para una simulación precisa. Como subrayan Villarreal et al. (2021), “un problema central en muchos contextos de servicios es la administración del tiempo de espera”,

razón por la cual fue clave medir con precisión variables como tiempo en cola, tiempo en el sistema y tiempos de servicio en cada etapa.

Ajuste estadístico mediante SAS

En la segunda fase, los datos recolectados fueron analizados con el software SAS 9.4, siguiendo un flujo estructurado:

- Carga y depuración de datos: Los registros fueron importados y se verificó la ausencia de datos atípicos o inconsistentes.
- Estimación de parámetros: Para cada conjunto de datos (tiempos entre llegadas y tiempos de servicio), se estimaron los parámetros de las distribuciones candidatas utilizando el método de máxima verosimilitud.
- Pruebas de bondad de ajuste: Se aplicaron tres pruebas ampliamente utilizadas en investigación operativa:
    - *Kolmogorov-Smirnov (K-S):* Detecta diferencias máximas entre la distribución empírica y la teórica.
    - *Anderson-Darling (A^2):* Da mayor peso a los extremos, útil en distribuciones con colas pesadas.
    - *Cramér-von Mises (W^2):* Mide discrepancias promedio entre las dos distribuciones.

Se evaluaron las distribuciones Poisson, Beta, Exponencial, Gamma, Lognormal y Weibull, tanto para los tiempos entre llegadas como para los tiempos de servicio. Este conjunto de pruebas permitió documentar con precisión el ajuste estadístico de los datos y sentar las bases para la selección posterior de las distribuciones más representativas. Tal como indican Walpole et al. (2012), la correcta identificación de las distribuciones subyacentes es fundamental para la validez de los modelos de simulación.

Cabe destacar que la interpretación de los valores p y la selección final de las distribuciones se abordaron en la sección de resultados, manteniendo esta fase como puramente metodológica y libre de inferencias.

Simulación computacional

Con los parámetros derivados del análisis estadístico, se desarrollaron dos modelos complementarios de simulación:

- Simulación en Python (Google Colab): Se programó un sistema de colas para estimar métricas como utilización del servidor, longitud promedio de cola y tiempo promedio en el sistema. Se modelaron dos escenarios: M/G/1 (un servidor) y M/G/2 (dos servidores), incorporando las distribuciones obtenidas en la fase estadística. Este enfoque permitió realizar cálculos analíticos rápidos y comparar configuraciones del recurso humano.

- Simulación en FlexSim: Se creó un modelo tridimensional que replicó el proceso real de atención. El modelo incluyó elementos cómo:
  - Source: Para representar las llegadas de clientes según los tiempos de interarribo observados.
  - Queue: Para las colas de caja y máquina de postres, midiendo longitud y tiempo de espera.
  - Processor: Uno para la caja (registro y cobro de pedidos) y otro para la máquina de postres (preparación y entrega).
  - Sink: Para la salida del sistema tras la atención.

El operador principal fue configurado con funciones multitarea, simulando su desplazamiento entre caja y máquina, mientras que un segundo operador fue incorporado dinámicamente en las horas pico mediante un trigger de cambio de estado que modificaba su disponibilidad entre las 12:00 p.m. y las 2:00 p.m. Esta decisión metodológica se fundamenta en el enfoque de programación flexible propuesto por Zhang & Guhathakurta (2020), quienes destacan la eficacia del personal multihabilidad para mitigar cuellos de botella en entornos con demanda fluctuante.

El uso combinado de modelos analíticos en Python y modelos dinámicos en FlexSim responde a lo señalado por Negahban & Smith (2014), quienes recomiendan la integración de simulaciones de diferente naturaleza para obtener una visión más completa del sistema: mientras los modelos analíticos permiten cálculos rápidos y evaluación de escenarios, los modelos visuales facilitan la identificación de cuellos de botella y la validación empírica del flujo operativo.

Con estas tres fases, el estudio estableció una metodología sólida, combinando observación empírica, análisis estadístico y modelado computacional, creando un marco robusto para la evaluación del desempeño operativo del sistema y la formulación de estrategias de mejora.

Resultados

A continuación, se presentan los principales resultados:

Análisis de distribuciones

Inicialmente se realizó la recolección de datos. Se hizo observación directa durante el mes de Marzo de 2025, de los datos de hora de llegada de un cliente a la fila, del tiempo de inicio de atención en caja, del tiempo de servicio en caja y de despacho del helado.

Con base en esta información se hicieron los cálculos de tiempos entre llegadas y tiempos de servicio. Para el análisis de distribuciones que siguen estos tiempos en el sistema de colas del mostrador de postres de McDonald's Chapinero, se evaluaron varias distribuciones teóricas usando SAS Studio: Beta, Exponencial, Gamma, Lognormal y Weibull. Se aplicaron las pruebas de

bondad de ajuste Kolmogorov-Smirnov, Cramer-von Mises y Anderson-Darling para determinar la mejor adaptación de los datos.

Tabla 1

Distribución Gamma ajustada para le tiempo de servicio

Distribución Gamma ajustada para Tiempo de servicio (Tiempo de servicio)

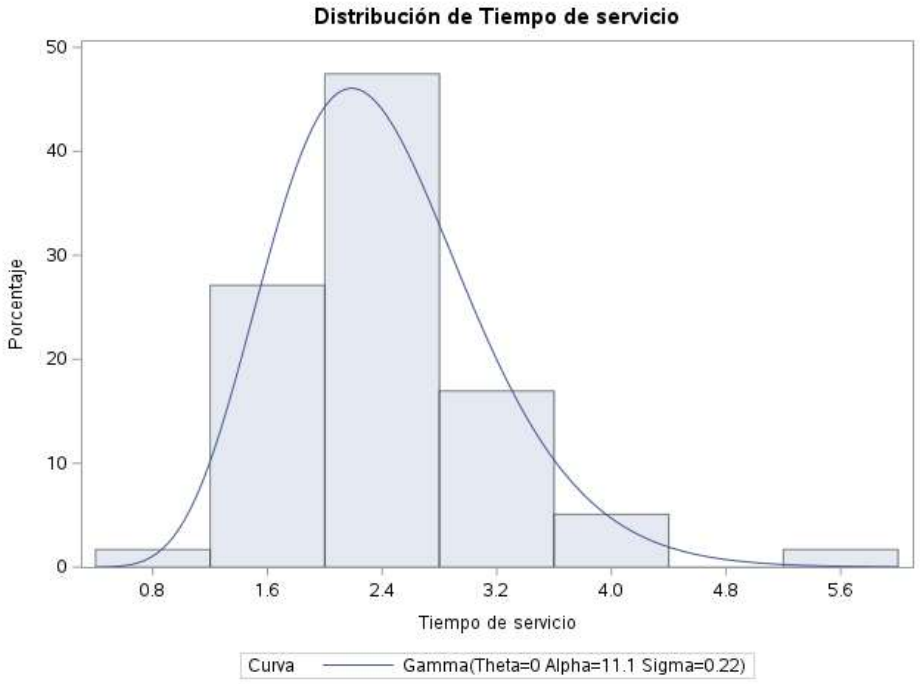
Tests de bondad de ajuste para la distribución Gamma				
Test	**Estadístico**		**P valor**	
Kolmogorov-Smirnov	**D**	0.08859021	**Pr > D**	>0.250
Cramer-von Mises	**W-Sq**	0.08991075	**Pr > W-Sq**	0.156
Anderson-Darling	**A-Sq**	0.58617865	**Pr > A-Sq**	0.132

Fuente: Elaboración propia.

La distribución Gamma mostró un buen ajuste, con valores p superiores a 0.05 en todas las pruebas (K-S: 0.250, W^2: 0.156, A^2: 0.132), lo que indica que no se rechaza la hipótesis de que los datos provienen de esta distribución. Por lo tanto, la distribución Gamma es adecuada para modelar el tiempo de servicio en este sistema de colas.

Figura 2

Distribución Gamma ajustada para el tiempo de servicio



Fuente: Elaboración propia.

La distribución Lognormal también presentó un buen ajuste a los datos de tiempo de servicio. Los resultados de las pruebas de bondad de ajuste fueron:

- Kolmogórov-Smirnov: D = 0.0863, Pr > D = > 0.150
- Cramer-Von Mises: W^2 = 0.0785, Pr > W^2 = 0.220
- Anderson-Darling: A^2 = 0.5642, Pr > A^2 = 0.142

Dado que los valores p en todas las pruebas son superiores a 0.05, no se rechaza la hipótesis nula, lo que indica que la distribución Lognormal también modela adecuadamente el tiempo de servicio.

Tabla 2
Distribución Lognormal ajustada para el Tiempo de Servicio

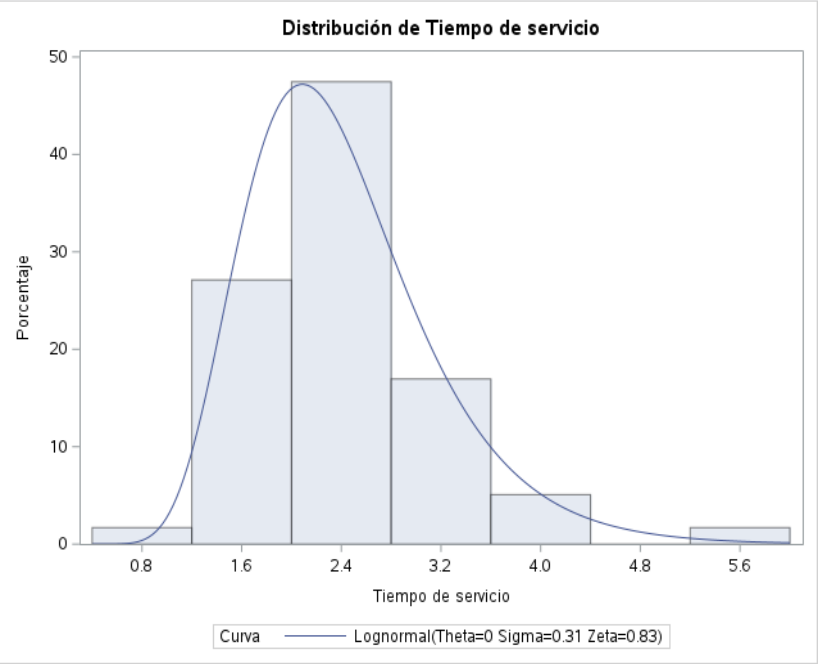
Distribución Lognormal ajustada para Tiempo de servicio (Tiempo de servicio)

Tests de bondad de ajuste para la distribución Lognormal				
Test	Estadístico		P valor	
Kolmogorov-Smirnov	D	0.08630562	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.07846358	Pr > W-Sq	0.220
Anderson-Darling	A-Sq	0.56424506	Pr > A-Sq	0.142

Fuente: Elaboración propia.

Figura 2

Distribución Lognormal ajustada para el Tiempo de Servicio



Fuente: Elaboración propia.

Las distribuciones Exponencial, Beta y Weibull no cumplieron con los criterios mínimos de ajuste para modelar el tiempo de servicio, ya que en todos los casos el test de Anderson-Darling arrojó valores p menores a 0.05, lo que indica diferencias significativas entre los datos reales y las distribuciones teóricas.

En cuanto al tiempo entre llegadas, se evaluaron las mismas distribuciones (Beta, Exponencial, Gamma, Lognormal y Weibull), pero ninguna presentó un ajuste adecuado, ya que todas obtuvieron

valores p inferiores a 0.05 en el test de Anderson-Darling. Esto sugiere que el comportamiento del tiempo entre llegadas no sigue un patrón probabilístico tradicional.

Análisis de la simulación

Comenzando con las simulaciones, la simulación en Python, implementada en Google Colab, permitió analizar el desempeño del sistema de colas del punto de venta de helados de McDonald's Chapinero, utilizando los datos recolectados durante marzo de 2025. A partir de una muestra de 59 clientes, se calcularon métricas iniciales del sistema: un total de 59 clientes atendidos, un tiempo promedio en cola de 1.87 minutos, un tiempo promedio en el sistema de 4.28 minutos, un tiempo promedio de servicio de 2.40 minutos, y un tiempo promedio entre llegadas de 6.74 minutos. Estas métricas reflejaron la congestión observada en el punto de venta, especialmente durante las horas pico (12:00 p.m. a 2:00 p.m.), motivando la simulación para evaluar mejoras operativas.

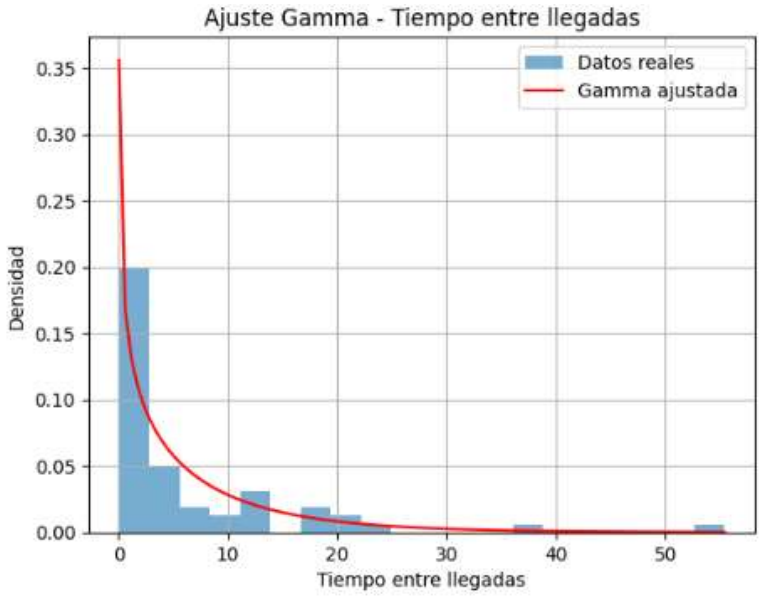
Figura 3
Métricas reales del sistema

```
Métricas generales del sistema:
Clientes totales: 59
Tiempo promedio en cola: 1.87
Tiempo promedio en el sistema: 4.28
Tiempo promedio de servicio: 2.40
Tiempo promedio entre llegadas: 6.74
```

Fuente: Elaboración propia en base a los datos.

El ajuste de la distribución Gamma se realizó para modelar los tiempos entre llegadas y los tiempos de servicio. Para los tiempos entre llegadas, se obtuvieron parámetros de shape (k) = 0.7051 y scale (θ) = 9.7210, indicando una alta variabilidad en las llegadas de clientes. La Figura 4 muestra el histograma de los datos reales junto con la distribución Gamma ajustada, confirmando un ajuste visual adecuado, aunque menos robusto que para los tiempos de servicio, como se observó en el análisis con SAS Studio.

Figura 1
Distribución Gamma ajustada para el Tiempo entre Llegadas

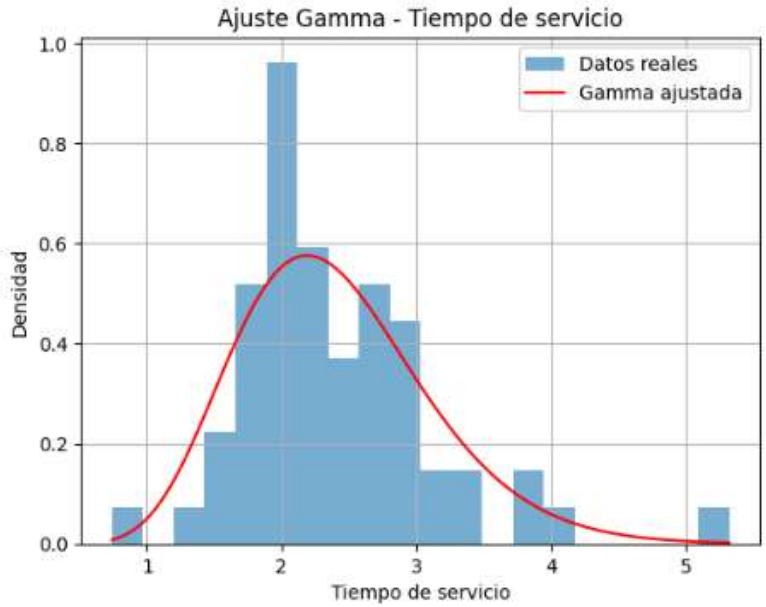


Fuente: Elaboración propia

Para los tiempos de servicio, se ajustó una distribución Gamma, cuyos parámetros se derivaron de los datos recolectados, permitiendo modelar la preparación de helados (por ejemplo, McFlurrys y conos) en el sistema de colas. La Figura 5 presenta el histograma correspondiente, mostrando una buena correspondencia entre los datos reales y la distribución teórica.

Figura 2

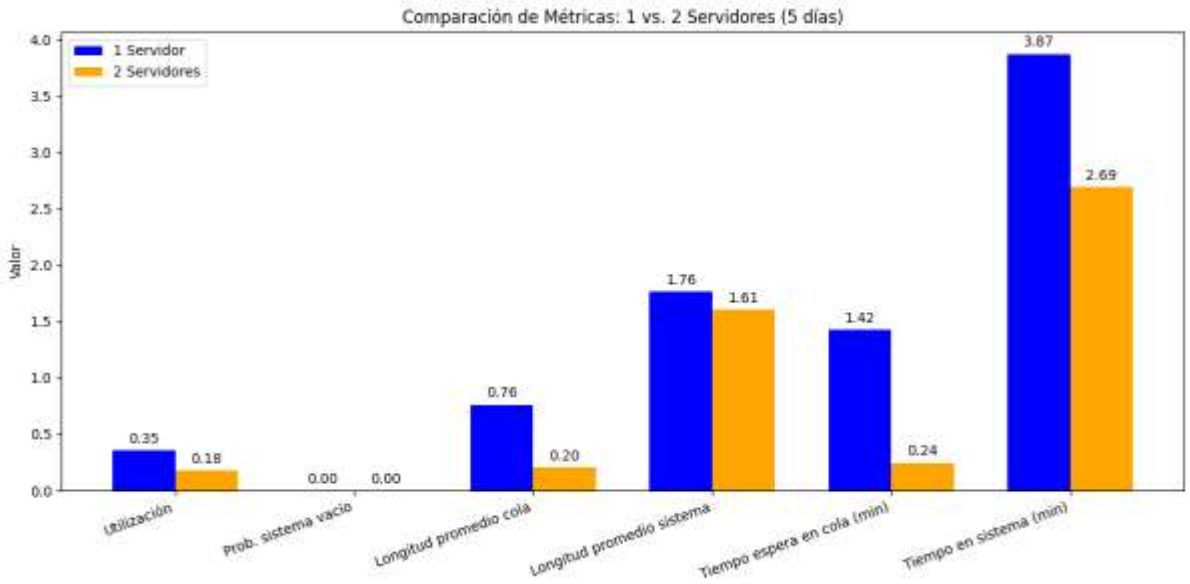
Distribución Gamma ajustada para el Tiempo de Servicio



Fuente: Elaboración propia

La Figura 6 ilustra una comparación de las métricas clave derivadas de la simulación en Python, ejecutada durante una semana laboral de 5 días con turnos de 8 horas, representando un turno típico en Colombia, comparando los escenarios de un servidor y dos servidores. Esta figura incluye la utilización del servidor, la longitud promedio de la cola, el tiempo promedio de espera en cola y el tiempo promedio en el sistema, ofreciendo una visión clara de las diferencias operativas entre ambas configuraciones. A continuación, se analizan estos datos para evaluar el impacto en el desempeño del sistema de colas del punto de venta de helados de McDonald's Chapinero.

Figura 3.
Comparación métricas principales M/G/1 vs M/G/2



Fuente: Elaboración propia en base a la simulación de Google Colab

Para el caso de un servidor, la utilización del 34% indica que el servidor está ocupado solo una tercera parte del tiempo, lo que podría reflejar una demanda intermitente o una capacidad subutilizada, aunque la probabilidad de sistema vacío del 0% sugiere que el sistema nunca estuvo sin clientes, manteniendo una actividad constante. La longitud promedio de la cola de 0.76 clientes y el tiempo promedio de espera en cola de 1.47 minutos, junto con un tiempo promedio en el sistema de 3.92 minutos y una longitud promedio del sistema de 1.76 clientes, muestran un desempeño moderado, con congestión evidente en horas pico que requiere atención.

Con la implementación de dos servidores, la simulación revela una utilización por servidor del 17%, lo que indica una distribución más equilibrada de la demanda, reduciendo la carga individual, mientras que la probabilidad de sistema vacío sigue siendo 0%, confirmando una demanda sostenida durante las 8 horas. La longitud promedio de la cola disminuye a 0.12 clientes y el tiempo de espera en cola se reduce drásticamente a 0.18 minutos, con un tiempo promedio en el sistema

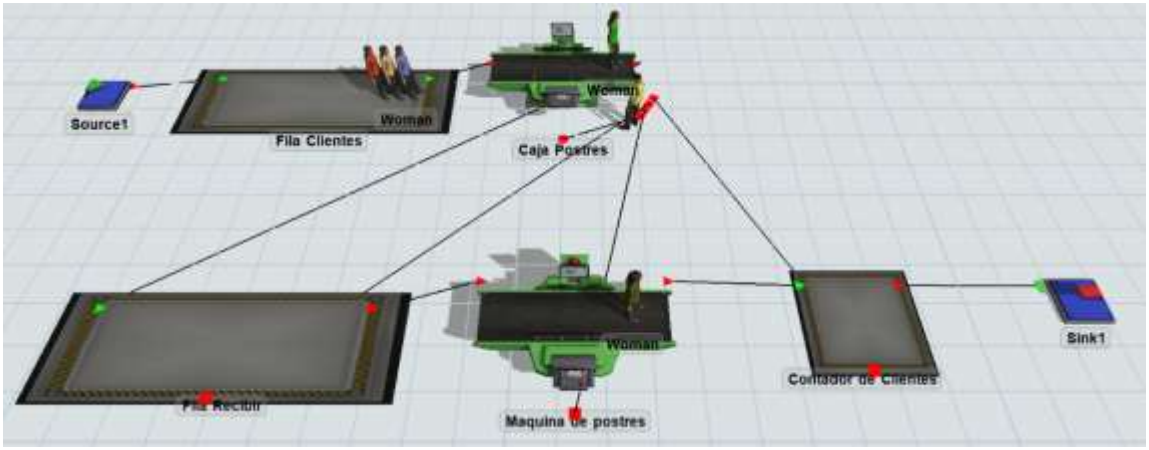
de 2.61 minutos y una longitud promedio del sistema de 1.59 clientes, reflejando una mejora significativa en el flujo y la experiencia del cliente, especialmente al mitigar cuellos de botella en momentos de alta afluencia.

La comparación entre ambos escenarios destaca que añadir un segundo servidor optimiza el sistema al reducir los tiempos de espera y la longitud de la cola, aunque la menor utilización por servidor sugiere que esta configuración podría ser excesiva si la demanda no justifica la inversión adicional. La ausencia de períodos vacíos en ambos casos subraya la consistencia de la demanda, mientras que la notable disminución de la espera con dos servidores valida esta estrategia como una solución efectiva para las horas pico.

Una vez identificadas las distribuciones, se procedió a modelar el sistema en el software FlexSim. El modelo base es el siguiente:

Figura 4.

Modelo de simulación en FlexSim



Fuente: Elaboración propia

En esta simulación se modelo la operación actual, un solo operador debe asumir dos funciones clave: la recepción de pedidos y cobro en la caja de postres, así como la preparación y dispensación del producto en la máquina correspondiente, Para el desarrollo de esta simulación se utilizaron diversos elementos que permiten representar el flujo operativo de la venta de postres en McDonald's. En primer lugar, se incorporó Source1, que funciona como la fuente de clientes. Este elemento genera la llegada de clientes al sistema y permite configurar la tasa de llegada, como el tiempo entre clientes, lo que facilita analizar el comportamiento del sistema bajo diferentes niveles de demanda.

Posteriormente, los clientes se dirigen a la Fila Clientes, que representa la cola de espera para la caja de postres. Este componente permite observar el tamaño de la fila y los tiempos de espera

antes de que los clientes sean atendidos, reflejando situaciones reales en las que se pueden formar cuellos de botella.

La atención inicial se lleva a cabo en el *Processor* denominado Caja Postres, donde un operador recibe el pedido y realiza el cobro del producto. Este elemento simula la interacción directa con el cliente y marca el inicio del proceso de servicio.

Una vez tomada la orden, los clientes pasan a la Fila Recibir, que representa el espacio donde esperan mientras el operador se desplaza a la máquina de postres para preparar el producto. Este elemento refleja la necesidad de gestión de tiempos y la coordinación del operador para atender múltiples tareas.

La preparación del producto se realiza en el *Processor* denominado Máquina de Postres, donde el mismo operador que tomó el pedido se encarga de dispensar el postre solicitado. Esta configuración permite analizar el impacto de la multitarea del empleado en los tiempos de atención y el flujo general del sistema.

Asimismo, se incluyó un Contador de Clientes, que funciona como un nodo de control encargado de registrar la cantidad de clientes atendidos. Esto facilita la medición del desempeño del sistema y el seguimiento de los resultados de la simulación.

Finalmente, los clientes que han recibido su producto pasan a Sink1, que representa la salida del sistema. Este elemento indica el cierre del proceso de atención, contabilizando los clientes que completaron satisfactoriamente el servicio.

Luego se realizó la configuración de todas las instancias de simulación incluida tasa de llegada, la tasa de servicio, los servidores activos y la lógica de flujo del sistema.

La jornada de simulación del sistema de colas en el mostrador de postres de McDonald's Chapinero, comprendió un horario laboral de 10:00 a.m. a 6:00 p.m., se analizaron los niveles de atención y salida de clientes a través de tres elementos clave: la caja de postres, la máquina de postres y el contador de clientes.

Tabla 1.

Tabla de Throughput – Clientes que salieron del sistema

Object	Throughput
Caja Postres	68
Maquina de postres	63
Contador de Clientes	63

Fuente: Elaboración propia.

Según los datos obtenidos en la Tabla 3, la caja de postres atendió un total de 68 clientes, mientras que 63 clientes fueron procesados por la máquina de postres y el contador final de salida también registró 63 clientes. Esta diferencia evidencia un desajuste entre los clientes que ingresan al sistema y los que logran completarlo.

La diferencia de cinco clientes entre la caja de postres y la máquina de postres sugiere la existencia de un cuello de botella en la máquina de postres, lo cual podría estar generando acumulación de clientes o incluso pérdida de algunos que no logran ser atendidos por completo.

A partir del *throughput* registrado, se estima que el sistema procesó, en promedio, un cliente cada 7.6 minutos, lo cual puede evaluarse frente a los datos reales para validar la fidelidad del modelo.

Posterior a ello se realizó la tabla 4 que presenta el análisis del flujo horario de clientes en el sistema de colas simulado.

Tabla 2

Tabla de Throughput – Clientes que salen del sistema cada hora

Object	Throughput
Caja Postres	8.50
Fila Clientes	8.50
Contador de Clientes	7.88

Fuente: Elaboración propia.

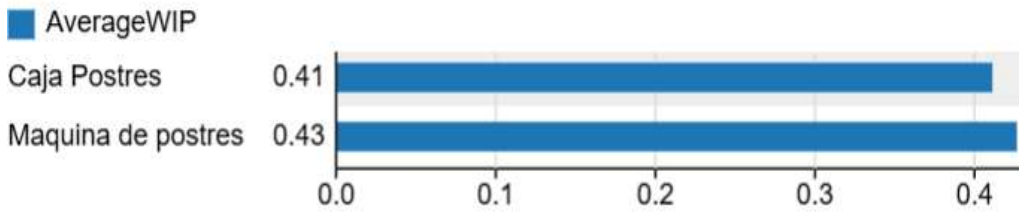
En este caso, se observa que tanto la caja de postres como la fila de clientes mantienen un *throughput* promedio de 8.50 clientes por hora, mientras que el contador de clientes, encargado de registrar las salidas del sistema, presenta un valor ligeramente inferior de 7.88 clientes por hora. Esta diferencia de 0.62 clientes por hora entre la caja, la fila y el contador de salida refuerza la evidencia de un cuello de botella o retención en la etapa final del servicio, probablemente en la máquina de postres, la cual actúa como intermediario entre la caja y la salida. Aunque los clientes son atendidos inicialmente con regularidad, el sistema no logra mantener esa eficiencia hasta el final del proceso, lo que ocasiona una acumulación temporal de personas en esta fase.

Este hallazgo es coherente con las tendencias observadas en los análisis de utilización de recursos, particularmente en la máquina de postres, que presenta picos de hasta el 50% de utilización en las horas de mayor demanda. La caída en el *throughput* entre la caja y el contador sugiere que el flujo de salida del sistema no depende exclusivamente del tiempo de atención inicial, sino que está condicionado por el desempeño del recurso encargado de la preparación y entrega del producto. Esto coincide con estudios previos sobre cuellos de botella en servicios multietapa, como los

descritos por Negahban & Smith (2014), quienes destacan que la etapa más lenta del proceso determina el ritmo de salida global del sistema.

Además, se analizó el promedio de clientes en el sistema, como se muestra en la Figura 8. Se observó que la máquina de postres concentra la mayor cantidad de personas en promedio (0.43), mientras que la caja de postres tiene un promedio de 0.41. Este resultado confirma que la máquina de postres representa un posible cuello de botella, alineándose con los resultados anteriores sobre tiempos de espera y con las simulaciones que señalan la necesidad de reforzar este punto del proceso. Al integrar estos indicadores *throughput,* utilización de recursos y promedio de clientes se obtiene una visión más completa de cómo la saturación de una etapa crítica afecta el desempeño global, reforzando la propuesta de implementar estrategias de asignación dinámica de personal, particularmente en los horarios de máxima afluencia.

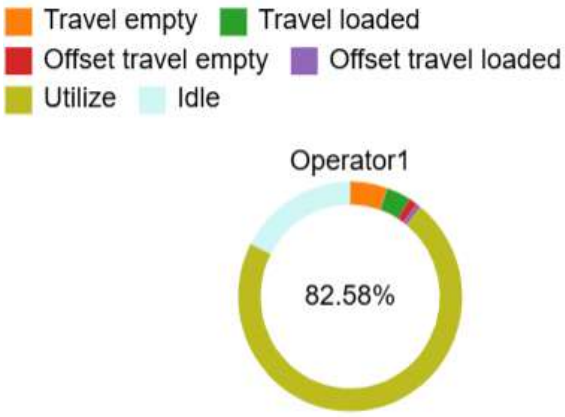
Figura 5.
Promedio de clientes en el sistema



Fuente: Elaboración propia.

Conociendo el comportamiento general del sistema, se procedió a realizar un análisis detallado del desempeño del operario, que constituye uno de los recursos críticos del punto de venta. En la Figura 9 se observa que la utilización del operador alcanza un 82,58%, lo cual indica que durante la mayor parte de su jornada laboral estuvo realizando actividades productivas directamente relacionadas con el proceso de atención al cliente. Este nivel de utilización es particularmente elevado si se compara con los estándares de carga recomendados para operaciones de servicios, donde valores cercanos al 80% suelen considerarse indicadores de sobrecarga operativa (Negahban & Smith, 2014). La elevada ocupación refleja una importante presión sobre el recurso humano, atribuida principalmente a la saturación del servicio en los horarios de mayor afluencia de clientes. Dicho comportamiento coincide con los picos de atención registrados entre las 12:00 p.m. y las 2:00 p.m., periodo que fue previamente identificado como crítico tanto por las observaciones realizadas en campo como por las entrevistas informales con el personal del establecimiento. Este hallazgo no solo permite dimensionar la magnitud del esfuerzo requerido por el operario, sino que también subraya la urgencia de explorar alternativas de redistribución de la carga de trabajo o incorporación de recursos adicionales en los momentos de máxima demanda. Además, esta información aporta insumos valiosos para el diseño de estrategias de programación dinámica del personal, orientadas a equilibrar la carga operativa y evitar el desgaste físico y mental de los trabajadores.

Figura 6.
Utilización del operario

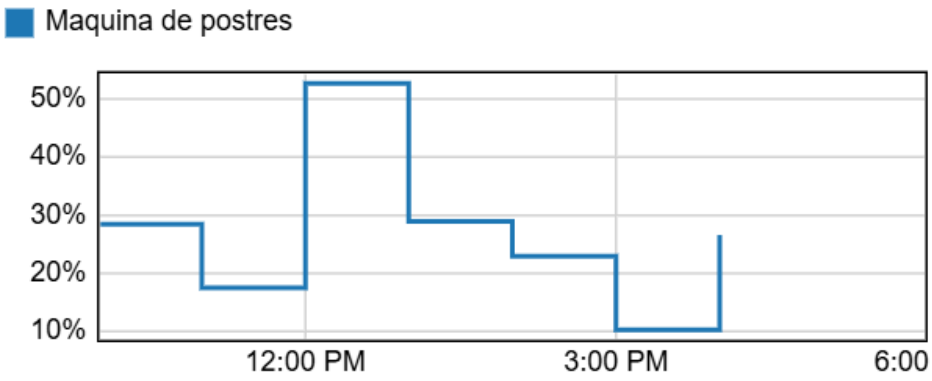


Fuente: Elaboración propia.

Dado que la máquina de postres constituye el principal cuello de botella dentro del sistema, se realizó un análisis pormenorizado de su utilización por intervalos horarios, presentado en la Figura 10, con el objetivo de proponer estrategias fundamentadas para la asignación de un operario de apoyo durante las horas de mayor carga de trabajo. La gráfica revela que la utilización de la máquina de postres presenta un patrón dinámico con variaciones significativas a lo largo de la jornada: inicia en un 30% entre las 12:00 p.m. y la 1:00 p.m., desciende temporalmente al 20%, y alcanza un pico máximo del 50% entre la 1:00 p.m. y las 2:00 p.m., periodo que coincide con el mayor flujo de clientes identificado en la fase de observación. Posteriormente, la utilización disminuye nuevamente al 30% de las 2:00 p.m. a las 3:00 p.m., experimenta un leve repunte al 35% entre las 3:00 p.m. y las 4:00 p.m., y finalmente se reduce drásticamente al 10% en el bloque comprendido entre las 4:00 p.m. y las 6:00 p.m. Este patrón de comportamiento confirma que el sistema enfrenta un pico operativo particularmente significativo al mediodía, seguido por una disminución progresiva hacia el final de la jornada.

Tales resultados sugieren la necesidad de reforzar los recursos operativos, específicamente mediante la incorporación de un segundo operario de apoyo en el intervalo de 1:00 p.m. a 2:00 p.m., cuando la máquina experimenta su máxima carga. La identificación de este intervalo crítico constituye un insumo esencial para fundamentar estrategias de programación flexible del personal, como las propuestas por Zhang & Guhathakurta (2020), cuyo objetivo es ajustar dinámicamente los recursos disponibles en función de la demanda real, optimizando así la eficiencia del servicio sin incurrir en costos excesivos durante las horas de baja afluencia.

Figura 7.
Utilización de la Maquina de postres por hora



Fuente: Elaboración propia.

La Figura 10 muestra el comportamiento horario de la utilización de la máquina de postres durante la jornada observada, lo cual permite identificar con claridad los momentos de mayor presión sobre este recurso. Tal como se aprecia, la utilización mantiene valores moderados durante el inicio del día, con una tasa cercana al 30% entre las 11:00 a.m. y el mediodía, lo que refleja un flujo de clientes estable y manejable. Sin embargo, al llegar al intervalo comprendido entre las 12:00 p.m. y las 2:00 p.m., la utilización se eleva drásticamente hasta alcanzar picos cercanos al 50%, evidenciando una concentración significativa de la demanda en este horario crítico. Posteriormente, la carga operativa disminuye progresivamente, oscilando entre el 10% y el 30%, con descensos notables después de las 3:00 p.m., lo que indica el retorno a un escenario de baja demanda.

Este patrón respalda los hallazgos obtenidos en la fase de recolección de datos y análisis estadístico, donde se identificó que el periodo del mediodía constituye el cuello de botella operativo más relevante del sistema. En este sentido, el incremento abrupto de la utilización durante el mediodía refleja no solo una mayor afluencia de clientes, sino también limitaciones en la capacidad de respuesta del recurso humano asignado a esta tarea, particularmente porque un único operario debe atender de forma simultánea la toma de pedidos en la caja y la preparación de los postres en la máquina.

A partir de este diagnóstico, se diseñó en FlexSim una propuesta de mejora orientada a optimizar el flujo de atención, incorporando un segundo operario exclusivamente durante el intervalo de mayor afluencia (12:00 p.m. a 2:00 p.m.). Esta estrategia, representada en la figura siguiente, consiste en la programación dinámica del operario adicional, quien permanece inactivo en los horarios de baja demanda (indicado con una camisa de color rojo) y pasa a desempeñar funciones activas en la máquina de postres durante el intervalo crítico (cambiando su camisa a color amarillo), reforzando así la capacidad del sistema.

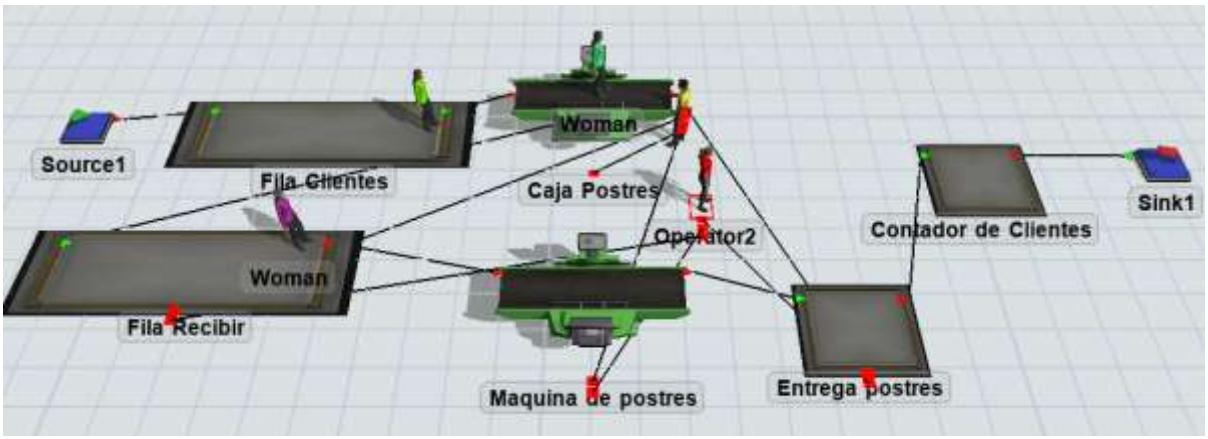
La justificación de esta intervención radica en que una utilización del 50% en un único recurso clave, como la máquina de postres, puede comprometer la fluidez del servicio y prolongar los tiempos de espera, generando insatisfacción en los clientes. Al incluir un segundo operario, se busca distribuir

la carga de trabajo, reducir el tiempo de permanencia de los clientes en el sistema y prevenir la saturación de las operaciones en los momentos de mayor presión. Esta medida se enmarca en los principios de programación flexible de personal (Zhang & Guhathakurta, 2020), ampliamente aplicados en entornos de servicios con demanda variable, donde la reasignación dinámica de recursos permite responder de manera más eficiente a los picos operativos.

En síntesis, el análisis de la Figura 10 no solo evidencia el comportamiento de la máquina de postres, sino que también sustenta la necesidad de reforzar el recurso humano en horas pico, proponiendo una solución de ajuste dinámico de personal que optimiza el desempeño del sistema, mejora la atención al cliente y promueve una operación más equilibrada.

Figura 8.

Modelo de simulación propuesta de mejora en FlexSim



Fuente: Elaboración propia.

La introducción de este operario adicional tuvo como objetivo mejorar el flujo de atención y reducir la saturación del sistema en las horas clave. Los resultados de la simulación fueron positivos: se observó una redistribución significativa de la carga de trabajo, donde la utilización del Operador 1 se redujo al 43% y la del Operador 2 se estabilizó en 31,69% como se puede observar en la Figura 12.

Figura 9

Análisis de la propuesta de mejora

Fuente: Elaboración propia.

Esta redistribución no solo disminuye el desgaste del personal, sino que también asegura la disponibilidad del recurso humano para responder de forma más ágil ante picos de demanda.

El análisis de la Figura 12 permite observar que, además de la reducción en los porcentajes de utilización, también disminuyeron los tiempos en estado “Travel loaded” y “Travel empty”, lo cual indica que el desplazamiento operativo fue optimizado. Asimismo, el tiempo en estado “Idle” (inactividad) aumentó levemente, lo que refleja una mayor holgura operativa y capacidad para absorber incrementos inesperados en la demanda sin comprometer la continuidad del servicio.

En conjunto, esta configuración permitió reducir los tiempos de espera de los clientes y mejorar el throughput del sistema, lo que se traduce en una atención más rápida y eficiente. A nivel estratégico, esta mejora favorece un modelo de operación más balanceado y sostenible, en el que los recursos se asignan de manera dinámica según los patrones reales de demanda. Esto no solo impacta la satisfacción del cliente, sino que también contribuye al bienestar del personal, reduciendo la fatiga y el riesgo de errores en jornadas de alta exigencia.

De lo cual podemos comparar que los resultados obtenidos en Python y FlexSim permiten observar con mayor claridad el impacto de las mejoras propuestas sobre el sistema. La simulación en Python proporcionó métricas clave bajo un enfoque analítico (modelos M/G/1 y M/G/2), como el tiempo promedio de espera en cola (1.47 minutos con un servidor y 0.18 minutos con dos servidores), el tiempo promedio en el sistema (3.92 minutos con un servidor y 2.61 minutos con dos servidores) y la utilización del servidor (34% en un servidor y 17% por servidor en el escenario de dos servidores). Estos valores evidencian que la incorporación de un segundo servidor reduce significativamente la saturación, optimizando el flujo de atención.

Para enriquecer este análisis, en la Tabla 6 se presentan comparativamente los principales indicadores obtenidos en la simulación analítica en Python, la simulación dinámica en FlexSim y los valores observados en el sistema real.

Tabla 5.
Comparación de indicadores entre datos reales, Python y FlexSim

Indicador	Datos reales	Python M/G/1	Python M/G/2	FlexSim (1 operario)	FlexSim (2 operarios)
Tiempo promedio en cola (min)	1.87	1.47	0.18	2.1	0.9
Tiempo promedio en el sistema (min)	4.28	3.92	2.61	5.0	3.1
Clientes atendidos (por jornada)	59	65	72	63	70
Utilización del operario (%)	—	34%	17%	82.58%	43.27% (Op1) / 31.69% (Op2)
Throughput (clientes/hora)	—	7.4	9.0	7.88	8.75

Fuente: Elaboración propia con base en los datos del estudio.

Por su parte, la simulación en FlexSim brindó una visión más detallada y visual del proceso operativo, integrando las limitaciones físicas del sistema y las interacciones multitarea del operario. Se identificó un cuello de botella en la máquina de postres, reflejado en la diferencia entre los clientes atendidos en la caja (68) y los que completaron el proceso (63), así como una alta utilización del operario (82.58%), especialmente entre la 1:00 p.m. y las 2:00 p.m., cuando la máquina alcanzó una utilización del 50%. Con la incorporación de un segundo operario entre las 12:00 p.m. y las 2:00 p.m., la utilización del Operador 1 disminuyó al 43% y la del Operador 2 se estabilizó en 31.69%, logrando una redistribución más equilibrada de la carga de trabajo y una mejora sustancial en el flujo de clientes.

En síntesis, mientras que Python proporcionó una aproximación cuantitativa precisa para evaluar diferentes configuraciones del sistema, FlexSim permitió validar esos escenarios en un entorno dinámico tridimensional, identificando cuellos de botella y visualizando los efectos de la multitarea en el desempeño. Ambos enfoques se complementan, ofreciendo una base sólida para la toma de decisiones estratégicas sobre la asignación de personal y la capacidad instalada.

Limitaciones

A pesar del nivel de detalle metodológico alcanzado, es necesario reconocer limitaciones que condicionan la validez y aplicabilidad de los hallazgos. En primer lugar, la representatividad temporal de los datos es limitada: la recolección se restringió a tres jornadas específicas, lo que impide captar las variaciones estacionales o asociadas a eventos especiales. Estudios recientes en el sector retail han demostrado que la demanda varía significativamente según la temporada e incluso en torno a días festivos o campañas promocionales, y que los modelos tradicionales pueden subestimar o ignorar estas fluctuaciones (Huber & Stuckenschmidt, 2020). Así, ampliar el horizonte de observación a diferentes meses y condiciones operativas resulta esencial para garantizar que

los modelos de simulación reflejen fielmente la dinámica real del sistema y no estén sesgados por condiciones excepcionales.

Otro aspecto crítico radica en la configuración del modelo de simulación en FlexSim, el cual se enfocó exclusivamente en la atención de postres, caja, máquina y operario sin integrar otros módulos interdependientes del restaurante, como cocina general o servicio *drive-thru*. Según la literatura especializada, los sistemas de atención al cliente en entornos complejos requieren modelos de colas conectadas, ya que las interacciones entre estaciones pueden generar efectos acumulativos que afectan la validez de las recomendaciones operativas (Siebers et al., 2008). Por lo tanto, futuras investigaciones deberían considerar el diseño de modelos más integrales que representen toda la red operativa del establecimiento.

Asimismo, aunque se realizó un análisis de ajuste paramétrico usando distribuciones clásicas como Poisson, Gamma y Weibull, no se exploraron alternativas más flexibles como modelos no paramétricos o enfoques de aprendizaje automático, especialmente en el caso de las llegadas de clientes. En entornos altamente dinámicos, los patrones de llegada suelen presentar comportamientos que se desvían de las distribuciones estándar, por lo que resulta más adecuado emplear aproximaciones que consideren distribuciones generales o mixtas para representar con mayor precisión la variabilidad del sistema (Chaves & Gosavi, 2022).

Por último, aunque la propuesta de incorporar un segundo operario durante los picos operativos ofrece mejoras en indicadores de eficiencia, no se evaluó su viabilidad desde un enfoque económico. La literatura especializada indica que un análisis costo-beneficio riguroso es indispensable para justificar la implementación de intervenciones operativas, como la adición de recursos humanos en un sistema de atención, pues solo de esta manera se puede evaluar si las mejoras en eficiencia justifican los recursos invertidos. En un estudio reciente, Rotunno et al. (2023) aplicaron simulación por eventos discretos combinada con análisis costo-beneficio en un caso real de logística portuaria, demostrando que, con base en los resultados operativos simulados, se puede determinar priorización de inversiones en equipo o personal según su rentabilidad calculada. Este enfoque permite ponderar no solo el rendimiento operativo sino también el impacto económico de implementar cambios. Por tanto, en futuras investigaciones, al igual que en estudios como el de Rotunno et al. (2023), sería fundamental integrar los resultados de la simulación con una evaluación financiera explícita (por ejemplo, retorno esperado, reducción de costos o valor económico agregado), para garantizar que las recomendaciones operativas no solo mejoren el desempeño, sino que también sean viables desde una perspectiva económica.

En conjunto, estas limitaciones abren claros caminos para investigaciones futuras: ampliar el periodo de recolección de datos, modelar redes interconectadas, explorar métodos predictivos avanzados, validar empíricamente los modelos y realizar análisis económico-operativos integrales. Estas líneas de trabajo fortalecerán significativamente la relevancia y el valor estratégico del estudio

Conclusiones

El uso combinado de herramientas como SAS para el análisis estadístico, Python para el modelado analítico y FlexSim para la simulación dinámica permitió abordar el problema de manera integral, ofreciendo una visión detallada del funcionamiento del sistema de atención en el punto de venta de postres de McDonald's Chapinero. La identificación de la distribución Gamma como la más representativa para los tiempos de servicio fue un hallazgo clave, pues proporcionó un insumo sólido y estadísticamente válido para alimentar los modelos de simulación, garantizando resultados cercanos a la realidad operativa y sustentando las decisiones de mejora con base en evidencia cuantitativa.

La simulación en Python, mediante los modelos M/G/1 y M/G/2, permitió calcular indicadores fundamentales como el tiempo promedio de espera en cola (1.47 minutos en un servidor y 0.18 minutos en dos servidores), el tiempo promedio en el sistema (3.92 minutos con un servidor y 2.61 minutos con dos) y la utilización del servidor (34% en un servidor y 17% en dos). Estos resultados demostraron que la incorporación de un segundo servidor reduce significativamente la saturación del sistema y mejora el flujo de atención, lo que impacta directamente en la experiencia del cliente al disminuir los tiempos de espera.

Por su parte, la simulación en FlexSim permitió validar estos resultados en un entorno tridimensional y dinámico, incorporando las restricciones físicas del sistema y las interacciones multitarea del personal. Los resultados confirmaron la existencia de un cuello de botella en la máquina de postres y una alta utilización del operario (82.58%) en el escenario base. La propuesta de incorporar un segundo operario exclusivamente en el horario crítico (12:00 p.m. – 2:00 p.m.) redujo la utilización del Operador 1 al 43% y distribuyó de manera más equilibrada la carga de trabajo, mejorando el *throughput* del sistema y reduciendo los tiempos de espera. La comparación entre escenarios demuestra que la adición de un segundo operario durante las horas pico reduce en más del 50% el tiempo de espera y aumenta el número de clientes atendidos por jornada, logrando un sistema más eficiente y equilibrado.

Estos resultados ponen de manifiesto el valor de implementar estrategias de asignación dinámica de personal basadas en patrones reales de demanda, lo que contribuye a mejorar tanto la productividad como la experiencia del cliente. No obstante, este estudio no está exento de limitaciones. La recolección de datos se restringió a tres jornadas específicas, lo que puede no capturar toda la variabilidad estacional o los efectos de campañas promocionales. Asimismo, el modelo de simulación se centró únicamente en la operación del punto de postres, sin integrar otras áreas del restaurante, como la cocina principal o el servicio *drive-thru*, cuyas interacciones podrían afectar la dinámica general del sistema. Además, aunque se evaluó la mejora operativa con la incorporación de un segundo operario, no se realizó un análisis costo-beneficio que permita valorar la viabilidad económica de dicha intervención, aspecto que sería clave en la toma de decisiones gerenciales.

Finalmente, este estudio evidencia que la combinación de modelos analíticos y simulación visual no solo facilita la identificación de problemas estructurales, sino que también aporta evidencia sólida para la toma de decisiones estratégicas. La metodología propuesta es escalable y puede aplicarse a otros puntos de venta de McDonald's o a negocios con dinámicas similares, consolidando un modelo de gestión operativa más eficiente, flexible y orientado a la mejora continua. Futuras investigaciones podrían ampliar el horizonte temporal de los datos, integrar modelos de colas interconectadas y realizar análisis económico-financieros que complementen los resultados operativos, fortaleciendo así el valor práctico de las recomendaciones generadas.

Referencias

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2010). *Discrete-Event System Simulation* (4th ed.). Pearson Education.

Camelo, G. R., Coelho, A. S., & Borges, R. M. (2010). Aplicación de la teoría de colas y simulación al embarque de mineral de hierro y manganeso en la terminal marítima de Ponta da Madeira. *Revista Gestão Industrial, 6(3)*, 63–78. https://revistas.utfpr.edu.br/revistagi/article/view/661/532

Chaves, C., Gosavi, A. On general multi-server queues with non-poisson arrivals and medium traffic: a new approximation and a COVID-19 ventilator case study. Operational Research. An International Journal, *22,* 5205–5229. https://doi.org/10.1007/s12351-022-00712-2

Corral, J. A. (2024). *La internacionalización del producto y su análisis en diferentes mercados* [Trabajo de grado, Universidad de Valladolid]. *UVaDOC*. https://uvadoc.uva.es/handle/10324/62292

Gómez, F. A. (2008). Aplicación de teoría de colas en una entidad financiera: Herramienta para el mejoramiento de los procesos de atención al cliente. *Revista Universidad EAFIT, 44(150)*, 51–63. https://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/view/154

Gross, D., & Harris, C. M. (2008). *Fundamentals of queueing theory* (4th ed.). Wiley.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting, 36*(4), 1420–1438. https://doi.org/10.1016/j.ijforecast.2020.02.005

Law, A. M., & Kelton, W. D. (2014). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education.

Law, A. M. (2019). *Simulation modeling and analysis* (5th ed.). McGraw-Hill.

Llorente, S., Puente, F. J., Alonso, M., & Arcos, P. I. (2001). Aplicaciones de la simulación en la gestión de un servicio de urgencias hospitalario. *Emergencias, 13(2)*, 90–96. https://revistaemergencias.org/wp-content/uploads/2023/08/Emergencias-2001_13_2_90-6.pdf

Narváez-Gómez, J. E., Ordoñez-Luna, W. A., & Paz-Ruiz, N. E. (2018). Análisis y simulación de tiempos de espera aplicando teoría de colas en una terminal de transportes. *Publicaciones e Investigación, 12(2)*, 35–43. https://repository.unad.edu.co/handle/10596/29722

Negahban, A., & Smith, J. S. (2014). Simulation for manufacturing system design and operation: Literature review and analysis. *Journal of Manufacturing Systems, 33*(2), 241–261. https://doi.org/10.1016/j.jmsy.2013.12.007

Ortiz, J. E., & Serrano, L. Á. (2006). Simulación de sistemas de transporte público masivo. *Revista Ingeniería e Investigación, 26(1),* 51–57. https://www.redalyc.org/pdf/643/64326106.pdf

Pérez, J. F., & Riaño, G. (2007). Análisis de colas para el diseño de una cafetería mediante simulación de eventos discretos. *Revista de Ingeniería, 1 (25)*, 12–21. https://doi.org/10.16924/revinge.25.2

Portilla, L. M., Arias, L., & Fernández, S. A. (2010). Análisis de líneas de espera a través de teoría de colas y simulación. *Scientia et Technica, 17(46),* 56–61. https://www.redalyc.org/articulo.oa?id=84920977012

Raj, J. (2025). Time-Dependent Queuing Model for Traffic Congestion Using Mt/D/1/K: Simulation and Policy Insights. *arXiv preprint* arXiv:2501.14132. https://doi.org/10.48550/arXiv.2501.14132

Roark, G., Chiodi, F. J., Petesch, G., Pastore, E., & dos Santos, C. H. (2019). Modelaje y simulación computacional con FlexSim de un proceso de despacho y expedición en una industria cementera argentina. *XXXIX Encontro Nacional de Engenharia de Produção*, Sao Paulo, 15 al 18 de octubre, 1–10. DOI:10.14488/ENEGEP2019_TN_STO_292_1648_38325

Rotunno, G., Lo Zupone, G., Cermineo, L., & Fanti, M. P. (2023). Discrete event simulation as a decision tool: A cost–benefit analysis case study. *Journal of Simulation, 18*(3), 378–394. https://www.tandfonline.com/doi/full/10.1080/17477778.2023.2167618

SAS Institute Inc. (2024). *SAS Studio: User's guide*. https://documentation.sas.com/doc/en/webeditorcdc/3.8/webeditorug/titlepage.htm

Singer, M., Donoso, P., & Scheller-Wolf, A. (2008). Una introducción a la teoría de colas aplicada a la gestión de servicios. *Revista ABANTE, 11(2)*, 93–120. https://www.ceop.cl/wp-content/uploads/2010/11/Una-Introducci%C3%B3n-a-la-Teor%C3%ADa-de-Colas.pdf

Siebers, P. O., Aickelin, U., Celia, H., & Clegg, C. W. (2008). Simulating customer experience and word-of-mouth in retail—A case study. *Simulation Modelling Practice and Theory, 16*(9), 913–927. https://arxiv.org/pdf/1003.3784

Toneguzzi, M. (2022, octubre 18). McDonald's Canada expanding walk-up window service concept with pedestrians prioritized. *Retail Insider*. https://retail-insider.com/retail-insider/2022/10/mcdonalds-canada-expanding-walk-up-window-service-concept-with-pedestrians-prioritized/

Villarreal, F. L., Bernal, M. L., & Montenegro, D. I. (2021). Teoría de colas y líneas de espera, un reto empresarial en el mejoramiento continuo de los servicios. *Ciencia Latina Revista Científica Multidisciplinar, 5(5)*, 8418–8421. https://doi.org/10.37811/cl_rcm.v5i5.933

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. E. (2012). *Probability and statistics for engineers and scientists* (9th ed.). Pearson Education.

Zhang, R., & Guhathakurta, P. (2020). Multi-skilled workforce scheduling: A simulation-based approach. *Computers & Industrial Engineering, 149*, 106847. https://doi.org/10.1016/j.cie.2020.106847