
Pontificia Universidad
JAVERIANA
Cali

# SIMULACIÓN DE LÍNEAS DE ESPERA EN UNA EMPRESA DE SERVICIOS PÚBLICOS DEL VALLE DEL CAUCA

**Programa de Maestría en Ingeniería**

**Presentado por:**

**RAFAEL EDUARDO TORRES VARGAS**

**MILTON MAURICIO ZÚÑIGA MOSQUERA**

**Dirigido por:**

**JONNY JAIR PANTOJA DIAZ M.Sc.**

**Pontificia Universidad Javeriana Cali**

**Facultad de Ingeniería y Ciencias**

**Enero de 2025**

**Simulación de líneas de espera en una empresa de servicios públicos del Valle del Cauca**

**Rafael E. Torres Vargas** [1]
**Milton M. Zúñiga Mosquera** [2]

[1]Pontificia Universidad Javeriana Cali, Colombia. correo electrónico: retorres@javerianacali.edu.co
[2] Pontificia Universidad Javeriana Cali, Colombia. correo electrónico: mmzuniga@javerianacali.edu.co

**Resumen:** Los centros de atención presencial en empresas de servicios públicos enfrentan desafíos operativos derivados de alta variabilidad en la demanda y heterogeneidad en los tiempos de servicio. Este estudio desarrolló y validó un modelo de simulación de eventos discretos para analizar el sistema de atención de una empresa del Valle del Cauca, Colombia, que gestiona seis tipos de trámites atendidos por 13 grupos de funcionarios. Utilizando más de 35,000 registros operativos, se ajustaron distribuciones empíricas de llegadas y tiempos de servicio, y se construyó un modelo en FlexSim siguiendo lineamientos STRESS-DIS. El diagnóstico inicial reveló un cuello de botella crítico en "Trámites y Solicitudes" (75% del volumen total), con utilización del 100%, tiempos máximos de espera de 69 minutos y colas de hasta 63 usuarios, mientras otras categorías operaban con capacidad ociosa. Se evaluaron dos rediseños: uno basado en especialización por macroprocesos (Propuesta 1) y otro con funcionarios generalistas (Propuesta 2). Ambas alternativas incrementaron el throughput entre 48% y 49% (de 158 a 234-236 usuarios/día). La Propuesta 1 redujo la saturación de colas de tres a una sola categoría y mantuvo la utilización promedio bajo 70%, logrando mejor equilibrio en la distribución de carga con tiempos de atención menos variables. La Propuesta 2, aunque marginalmente superior en throughput, concentró la carga en dos grupos (utilización >81%), aumentando el riesgo de sobrecarga. Se recomienda implementar la Propuesta 1 en un piloto controlado, dado su balance entre eficiencia, control operativo y sostenibilidad a largo plazo. El modelo desarrollado constituye una herramienta reutilizable para evaluar escenarios futuros ante cambios en demanda o políticas de servicio.

**Palabras clave:** Simulación de eventos discretos, teoría de colas, tiempos de espera, eficiencia operativa, cuellos de botella

## 1. Introducción

Se requiere abordar el problema de los altos tiempos de atención de clientes y elevada demanda de solicitudes presenciales en los puntos de atención, con el fin de plantear una propuesta óptima y eficiente que evite deterioro de imagen institucional y mejore la satisfacción de los usuarios. Dichas propuestas no deben implicar un incremento del personal actual.

En los centros de atención de servicios públicos, la experiencia de espera se caracteriza por ser prolongada e impredecible. Los usuarios no llegan de forma uniforme a lo largo del día; por el contrario, se observan periodos de baja afluencia alternados con picos de congestión. Adicionalmente, los tiempos de servicio varían

significativamente entre los distintos tipos de trámites con diferencias de desempeño entre gestores. En conjunto, estas condiciones explican la formación de cuellos de botella que deterioran la experiencia del usuario y la estabilidad operativa. Estudios recientes en teoría de colas y operaciones reconocieron estas fuentes de variabilidad y su impacto en desempeño, utilización y tiempos de espera [1] ,[2] .

La teoría de colas nos da herramientas para analizar este tipo de situaciones. Relaciona cuántos clientes llegan, cuánto tarda cada servicio y cuánta capacidad hay disponible [2] . Pero hay un problema: los modelos tradicionales asumen que las llegadas siguen un patrón Poisson, que los tiempos de servicio son exponenciales y que todo funciona a tasas constantes. Y la realidad casi nunca es así [3] , [4] . En la práctica hay interrupciones, cambios según la época del año, diferencias marcadas entre un trámite y otro, entre un funcionario y otro. Por eso la simulación de eventos discretos resulta más útil: Permite construir un modelo que se parece mucho más al sistema real y probar distintas soluciones sin tocar la operación del día a día [5] [6] .

La simulación se utiliza para analizar y proponer mejoras de estos sistemas en distintos sectores. En salud, por ejemplo, distintas intervenciones basadas en simulación de eventos discretos redujeron tiempos de espera y mejoraron la estabilidad del servicio cambiando la forma en que se organizan los procesos [7] [8] . En servicios públicos, hay casos documentados donde se logró reducir hasta un 28% el tiempo total de atención al apoyarse en modelos de simulación [9] . También en manufactura y otros servicios se ha visto cómo herramientas como FlexSim ayudan a representar escenarios complicados y a usar mejor los recursos disponibles [10] [12] .

Detrás de estos trabajos hay una base metodológica sólida. Se usan técnicas de modelado estocástico, se valida todo estadísticamente y se analiza cómo la incertidumbre en los datos afecta los resultados [13] Hay estudios clásicos sobre cómo ajustar distribuciones de probabilidad a los datos reales [12] y sobre cómo asegurarse de que un modelo de simulación realmente representa lo que sucede en la realidad [13] . Este estudio sigue esa misma línea. Se desarrolló un modelo de simulación para un centro de atención de servicios públicos en el Valle del Cauca usando información real de su operación. El objetivo fue caracterizar el sistema, identificar distribuciones de entrada y servicio que representaran adecuadamente su comportamiento e instalar una línea base cuantitativa para proponer mejoras operativas posteriores bajo incertidumbre [2] [5] [2] [14] . Existe un estudio donde se busca solucionar un problema de programación en reparaciones donde el reto es optimizar la secuencia y reducir tiempos y mejorar la eficiencia [15] .

El diseño de este estudio de simulación inicia con la definición clara de reducir los tiempos de espera de los clientes y la mejora de la eficiencia de los funcionarios que atienden los diferentes tramites [16] .

## 2. Metodología

### 2.1. Diseño del estudio

El estudio siguió el ciclo recomendado para proyectos de simulación: formulación del problema y conceptualización, especificación y construcción del modelo, estimación de parámetros, verificación, validación y experimentación [5] [6] . Para la transparencia del reporte se adoptaron los lineamientos STRESS-DIS (Strengthening The Reporting of Empirical Simulation Studies) [17] y los

principios de verificación y validación [13] , con el fin de facilitar la reproducibilidad. Existe un análisis de sensibilidad donde se busca determinar la cantidad óptima de puntos de atención teniendo en cuenta el costo de espera [18] **¡Error! No se encuentra el origen de la referencia.** El impacto de la decisión humana es importante en los análisis ya que se debe considerar la percepción errónea del tiempo, este autor incorpora el comportamiento humano en la simulación, también se debe tener en cuenta la velocidad de atención de los funcionarios de atención, las simulaciones que consideran los factores psicológicos son más realista para el diseño del modelo [7] **¡Error! No se encuentra el origen de la referencia.**.
El uso de la simulación se debe utilizar como una herramienta de evaluación no de solución, esta herramienta permite mejorar el rendimiento del proceso de atención ya que permite modelar procesos complejos y tener varios escenarios, permite optimizar recursos y tomar decisiones basadas en datos [16] .

## 2.2. Contexto y Sistema Bajo Estudio

El estudio se realiza en un centro de atención presencial de una empresa de servicios públicos ubicada en Cali, Colombia. El sistema gestiona seis tipos de trámites: Financiaciones, PQR Verbales, PQR Escritos, PqrRRP y Silencioso, Trámites y Solicitudes y Ventas. El personal se organizó en trece grupos (G01–G13) según complejidad, pe. G01 ejecuta 6 tipos de procesos y tiene la mayor cantidad de atenciones, G03 ejecuta 5 tipos de procesos y está en segundo lugar con cantidad de atenciones, G09 ejecuta 2 tipos de procesos y está en tercer lugar con cantidad de atenciones, G04 ejecuta 4 tipos de procesos y está en cuarto lugar con cantidad de atenciones, G02 ejecuta 5 tipos de procesos y está en quinto lugar con cantidad de atenciones, etc. La combinación de grupos y subcategorías de trámite generó 42 configuraciones de atención con dinámicas y tiempos de servicio diferenciados.
Vemos que el proceso tiene datos de variabilidad en la llegada de usuarios, diferentes tipos de tramites, recursos con diferentes habilidades y horarios definidos [16]

## 2.3. Arquitectura Conceptual del Sistema

Se representó un sistema abierto con una cola inicial y un clasificador que enrutó a cada usuario a uno de los 42 procesadores (unidades de servicio) según su trámite. Las llegadas se modelaron como un proceso puntual donde los usuarios llegan a cierta tasa λ y las salidas como abandonos del sistema tras el servicio. La propiedad PASTA (Poisson Arrivals See Time Averages) se consideró cuando procedió y se verificó empíricamente sobre los datos [2] [19] , al aplicar la simulación con FlexSim, se modela el proceso actual y permite evaluar las métricas claves como tiempos de espera, variabilidad en los flujos de clientes y lograr la eficiencia en la atención de casos según su prioridad, con la simulación implica un rediseño del proceso como se evidencia en el artículo [13] . La simulación con Flexim permitió ver el impacto de las recomendaciones de mejoras y facilito las acciones correctivas y de mejora [5]
Se debe tener en cuenta los clientes, las estaciones de trabajo, áreas de espera y el personal de atención [16] .

## 2.4. Datos y Preparación

Se trabajó con más de 35.000 registros correspondientes a aproximadamente seis meses de operación. Cada registro incluyó sello temporal de llegada, tiempo de servicio, intervalo entre llegadas, identificador del gestor, tipo de trámite y fecha/hora. Se

procesaron datos con Python 3.11 y librerías pandas, numpy, matplotlib y scipy.stats. Se estandarizaron tiempos a minutos, se exploró completitud y consistencia, y se filtraron atípicos mediante el rango intercuartílico (Tukey) adaptado a contexto operativo. Finalmente, se agruparon observaciones por las 42 combinaciones grupo–subcategoría para el modelado de entradas y servicio.

## 2.5. Modelado Estocástico de Entradas

### 2.5.1. Proceso de llegadas

Se estimó la tasa promedio de llegadas a partir de los tiempos entre llegadas. Cuando el contraste estadístico apoyó un proceso Poisson, se modeló el intervalo $X$ con distribución exponencial:

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0 \quad (1)$$

La idoneidad del supuesto se evaluó con pruebas Kolmogorov–Smirnov (K-S) y Anderson–Darling (A-D) en los subconjuntos pertinentes [5] [20]

### 2.5.2. Tiempos de servicio

Para cada una de las 42 combinaciones se ejecutó el flujo: extracción de subconjunto, estadística descriptiva (media, mediana, desviación estándar, coeficiente de variación), ajuste paramétrico y selección de modelo. Las familias candidatas fueron Exponencial, Gamma, Lognormal, Weibull y Normal, estimadas por máxima verosimilitud (MLE) [5] [20] . Las funciones de densidad consideradas fueron:

$$\text{Gamma: } f(x) = \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \quad (2)$$

$$\text{Lognormal: } f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnln\, x - \mu)^2}{2\sigma^2}} \quad (3)$$

$$\text{Weibull: } f(x) = \frac{\alpha}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-(x/\lambda)^{\alpha}} \quad (4)$$

$$\text{Normal: } f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

La bondad de ajuste se evaluó con K-S, A-D y Chi-cuadrado de Pearson, considerando umbral $\alpha$ = 0.05 [5] [20] . Para balancear ajuste y parsimonia se aplicaron AIC y BIC:

$$AIC = 2k - 2\, ln\, ln\, (L) \quad (6)$$

$$BIC = k\, ln\, ln\, (n)\, - 2\, ln\, ln\, (\hat{L}) \quad (7)$$

donde $k$ fue el número de parámetros, $n$ el tamaño muestral y $\hat{L}$ la verosimilitud maximizada [14] **¡Error! No se encuentra el origen de la referencia.**. Los parámetros seleccionados se incorporaron posteriormente al modelo de simulación.

## 2.6. Implementación computacional

El modelo se implementó en FlexSim, con los componentes: Source (generación de llegadas con el proceso estimado), Queue (cola inicial FCFS – First Come, First Served), Processor (clasificador) para enrutar según trámite y 42 Processors específicos con distribuciones de servicio estimadas; y finalmente, Sink registró salidas y métricas del sistema [10] . Se parametrizó, para cada procesador, el número de servidores, la capacidad y la distribución de servicio. Se definieron escenarios para analizar cambios en dotación y reglas de asignación. El proceso se representa en la Figura 1, desde la llegada del usuario hasta su salida del sistema, incluyendo la asignación a grupos y tipos de trámite:

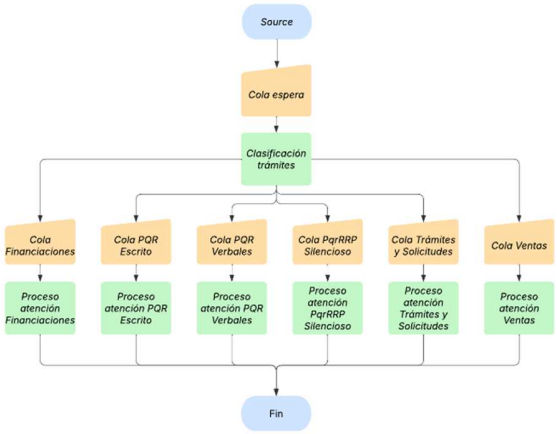


Figura 1. Flujograma del proceso actual desarrollado en simulación

### 2.7. Verificación, validación y diseño experimental

La verificación se realizó mediante revisión de trazas, chequeos de conservación de flujo y pruebas unitarias de lógica de enrutamiento [13] . La validación comparó métricas simuladas vs. observadas (tiempos medios, varianzas y niveles promedio en cola/sistema) con contrastes sobre medias y varianzas en estado estable [5] . Se aplicó calentamiento de Welch para remover sesgos transitorios y se ejecutaron réplicas independientes con intervalos de confianza del 95% sobre indicadores de desempeño [5] [21] . Con el fin de fortalecer las inferencias bajo incertidumbre de insumos, se atendieron recomendaciones recientes de validación de modelos de entrada y análisis de riesgo [14] [22] .

### 2.8. Métricas de evaluación

Se extrajeron indicadores estándar: tiempos promedio de espera en cola ($W_q$), tiempo total en el sistema ($W_s$), espera máxima observada ($W_{max}$), niveles promedio en cola ($L_q$) y sistema ($L_s$), utilización por procesador y global ($\rho$), y throughput horario/diario. Estos indicadores permitieron cuantificar la línea base y evaluar sensibilidad a cambios en demanda y capacidad [5] [6]

### 2.9. Consideraciones éticas y limitaciones

Los datos se anonimizaron de forma irreversible y se cumplieron las obligaciones de protección de datos personales aplicables en Colombia. El modelo asumió una tasa promedio de llegadas a nivel diario, homogeneidad estacional en el periodo observado e independencia entre llegadas y servicio; no se modeló abandono por impaciencia. Estas simplificaciones, habituales en estudios SED, no invalidaron la línea base, pero señalaron rutas de refinamiento [12] [14]

## 3. Resultados

En esta sección se presentan los hallazgos del modelo inicial de simulación, organizados por subcategorías de trámite y por grupos de atención. Se resumen las métricas clave de desempeño tales como, tiempos de atención, throughput, estado de las colas, se incluyen tablas y figuras comparativas para facilitar el análisis [23]

En la Tabla 1 se sintetiza el desempeño de cada una de las seis subcategorías de trámite modeladas. Se reporta el número de usuarios atendidos (throughput) durante la simulación [23] , el porcentaje de tiempo que la cola estuvo ocupada (utilización de cola), el contenido promedio y máximo de usuarios en cola, los tiempos promedio y máximos de permanencia (staytime) en el sistema para cada tipo de trámite. Se observan contrastes marcados: la categoría Trámites y Solicitudes concentró el mayor volumen de demanda (118 usuarios atendidos, ~75% del total) y presentó

la cola más congestionada, con un 48,5% del tiempo con al menos un usuario esperando y un tamaño promedio de ~16 personas en fila (máximo de 63) durante la jornada simulada. En concordancia, su throughput horario fue el más alto y sus servidores operaron cercanos al límite de capacidad, reflejando un cuello de botella significativo en esta subcategoría. Por el contrario, trámites de Financiaciones, PQR Escritos y PqrRRP Silencioso mostraron demandas muy bajas (solo 3–4 usuarios atendidos cada uno) y prácticamente no generaron espera en cola (utilización ~0%). Esto sugiere que la capacidad asignada a estas categorías fue suficiente o excedente respecto a su llegada de usuarios, evitando acumulación de filas. Los trámites de Ventas tuvieron un nivel de actividad intermedio (10 atenciones) con una cola poco frecuente (cola ocupada 6,4% del tiempo, máximo 5 clientes) y tiempos de atención moderados. PQR Verbales fue la única categoría de peticiones/quejas que mostró congestión apreciable: atendió 19 usuarios y su cola estuvo ocupada ~30% del tiempo, con hasta 10 usuarios esperando en el pico; sin embargo, su nivel promedio de cola se mantuvo bajo (1,4 en espera).

Estos resultados cuantifican cómo las categorías de mayor volumen (Trámites y Solicitudes) o con capacidad limitada (PQR Verbales) tienden a formar colas más largas y persistentes, mientras que categorías de baja demanda no alcanzan a saturar su recurso asignado.

Tabla 1. Comparación de los resultados de los escenarios propuestos

| **Criterio** | **Modelo Inicial** | **Propuesta 1** | **Propuesta 2** |
|---|---|---|---|
| Throughput**¡Error! No se encuentra el origen de la referencia.** total | 158 usuarios | 234 usuarios | 236 usuarios |
| Colas saturadas (>30%) | 3 (PQR Verbales, Trámites, Ventas) | 1 (solo Trámites) | 5 (G01, G02, G03, G04, G05) |
| Utilización máxima | Alta en grupos PQRs | Moderada en mayoría | Crítica (>80%) en G01, G02 |
| Balance de carga | Muy desigual | Equilibrado por tipo de trámite | Desigual (grupos primeros más cargados) |
| Tiempos promedio altos | PQRs > 30 min | PQRs y Trámites > 25 min | G2/G3 > 25 min; gran dispersión |
| Tamaño promedio de colas | Hasta 16 usuarios (Trámites) | Máximo 1.4 | 0.22–0.82 |
| Complejidad operativa | Alta (42 rutas de asignación) | Media (12 flujos) | Baja (8 grupos, sin clasificación) |

En la Tabla 1 se incluyen el número de usuarios atendidos, utilización de la cola (% tiempo con al menos un usuario esperando), longitud promedio y máxima de la cola, tiempos de permanencia promedio y máximo en el sistema (en minutos) para cada tipo de trámite.

Analizando los tiempos de atención (staytime) por trámite [23] , se evidencia que las peticiones/quejas formales tuvieron las demoras más prolongadas. En PQR Verbales, el tiempo medio de permanencia en el sistema fue de ~28 minutos y llegó hasta 58 minutos en el peor caso (Tabla 1). De manera similar, en PQR Escritos el tiempo máximo alcanzó ~43 minutos. Estas demoras responden a que pocos grupos de funcionarios están habilitados para estos trámites (ver análisis por grupo más adelante [1] ) y a la variabilidad inherente en su atención, lo cual propicia filas incluso con bajas llegadas.

En contraste, el trámite masivo de Solicitudes a pesar de su gran volumen logró tiempos de atención individuales más contenidos (promedio ~16 min). Esto se debe a la atención

paralela por múltiples grupos de servidores en esta categoría [1] , lo que diluye la espera por usuario aun cuando la cola total sea larga. En trámites de Financiaciones y Ventas, los tiempos de atención fueron relativamente cortos (promedios ~13–15 min) y con poca variabilidad, coherente con una atención ágil y sin congestión. En general, los resultados confirman que, bajo alta utilización de los servidores, los tiempos de espera crecen de forma no lineal; por ejemplo, Trámites y Solicitudes operó cerca del 100% de uso de su recurso gran parte del tiempo, generando esperas prolongadas y una cola acumulada (32 usuarios aún en fila al final del día). Por su parte, categorías con utilización ρ muy baja (≈0) no registraron espera significativa, atendiendo a los usuarios casi de inmediato.

Tabla 2. Tiempo promedio máximo de procesamiento en el sistema por tipo de trámite.

| Subcategoría | Tiempo prom. Máx de proceso (Min) |
|---|---|
| Financiaciones | 16.74 |
| PQR Escrito | 39.84 |
| PQR Verbales | 36.39 |
| PqrRRP Silencioso | 33.85 |
| Trámites y Solicitudes | 21.33 |
| Ventas | 34.18 |

Se aprecia en la Tabla 2 que los trámites de PQR (Peticiones, Quejas y Reclamos) presentan los mayores tiempos medios de atención, superando los 25–30 minutos, mientras que categorías operativas como Ventas y Solicitudes se resuelven en tiempos más cortos (~15 minutos en promedio). Los trámites con baja demanda (Financiaciones, PQR Escritos, PqrRRP Silencioso) mostraron tiempos reducidos al no generarse espera.

Tabla 3. Porcentaje de utilización de la cola por tipo de trámite.

| Subcategoría | Utilización de la Cola (%) |
|---|---|
| Financiaciones | 0.00 |
| PQR Escrito | 0.00 |
| PQR Verbales | 30.24 |
| PqrRRP Silencioso | 0.00 |
| Trámites y Solicitudes | 48.53 |
| Ventas | 6.39 |

El indicador de la Tabla 3 refleja la fracción del tiempo simulado en que cada cola estuvo ocupada (con al menos un cliente esperando). Se observa que Trámites y Solicitudes tuvo la cola ocupada casi la mitad del tiempo (~48%), evidenciando congestión severa, seguida de PQR Verbales (~30%). Las demás categorías estuvieron mayormente sin espera (valores por debajo de 7%, cercanos a cero), lo que indica suficiencia de capacidad.

Complementando el análisis por trámites, la Tabla 2 presenta los tiempos de atención desagregados por grupo de atención para la categoría más concurrida (Trámites y Solicitudes). Se listan los ocho grupos de funcionarios (G01–G08) definidos según nivel de complejidad. Se reportan sus tiempos promedio, mínimo y máximo de servicio para los usuarios de esta subcategoría. Destaca que existe variabilidad significativa entre grupos: algunos mostraron un desempeño más eficiente, como el Grupo G07 con un tiempo promedio de solo 9,63 min en este trámite (el más bajo de todos) y G06 con 10,92 min, indicando alta rapidez de atención. En cambio, grupos como G02 y G03 tuvieron los promedios más altos (por encima de 20 min) y episodios de atención mucho más prolongados (hasta ~75 min en el caso de G03). Esto sugiere diferencias tanto en la disponibilidad como en la productividad/habilidad de los grupos: por ejemplo, G07 y G06 podrían ser unidades especializadas en solicitudes comunes,

resolviendo casos con rapidez, mientras que G02 y G03 posiblemente enfrentaron casos más complejos o simultaneidad de tareas que les generaron demoras mayores. El Grupo G01 —que atiende prácticamente todas las categorías de trámite (ver Figura 2)— presentó un desempeño intermedio en Solicitudes (17,3 min promedio), pero en otras categorías más complejas su rendimiento fue notablemente menor: por ejemplo, G01 promedió ~36–40 min en trámites de PQR, significativamente por encima de otros grupos más especializados en esas áreas [23] . Esto indica que los grupos polivalentes como G01 tienden a sobrecargarse, atendiendo múltiples tipos de trámites a costa de mayores tiempos por usuario en algunos de ellos. Por otro lado, los grupos dedicados exclusivamente a una sola categoría (p. ej. G05–G08 en Solicitudes) no diversifican su atención, pero pueden enfocarse y resolver más rápido cada caso de su especialidad.

Tabla 4. Tiempo de atención en trámites de Solicitudes desagregado por grupo de atención.

| Grupo | Tiempo promedio (min) | Tiempo mínimo (min) | Tiempo máximo (min) |
|---|---|---|---|
| G01 | 17.32 | 3.76 | 34.33 |
| G02 | 21.33 | 0.93 | 64.66 |
| G03 | 20.83 | 3.34 | 75.01 |
| G04 | 16.33 | 2.18 | 38.86 |
| G05 | 17.74 | 2.41 | 68.83 |
| G06 | 10.92 | 2.06 | 28.55 |
| G07 | 9.63 | 2.16 | 48.38 |
| G08 | 17.05 | 1.69 | 62.67 |

La Tabla 4 muestra el tiempo promedio, mínimo y máximo que cada grupo G01–G08 tomó en atender a los usuarios de esta subcategoría durante la simulación. Se evidencian diferencias de desempeño importantes entre grupos, asociadas a distintos niveles de eficiencia o carga de trabajo.

En términos de utilización de los grupos de servicio, los resultados indican que la carga no estuvo distribuida homogéneamente. Los grupos dedicados a trámites con baja afluencia permanecieron ociosos gran parte del tiempo. Por el contrario, los grupos que atendieron múltiples categorías –notablemente G01, G02, G03 estuvieron ocupados durante la mayor parte de la jornada simulada (ρ cercano a 1,0), ya que además de las Solicitudes debieron cubrir trámites de PQR, Ventas y Financiaciones. Esta alta utilización de los recursos condujo, como era de esperarse, a un aumento en los tiempos de espera y a la formación de colas en las categorías correspondientes. Por ejemplo, G01 (el grupo más versátil) atendió usuarios de todas las subcategorías y operó al límite de su capacidad, lo cual se reflejó en demoras considerables especialmente en los trámites más complejos que cubría. En contraste, grupos especializados como G06 y G07, al enfocarse solo en Solicitudes y contar con menores tiempos de servicio, lograron atender más usuarios por unidad de tiempo, reduciendo sustancialmente la espera en esa cola [7] . Estos hallazgos ponen de manifiesto la importancia de equilibrar la asignación de personal con la demanda por tipo de trámite: una configuración subóptima **¡Error! No se encuentra el origen de la referencia.** (p. ej., pocos grupos para trámites de alta demanda, o grupos multitarea sobrecargados) deriva en un desempeño desigual entre filas, con algunas prácticamente vacías y otras desbordadas.

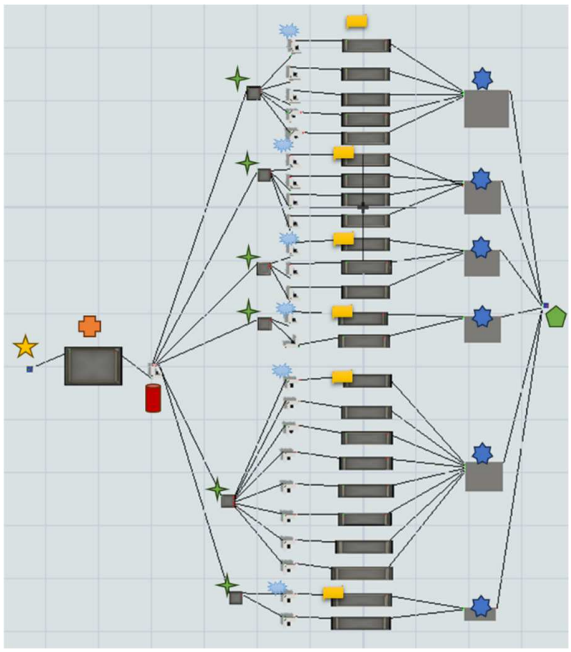

Figura 2. Vista superior del modelo inicial en FlexSim

Como se observa en la Figura 2 en el flujograma del proceso de atención simulado en FlexSim, los usuarios ingresan por una fuente común ( ) y pasan a una cola inicial única (FCFS) . Un Clasificador ( ) determina el tipo de trámite de cada usuario y lo enruta a la cola específica de su subcategoría (Financiaciones, PQR Escritos, PQR Verbales, PqrRRP Silencioso, Trámites y Solicitudes o Ventas) . Cada cola de trámite alimenta a uno o más Grupos de atención (G01–G08 ) capacitados para ese tipo de caso; por ejemplo, G01 y G02 atienden Financiaciones, G01, G02 y G03 atienden PQR Escritos, etc. (en total se modelaron 24 configuraciones de servidor distintas combinando grupo y trámite). Los recuadros Grupo G representan servidores (unidades de servicio) que pueden estar ocupados o libres. Una vez completada la atención y para validar las métricas del modelo, los usuarios salen hacia una cola específica para cada Processor ( ) y las colas comunes se unen en una ( ) para determinar las métricas de output que pasan para cada uno de los trámites. Finalmente, los usuarios salen del sistema por el Sink ( ) registrándose su tiempo de permanencia [10] . Este diagrama ayuda a visualizar cuellos de botella: por ejemplo, la rama de Trámites y Solicitudes concentra múltiples grupos, pero aun así sufrió saturación en la simulación [23] , lo que sugiere la necesidad de reforzar capacidad o mejorar el proceso en ese segmento.

En síntesis, los resultados del modelo de simulación permiten identificar las áreas críticas en el centro de servicios públicos estudiado. La subcategoría de Trámites y Solicitudes evidenció ser el punto más congestionado del sistema con altos niveles de utilización y esperas prolongadas, mientras que trámites especializados de baja frecuencia operaron con holgura. Así mismo, se observaron diferencias de desempeño entre grupos de funcionarios, atribuidas a la multifuncionalidad y a las distintas velocidades de servicio. Estas observaciones proporcionan una línea base cuantitativa para proponer mejoras: por ejemplo, redistribuir recursos hacia los trámites de mayor demanda o capacitar a más grupos en la resolución de PQR podría equilibrar la carga y reducir los tiempos de espera [10] . En la siguiente sección se discutirán implicaciones de estos hallazgos y posibles intervenciones, a la luz de trabajos previos que reportan reducciones de hasta 28% en tiempos de atención aplicando cambios informados por simulación.

### 3.1. Propuesta 1

A partir del rediseño estructural representado en la nueva arquitectura del modelo FlexSim (ver Figura 3) **¡Error! No se encuentra el origen de la referencia.**, se implementó una consolidación de categorías de atención y una redistribución de los recursos hacia tres

macroprocesos: Financiaciones/Ventas, PQR (que incluye PQR Escritos, PQR Verbales y PqrRRP Silencioso) y Trámites y Solicitudes. Esta reconfiguración simplificó los flujos de atención y permitió concentrar capacidad operativa en las áreas de mayor demanda.

### 3.1.1. Comportamiento General del Sistema

La Tabla 5 muestra el throughput o número de usuarios atendidos por cada macroproceso [23] . La categoría de Trámites y Solicitudes mantuvo su posición como la de mayor volumen con 194 casos resueltos, lo que representa aproximadamente el 78% del total del sistema. En comparación con la línea base, este valor refleja un aumento en capacidad efectiva sin necesidad de aumentar significativamente el personal.

Por otro lado, PQR concentró 31 atenciones y Financiaciones/Ventas agrupó 9 casos, ambas con bajo nivel de congestión, lo cual indica una atención oportuna en estos segmentos sin cuellos de botella [9] .

Tabla 5. Usuarios atendidos por tipo de trámite

| Tipo de trámite | Usuarios atendidos |
|---|---|
| Financiaciones/Ventas | 9 |
| PQR (Escrito, Verbales y Silencioso) | 31 |
| Trámites y Solicitudes | 194 |

### 3.1.2. Utilización de las Colas

La Tabla 6 presenta el porcentaje de utilización de las colas por cada categoría. A diferencia de la línea base (donde algunas colas como la de Trámites y Solicitudes alcanzaban hasta 48.5% de ocupación), la propuesta logró reducir la utilización de todas las colas a menos del 5%. En efecto, Trámites y Solicitudes fue la única cola con ocupación perceptible, registrando solo 4.96% del tiempo activa. Esto representa una mejora sustancial en términos de fluidez del sistema y reducción de la espera.

Tabla 6. Utilización de las colas por trámite

| Tipo de trámite | Utilización de cola (%) |
|---|---|
| Financiaciones/Ventas | 0.00 |
| PQR (Escrito, Verbales y Silencioso) | 0.00 |
| Trámites y Solicitudes | 4.96 |

### 3.1.3. Tamaño Promedio de la Cola

Complementando lo anterior, la Tabla 7 muestra el número promedio de clientes en espera por categoría. Se evidencia una drástica reducción en el tamaño promedio de cola, especialmente en Trámites y Solicitudes que pasó de 16.3 personas promedio a tan solo 0.10 usuarios, en PQR y Financiaciones/Ventas, el promedio fue 0.00, es decir, no se formaron filas.

Tabla 7. Tiempo promedio máximo de procesamiento por tipo de trámite

| Tipo de trámite | Tiempo prom. Máx. (Min) |
|---|---|
| Financiaciones/Ventas | 20.42 |
| PQR (Escrito, Verbales y Silencioso) | 40.06 |
| Trámites y Solicitudes | 25.20 |

### 3.1.4. Tiempos de Atención por Subprocesos

La Tabla 3 resume los tiempos promedio, mínimo y máximo de atención (staytime) registrados para los procesadores específicos en esta propuesta [23] . Los valores se mantuvieron dentro de los rangos observados en la línea base, con algunas mejoras significativas en uniformidad y reducción de máximos. Por ejemplo:

- Los tiempos de atención en Trámites_Solicitudes_7 fueron particularmente bajos (promedio de 12.06 min),

- Mientras que los picos más altos fueron en Trámites_Solicitudes_12 con un máximo de 69.75 min, aunque se registraron con menor frecuencia.

Esta variabilidad indica que el modelo aún conserva ciertas ineficiencias en tareas complejas, pero ha logrado contener el impacto general de las demoras.

Tabla 8. Tiempos de atención por procesador.

| Procesador | Tiempo Promedio (min) | Tiempo Mínimo | Tiempo Máximo |
|---|---|---|---|
| Financiaciones_Ventas1 | 20.42 | 1.25 | 47.63 |
| Financiaciones_Ventas2 | 18.99 | 10.21 | 33.43 |
| PQR_1 | 25.21 | 9.85 | 61.31 |
| PQR_2 | 26.03 | 3.11 | 53.05 |
| PQR_3 | 40.06 | 10.88 | 56.58 |
| PQR_4 | 32.52 | 29.12 | 41.05 |
| Trámites_Solicitudes_1 | 15.37 | 2.38 | 41.05 |
| Trámites_Solicitudes_2 | 16.59 | 4.17 | 54.09 |
| Trámites_Solicitudes_3 | 21.32 | 3.90 | 44.94 |
| Trámites_Solicitudes_4 | 16.04 | 2.16 | 61.69 |
| Trámites_Solicitudes_5 | 24.23 | 3.88 | 57.55 |
| Trámites_Solicitudes_6 | 15.86 | 2.55 | 41.21 |
| Trámites_Solicitudes_7 | 12.06 | 2.05 | 30.14 |
| Trámites_Solicitudes_8 | 16.27 | 3.89 | 63.80 |
| Trámites_Solicitudes_9 | 15.06 | 4.83 | 38.74 |
| Trámites_Solicitudes_10 | 21.73 | 4.08 | 66.49 |
| Trámites_Solicitudes_11 | 13.32 | 2.03 | 42.04 |
| Trámites_Solicitudes_12 | 25.20 | 3.20 | 69.75 |

### 3.1.5. Distribución Visual del Proceso

La siguiente imagen muestra el nuevo diseño del modelo simulado en FlexSim. Se observa una mayor centralización del clasificador y una asignación más eficiente de colas y procesadores [10] , con agrupaciones claras por tipo de trámite:

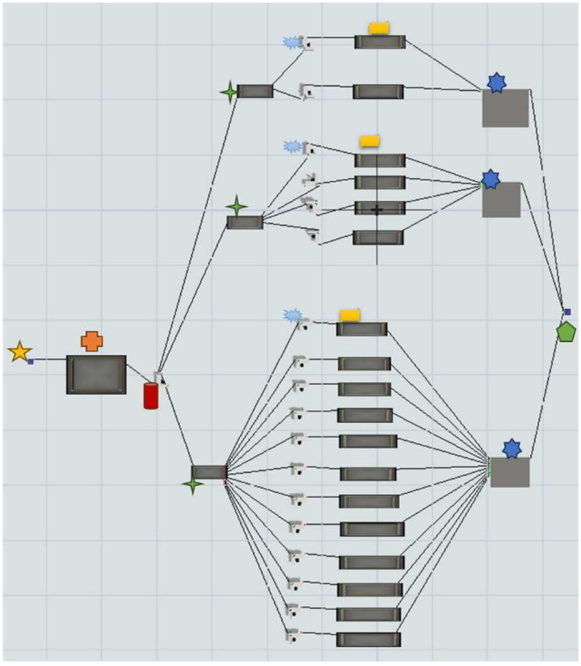

Figura 3. Vista superior en FlexSim Propuesta 1

## 3.2. Propuesta 2

Esta propuesta elimina la segmentación por tipo de trámite y rediseña el modelo para que todos los grupos funcionales (G01–G08) puedan atender cualquier tipo de solicitud (Financiaciones, PQRs, Trámites, Ventas, etc.). Los usuarios ingresan al sistema, pasan por un único clasificador y son enviados al primer grupo disponible sin distinción del trámite específico. Esta estrategia busca maximizar la flexibilidad operativa, reducir la ociosidad del personal y agilizar la atención de punta a punta.

### 3.2.1. Throughput por Grupo Funcional

La Figura 4. Usuarios atendidos por grupo funcional muestra cuántos usuarios fueron atendidos por cada grupo. El Grupo G01 resolvió la mayor cantidad de casos (50 usuarios), seguido por G03 (40) y G02 (35). Esto refleja una tendencia en la cual los primeros grupos disponibles o los más ágiles, absorben

mayor carga del sistema. Los grupos G07 y G08, en cambio, participaron menos intensamente (~15–23 casos), posiblemente por su menor disponibilidad, menor velocidad de atención o porque no fueron liberados con la misma rapidez [23] .

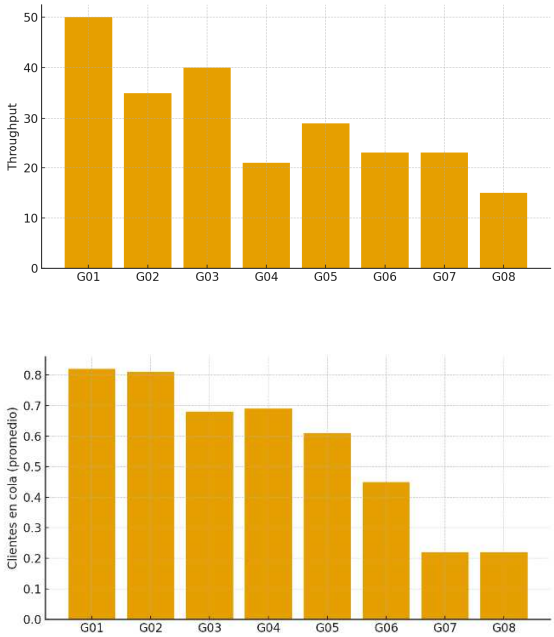


Figura 4. Usuarios atendidos por grupo funcional

La Figura 5. Utilización de los grupos funcionales indica el porcentaje de tiempo que cada grupo estuvo ocupado (utilización del servidor). Los grupos G01 y G02 alcanzaron una utilización crítica (>81%), lo cual señala una sobrecarga potencial, mientras que G03–G05 se mantuvieron también en niveles elevados (~61%–69%). En contraste, G07 y G08 trabajaron en rangos mucho más bajos (21%–22%), lo que sugiere que, si bien están disponibles, no están siendo aprovechados al mismo ritmo, ya sea por políticas internas o tiempos de procesamiento más lentos [15] .

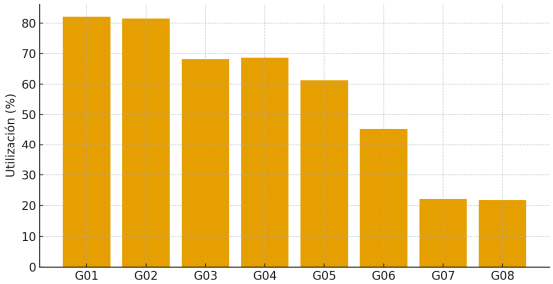


Figura 5. Utilización de los grupos funcionales

### 3.2.2. Tamaño Promedio de las Colas por Grupo

La Figura 6. Tamaño promedio en cola por grupo funcional muestra el número promedio de usuarios esperando por grupo. Se confirma el patrón anterior: G01 y G02 acumularon las colas más largas (~0.82 usuarios promedio), mientras que los grupos de menor uso mantuvieron colas casi inexistentes (~0.2). Esta asimetría en la carga sugiere que, aunque la estrategia reduce los pasos intermedios (como clasificación o segmentación), aún persiste un desequilibrio en la distribución de usuarios, probablemente porque algunos servidores se liberan más rápido que otros.

Figura 6. Tamaño promedio en cola por grupo funcional

### 3.2.3. Tiempos de Atención por Procesador

A continuación, en la Tabla 9 se listan los tiempos de atención por grupo, reflejando también la heterogeneidad del sistema [13] . Algunos grupos (como G7 y G8) tienen tiempos medios más bajos, mientras que G2 y G3 presentan mayor variabilidad.

Tabla 9. Tiempos de atención por procesador.

| Grupo | Procesador | Promo (min) | Mín | Máx |
|---|---|---|---|---|
| G1 | G1_Processor1 | 18.87 | 3.64 | 55.26 |
| G1 | G1_Processor2 | 18.67 | 2.96 | 99.27 |
| G2 | G2_Processor1 | 21.99 | 2.50 | 51.59 |

| | | | | |
|---|---|---|---|---|
| G2 | G2_Processor2 | 30.02 | 5.60 | 129.23 |
| G3 | G3_Processor1 | 25.64 | 2.83 | 77.03 |
| G3 | G3_Processor2 | 18.84 | 2.61 | 60.24 |
| G4 | G4_Processor1 | 18.05 | 3.68 | 55.10 |
| G5 | G5_Processor1 | 12.83 | 2.11 | 41.36 |

### 3.2.4. Estructura del Modelo de Simulación

La imagen del modelo en FlexSim evidencia que, a diferencia de propuestas anteriores, no existe una etapa de clasificación ni colas diferenciadas por tipo de trámite. El clasificador único redirige a cualquier grupo disponible. Esto simplifica el flujo y puede reducir el tiempo total de espera [10] pero al mismo tiempo puede acentuar la desigualdad en carga de trabajo si no se aplica una política de asignación balanceada (por ejemplo, round robin o least loaded first [15] ).

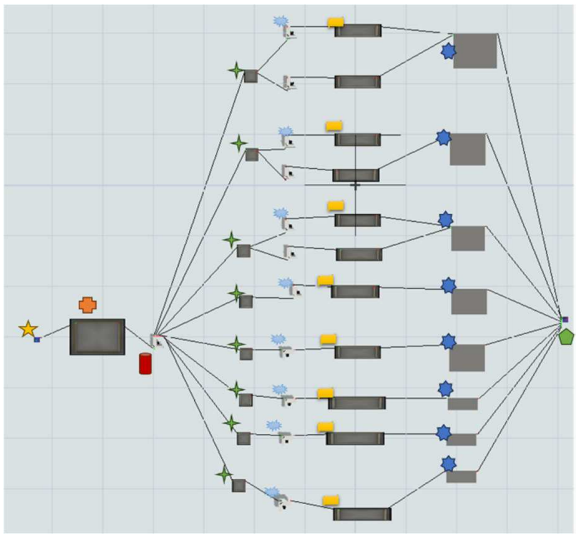

Figura 7. Modelo FlexSim propuesta 2 – Grupos funcionales generalistas.

## 3.3. Análisis

### 3.3.1. Descripción General de los Modelos

- Modelo inicial: Clasificación por tipo de trámite (Financiaciones, PQRs, Ventas, etc.) y asignación específica por grupo.
- Propuesta 1: Agrupación por tipo de trámite con múltiples servidores por categoría (12 servidores), sin clasificación final, generando que los funcionarios sean más dedicados a algún trámite específico [1] .
- Propuesta 2: Asignación directa a primer grupo funcional disponible (8 servidores generalistas), sin clasificación. Esto pretende no generar una clasificación por trámite, sino que se agrupan por funcionarios y éstos atienden cualquier solicitud.

### 3.3.2. Análisis de Métricas

**a. Eficiencia Operacional (Throughput Total)**

El throughput refleja la cantidad de usuarios atendidos exitosamente [23]

- El modelo inicial atendió 158 usuarios, limitado por cuellos de botella en PQRs y Trámites.
- Propuesta 1 alcanzó 234 usuarios, mejorando la asignación de servidores por categoría.
- Propuesta 2 logró 236 usuarios, apenas superior, gracias a la eliminación de rutas múltiples.

**Resultado**: Ambos rediseños superan ampliamente al modelo original en capacidad de atención [10]

**b. Utilización y Saturación de Recursos**

- En el modelo inicial, las colas de PQR Verbales y Trámites presentaron saturación con niveles de uso superiores al 30%,

especialmente en colas con 16 usuarios promedio.

- Propuesta 1 logró un equilibrio mejorado, con saturación solo en la cola de Trámites y Solicitudes (48%), manteniendo bajo uso en las demás.
- Propuesta 2 mostró mayor desigualdad: G01 y G02 operaron con más de 80% de ocupación y cinco de ocho grupos superaron el 60% de utilización, lo cual aumenta el riesgo de sobrecarga local.

**Resultado**: Propuesta 1 presenta mejor control de saturación.

**c. Tiempos de Atención**

En el modelo original, los tiempos promedio más altos se registraron en

- PQR Verbales y Escritos, alcanzando hasta 40 minutos.
- En Propuesta 1, persisten tiempos altos para algunos trámites (hasta 40 minutos), pero se diversifican entre 12 procesadores.
- En Propuesta 2, algunos servidores (como G2_Processor2 con 30.02 minutos promedio y 129.23 máximo) presentan altísima dispersión en tiempos, lo cual afecta la previsibilidad.

**Resultado**: Las dos propuestas mantienen tiempos similares, pero Propuesta 2 presenta mayor varianza entre grupos.

**d. Equilibrio en Distribución de Carga**

- El modelo inicial tiene 24 configuraciones entre grupo y trámite, lo cual dificulta la asignación balanceada y genera desequilibrios.
- Propuesta 1 agrupa trámites y servidores (12 canales únicos), lo que permite un balance más controlado y especializado [15] .
- Propuesta 2 otorga total libertad a los usuarios para ser atendidos por cualquier grupo, lo que termina sobrecargando los primeros servidores disponibles.

**Resultado**: Propuesta 1 equilibra mejor la carga y reduce el riesgo de ociosidad o sobreuso extremo.

**e. Complejidad del Modelo y Escalabilidad**

- El modelo original requiere gran cantidad de configuraciones manuales y clasificación por cada tipo de caso.
- Propuesta 1 reduce la complejidad al trabajar por categorías, facilitando su gestión.
- Propuesta 2 es el modelo más simple estructuralmente y es fácilmente escalable en número de operadores, pero requiere lógica adicional si se busca balancear la carga adecuadamente.

**Resultado**: En la Tabla 1 se muestra que la propuesta 2 es más simple; Propuesta 1 balancea mejor entre eficiencia y control.

**Recomendación Final**

La Propuesta 1 representa la opción más balanceada entre eficiencia, control operativo y escalabilidad. A pesar de no tener el throughput más alto [23] , logra:

- Reducción significativa en saturación de colas [7] .
- Buena utilización del personal sin sobrecarga extrema.
- Manejo especializado de trámites con lógica de asignación clara [23] .
- Menor dispersión en tiempos de atención.

En cambio, Propuesta 2, aunque ofrece una estructura más sencilla y marginalmente superior en throughput [23] , muestra signos de concentración de carga en pocos grupos, lo cual puede convertirse en un cuello de botella ante incrementos de demanda o en contextos de alta exigencia.

Por tanto, se recomienda adoptar la Propuesta 1 como modelo base para implementación o prueba piloto, incorporando métricas de balanceo dinámico y priorización si fuese necesario en etapas futuras.

## 4. Discusión

Este estudio desarrolló y validó un modelo de simulación de eventos discretos para analizar y rediseñar el sistema de atención presencial de una empresa de servicios públicos del Valle del Cauca. Los hallazgos principales muestran que la configuración inicial presentaba desequilibrios operativos significativos, con saturación crítica en la categoría "Trámites y Solicitudes" (utilización >100%, colas de hasta 63 usuarios) mientras otras categorías operaban con capacidad ociosa. Las dos propuestas de rediseño evaluadas incrementaron el throughput entre 48% y 49% respecto a la línea base, aunque con trade-offs distintos en términos de balance de carga y complejidad operativa.

### 4.1. Interpretación de los Hallazgos Principales

El cuello de botella identificado en "Trámites y Solicitudes" responde a la combinación de alta demanda (75% del volumen total) con capacidad insuficiente para absorber la variabilidad inherente en los tiempos de servicio. Este comportamiento es consistente con la teoría de colas, donde la utilización del servidor cercana o superior al 100% ($\rho \geq 1$) genera crecimiento no lineal en los tiempos de espera y longitud de cola [2], [3]. Los tiempos máximos observados de hasta 69 minutos en el modelo inicial confirman que el sistema operaba en régimen inestable durante períodos de alta demanda. La Propuesta 1, basada en especialización por macroprocesos, logró reducir la saturación de colas de 3 a 1 categoría y mantener la utilización de personal bajo 70% en la mayoría de los grupos. Este resultado se alinea con estudios previos en servicios públicos donde la consolidación de categorías similares y la reasignación de recursos basada en simulación generaron reducciones de 28% en tiempos de atención [9]. La especialización funcional permite que los servidores desarrollen competencias específicas para cada tipo de trámite, reduciendo la variabilidad en los tiempos de servicio (CV menor) y consecuentemente, mejorando la estabilidad del sistema [2]. Por otro lado, la Propuesta 2 (generalistas) alcanzó el throughput más alto (236 usuarios) al eliminar restricciones de enrutamiento, pero concentró la carga en los primeros grupos disponibles (G01, G02 con utilización >81%). Este patrón es esperado en sistemas sin políticas activas de balanceo de carga: los servidores más rápidos o mejor posicionados tienden a absorber más trabajo, generando desigualdad en la utilización [14]. Si bien la simplicidad estructural de esta propuesta facilita la escalabilidad, la alta varianza en tiempos de atención entre grupos (de 12.83 a 30.02 minutos promedio) introduce impredecibilidad en la experiencia del usuario.

### 4.2. Comparación con la Literatura

Los resultados obtenidos son coherentes con reportes previos sobre aplicación de simulación de eventos discretos en servicios. Villarreal et al. [23] encontraron que modelos multicanal (M/M/m) en agencias de viaje revelaban tiempos de espera promedio de 9.18 minutos, y recomendaban inversiones tecnológicas y reasignación de actividades para reducir la carga. De manera similar, nuestro estudio identifica la necesidad de redistribuir recursos hacia trámites de alta demanda [24] demostraron que incrementar la capacidad de 7 a 8 ventanillas en un almacén redujo tiempos de espera en 75% (de 10.52 a 2.56 minutos). En nuestro caso, la Propuesta 1 no aumentó significativamente el personal, sino que reorganizó su asignación, logrando mejoras comparables mediante eficiencia estructural. Los hallazgos también resuenan con principios teóricos establecidos. La Ley de Little ($L = \lambda W$) predice que, para una tasa de llegada constante, reducir el tiempo en sistema (W) disminuye proporcionalmente el número de usuarios en el sistema (L). Nuestros resultados confirman esta relación: la Propuesta 1 redujo el tamaño promedio de cola de 16.3 a 0.10 usuarios en "Trámites y Solicitudes", acompañado de una reducción en tiempos de permanencia de ~30 a ~25 minutos en promedio.

### 4.3. Implicaciones Prácticas

Desde una perspectiva gerencial, la Propuesta 1 representa una alternativa viable para implementación piloto sin requerir contratación masiva de personal. La consolidación de las seis subcategorías originales en tres macroprocesos simplifica la capacitación cruzada y facilita la gestión de turnos y licencias. Sin embargo, su implementación exitosa depende de: (a) capacitar al personal existente en trámites que previamente no atendían, (b) ajustar los sistemas informáticos para permitir el acceso a los casos de las tres categorías, (c) comunicar claramente a los usuarios la nueva estructura para evitar confusión.

La Propuesta 2, aunque atractiva por su simplicidad operativa, requeriría mecanismos adicionales de balanceo de carga (como asignación round-robin o least-loaded-first [14]) para evitar la sobrecarga de grupos específicos. Sin estas políticas activas, el riesgo de burnout en G01 y G02 es considerable, lo cual podría traducirse en ausentismo, rotación de personal y deterioro en la calidad del servicio a largo plazo.

Más allá de las propuestas evaluadas, el modelo desarrollado constituye una herramienta de apoyo a la toma de decisiones que puede extenderse para evaluar escenarios futuros, tales como: (i) incremento estacional en demanda (ej.: períodos de facturación), (ii) fallas en sistemas informáticos que reduzcan la velocidad de atención, o (iii) expansión a nuevos servicios. La validación estadística del modelo y el uso de lineamientos STRESS-DIS [16] fortalecen la confianza en las proyecciones generadas.

### 4.4. Limitaciones del Estudio

A pesar del rigor metodológico, el estudio presenta limitaciones que deben reconocerse. Primero, el modelo asumió tasas de llegada promedio sin capturar patrones intradiarios o estacionales. Es posible que existan picos de demanda en horarios específicos (ej.: mediodía, fin de mes) que no fueron modelados explícitamente. La simulación de un solo día operativo puede no capturar la variabilidad de largo plazo, por lo que se recomienda extender el horizonte de simulación y

realizar análisis de sensibilidad sobre la tasa de llegadas [13]. Segundo, no se modeló el comportamiento estratégico de los usuarios, específicamente el abandono por impaciencia (reneging) o la elección de no ingresar al sistema ante colas largas visibles (balking) [3]. Estudios sobre percepción del tiempo de espera indican que los usuarios tienden a sobrestimar la duración real de la espera y los factores psicológicos influyen en la satisfacción más allá del tiempo objetivo [7], [25]. Incorporar estos fenómenos requeriría datos adicionales sobre tasas históricas de abandono y encuestas de satisfacción, los cuales no estaban disponibles en el momento del estudio. Tercero, aunque se realizó validación mediante comparación de indicadores del modelo con promedios históricos, no se ejecutó una validación externa con datos operativos post-implementación. La recomendación final de adoptar la Propuesta 1 se basa en evidencia simulada, pero su efectividad real solo puede confirmarse mediante un piloto controlado en el centro de atención. Cuarto, el modelo asumió independencia entre llegadas y tiempos de servicio, así como distribuciones estacionarias a lo largo del día. Si bien se ajustaron distribuciones empíricas a los datos históricos, no se realizaron pruebas formales de bondad de ajuste (ej.: Kolmogorov-Smirnov, Anderson-Darling) para cada una de las 42 configuraciones grupo-trámite debido a limitaciones de tamaño muestral en categorías de baja frecuencia. Esto introduce incertidumbre en las proyecciones, especialmente para trámites poco frecuentes como "Financiaciones" o "PqrRRP Silencioso". Finalmente, el análisis de costos económicos de implementación no fue abordado. Si bien la Propuesta 1 no requiere personal adicional, sí implica costos de capacitación, ajustes tecnológicos y posible resistencia al cambio organizacional. Un análisis costo-beneficio que cuantifique estos factores en términos monetarios fortalecería la argumentación para la toma de decisiones gerenciales.

## 5. Conclusiones

Este estudio demuestra que la simulación de eventos discretos constituye una herramienta efectiva para diagnosticar ineficiencias operativas en centros de atención de servicios públicos y evaluar alternativas de rediseño antes de su implementación. El análisis de más de 35,000 registros operativos permitió construir un modelo validado que identificó un cuello de botella crítico en la categoría "Trámites y Solicitudes", donde la demanda superaba consistentemente la capacidad instalada, generando tiempos de espera de hasta 69 minutos y colas de 63 usuarios en el peor escenario observado. Las dos propuestas de rediseño evaluadas incrementaron el throughput del sistema entre 48% y 49% (de 158 a 234-236 usuarios/día) sin requerir aumentos significativos en la plantilla de personal. La Propuesta 1, basada en especialización por macroprocesos, logró el mejor equilibrio entre eficiencia operativa, control de saturación y distribución equitativa de la carga de trabajo. Esta alternativa redujo la saturación de colas de tres a una sola categoría y mantuvo la utilización promedio del personal bajo 70%, mitigando riesgos de sobrecarga y deterioro en la calidad del servicio. Por el contrario, la Propuesta 2 (generalistas) alcanzó un throughput marginalmente superior, pero concentró la carga en pocos grupos funcionales, evidenciando alta variabilidad en tiempos de atención y riesgo de

desbalance operativo ante incrementos en la demanda. La contribución principal de este trabajo radica en demostrar que, en contextos de alta variabilidad en llegadas y tiempos de servicio, la reorganización estructural basada en evidencia simulada puede generar mejoras operativas comparables o superiores a las obtenidas mediante incremento de capacidad física. Este hallazgo es relevante para entidades con restricciones presupuestarias que buscan optimizar recursos existentes antes de invertir en expansión. Además, el modelo desarrollado constituye una plataforma reutilizable para evaluaciones futuras ante cambios en patrones de demanda, políticas de servicio o eventos disruptivos, cambios regulatorios y temas de atención exclusiva por temas de relevancia. Se recomienda que la empresa implemente la Propuesta 1 en un piloto controlado, comenzando con una de las sedes de atención y monitoreando indicadores clave (tiempo de espera promedio, utilización de servidores, satisfacción del usuario) durante al menos tres meses antes de su extensión a otras sedes. Paralelamente, se sugiere recolectar datos sobre abandono de usuarios y percepción subjetiva del tiempo de espera para refinar el modelo en iteraciones futuras. Adicionalmente, sería valioso explorar políticas de priorización dinámicas (ej.: fast-track para trámites simples, citas programadas para casos complejos) que podrían complementar la reorganización estructural propuesta. Trabajos futuros deberían abordar las limitaciones identificadas, en particular: (i) extender el horizonte de simulación para capturar variabilidad estacional, (ii) incorporar modelos de comportamiento del usuario (balking, reneging) mediante datos empíricos de abandono, (iii) realizar análisis costo-beneficio detallado considerando inversiones en capacitación y tecnología, y (iv) validar externamente los resultados mediante comparación con datos operativos post-implementación. Asimismo, explorar el uso de técnicas de optimización basadas en simulación (simulation-based optimization) podría identificar configuraciones aún más eficientes que las evaluadas en este estudio. En síntesis, este trabajo confirma que la simulación de eventos discretos, sustentada en datos operativos reales y principios de verificación y validación rigurosos, proporciona una base sólida para la toma de decisiones gerenciales en entornos complejos y variables característicos de los servicios públicos en Colombia. La simulación permite recrear los comportamientos reales de atención y en el proceso se encuentran acciones de mejora en calidad de datos y mejora de procesos que permiten mejorar los resultados de la simulaciones posteriores para seguir mejorando y aplicando con los cambios solicitados por los entes de control.

## 6. Referencias


[1] J. Bienstock, A. Heuer, Y. Zhang, International Journal of Paramedicine, 1 (2022) 1-15.

[2] F. Shortle, J.M. Thompson, D. Gross, C.M. Harris, Fundamentals of Queueing Theory, 5th ed., Wiley, New York, 2018.

[3] W. Whitt, Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues, Springer, New York, 2018.

[4] R.W. Wolff, Operations Research, 30 (1982) 223-231.

[5] M. Law, Simulation Modeling and Analysis, 6th ed., McGraw-Hill, New York, 2024.

[6] E.S. Hernández Gress, R. Calderón Andrade, Analítica Descriptiva y Estrategias para el Crecimiento Sostenible de las PYMES, Universidad Autónoma de Nuevo León, 2025, pp. 139-159.
[7] J.I. Vázquez-Serrano, N. Peimbert-García, M.L. Crespo-Knopfler, International Journal of Environmental Research and Public Health, 18 (2021) 12262.
[8] W.H. Fun, J. Hooi, A. Cai, P.S. Lim, International Journal of Environmental Research and Public Health, 19 (2022) 2202.
[9] M. Mohamed, H. Abdelrahman, A. El-Adawy, International Journal of Healthcare Management, 14 (2021) 213-221.
[10] J. Deng, C. Wang, X. Liu, Y. Zhang, Applied Sciences, 13 (2023) 5760.
[11] J. Soza-Parra, M.A. Mariño, R. Núñez-Letelier, Simulation Modelling Practice and Theory, 128 (2023) 102692.
[12] C.G. Corlu, A. Akcay, W. Xie, Operations Research Perspectives, 7 (2020) 100162.
[13] C.G. Corlu, B. Biller, Journal of Simulation, 14 (2020) 1-20.
[14] Y. Medjkane, M. Rihane, N. Chikhi, Proceedings of the Operational Research Society Simulation Workshop 2025, The OR Society, 2025.
[15] H. Geng, Manufacturing Engineering Handbook, 2nd ed., McGraw-Hill Education, New York, 2016, Ch. 12.
[16] T. Monks, C.S. Currie, B.S. Onggo, S. Robinson, M. Kunc, S.J.E. Taylor, PLOS ONE, 14 (2019) e0213370.
[17] A. Haghani, M. Hamedi, Transportation Research Record, 1783 (2002) 67-74.
[18] D.J. Daley, D. Vere-Jones, An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods, 2nd ed., Springer, New York, 2003.
[19] A.M. Law, W.D. Kelton, Simulation Modeling and Analysis, 3rd ed., McGraw-Hill, Boston, 2000.
[20] R. Sargent, Proceedings of the 2013 Winter Simulation Conference, IEEE Press, 2013, pp. 342-353.
[21] B. Nelson, Operations Research, 61 (2013) 1313-1329.
[22] D. Kelton, R. Sadowski, N. Zupick, Simulation with Arena, 6th ed., McGraw-Hill Education, New York, 2015.
[23] F.L. Villarreal Satama, M.L. Bernal, D.I. Montenegro Gálvez, Ciencia Latina Revista Científica Multidisciplinar, 5 (2021) 8418-8440.
[24] G.P. Loor Alcívar, S.M. Rodríguez Merchán, O.B. Santos Vásquez, B.J. Loor Alcívar, AlfaPublicaciones, 4 (2022) 22-38.
[25] M. Pereda, Dirección y Organización, 77 (2022) 31-39.

## 7. Nomenclatura

Siglas y Acrónimos

- DES: Discrete Event Simulation (Simulación de Eventos Discretos)
- FCFS: First Come, First Served (Primero en llegar, primero en ser atendido)
- STRESS-DIS: Strengthening the Reporting of Empirical Simulation Studies - Discrete Event Simulation
- V&V: Verificación y Validación

- MLE: Maximum Likelihood Estimation (Estimación por Máxima Verosimilitud)
- AIC: Akaike Information Criterion (Criterio de Información de Akaike)
- BIC: Bayesian Information Criterion (Criterio de Información Bayesiano)
- K-S: Kolmogorov-Smirnov (prueba de bondad de ajuste)
- A-D: Anderson-Darling (prueba de bondad de ajuste)

Variables de Teoría de Colas

- λ: Tasa de llegada de usuarios al sistema (usuarios/unidad de tiempo)
- μ: Tasa de servicio (usuarios atendidos/unidad de tiempo)
- ρ: Factor de utilización del sistema ($\rho = \lambda/\mu$)
- L: Número promedio de usuarios en el sistema
- Lq: Número promedio de usuarios en cola
- W: Tiempo promedio de permanencia en el sistema
- Wq: Tiempo promedio de espera en cola

Parámetros de Distribuciones de Probabilidad

- α, β: Parámetros de forma y escala (distribución Gamma, Weibull)
- μ, σ: Media y desviación estándar (distribuciones Normal, Lognormal)
- k: Número de parámetros del modelo
- n: Tamaño de la muestra
- $\hat{L}$: Verosimilitud maximizada