UNA INTRODUCCIÓN A LA TEORÍA DE COLAS APLICADA A LA GESTIÓN DE SERVICIOS

Marcos Singer [*]
Patricio Donoso[*]
Alan Scheller-Wolf [**]

ABSTRACT

Queuing theory studies the behavior of systems subject to different working conditions, which sometimes force clients to wait for service. It has a vast applicability, since it quantifies the dilemma of many companies and institutions between efficacy (provide a good service) and efficiency (reduce the cost). However, the models do not always have a direct interpretation, reducing their helpfulness. In order to link queuing theory with management, we first explain the relevance of waiting time to the quality of service. Then we identify a number of key performance indicators, related to efficacy, efficiency and the design of the system. For different configurations, we present models that relate these indicators and show examples and applications in factories, service, logistics and health organizations.

Keywords: queuing theory, service, efficiency, applications.
JEL Classification: C44, L60, L8, Y2

RESUMEN

La teoría de colas estudia el comportamiento de los sistemas de atención sujetos a diferentes condiciones de funcionamiento, en que los clientes a veces deben esperar por el servicio. Su aplicabilidad es muy amplia, pues cuantifica el dilema de muchas empresas e instituciones entre la eficacia (dar un buen servicio) y la eficiencia (mantener bajos los costos). Sin embargo, los modelos no siempre tienen una interpretación directa, haciendo que pierdan utilidad práctica. Para vincular la teoría de colas a la gestión de las organizaciones, en primer término explicamos la relevancia del tiempo de espera en la calidad del servicio. A continuación identificamos un conjunto de indicadores de desempeño, relacionados con la eficacia, con la eficiencia y con el diseño del sistema. Para diferentes configuraciones, presentamos modelos que vinculan estos indicadores y mostramos ejemplos y aplicaciones en empresas productivas, de servicio, de logística y de salud.

Palabras clave: teoría de colas, servicio, eficiencia, aplicaciones.

[*] Profesor de la Escuela de Administración de la Pontificia Universidad Católica de Chile.
[**] Profesor Tepper School of Business, Carnegie Mellon University

La teoría de colas es la rama de la investigación de operaciones que estudia el comportamiento de los sistemas de atención, en que los clientes eventualmente esperan por el servicio. Su fundador es el matemático danés Agner Erlang (1878-1929), quien aplicó en 1909 la teoría de las probabilidades al comportamiento de las conversaciones telefónicas. Este y otros trabajos permitieron comprender y controlar las redes de telefonía, cuyos altos costos obligaban a asignar de manera óptima los componentes electrónicos para mantener los tiempos de espera dentro de estándares aceptables. Actualmente, no obstante el costo del hardware es relativamente bajo, la teoría de colas sigue siendo relevante para las telecomunicaciones. Por ejemplo, orienta la administración de los centros de llamadas (*call-centers* en inglés), una industria que emplea aproximadamente un 3% de la fuerza laboral de EE.UU. y del Reino Unido y que crece a una tasa anual de 20% (Koole y Mandelbaum, 2002). Hoy en día los costos están principalmente determinados por el personal empleado, que para algunos servicios requiere un alto grado de especialización técnica y por ende su costo es significativo.

Los modelos de colas apoyan la toma de decisiones del centro de llamados al identificar y relacionar los indicadores de desempeño de interés del administrador (por ejemplo, la capacidad instalada) y los de interés de sus clientes (por ejemplo, el tiempo de espera). Los modelos también ayudan a mejorar la calidad del servicio, estimando e informando al cliente cuánto tiempo debe esperar hasta ser atendido. Salvo cuando el requerimiento de servicio es de extrema urgencia, a veces las personas valoran más la puntualidad que la rapidez. Después de realizar una encuesta a sus clientes, una compañía que distribuye gas licuado (Sección C.B) decidió cambiar su estrategia competitiva: desde intentar ser la más rápida en despachar a ser la más confiable. Incluso en el caso de las llamadas a la policía de menor gravedad, los ciudadanos prefieren que se les informe verazmente que el radiopatrulla llegará en una hora más, a que se les deje en la incertidumbre y que se les atienda mucho antes de la hora (Larson, 1987).

La aplicabilidad de la teoría de colas es muy amplia en la administración de las organizaciones, pues el dilema entre la eficacia (dar un buen servicio) y la eficiencia (hacerlo con pocos recursos) es universal. Sin embargo, la formulación de los modelos es con frecuencia críptica. Las variables y parámetros no siempre tienen una interpretación directa al mundo concreto, haciendo que los modelos pierdan utilidad práctica. El objetivo de este artículo es vincular la teoría de colas a la gestión de las empresas y organizaciones. Para ello, la Sección I explica el impacto del tiempo de espera en la calidad del servicio. La Sección II identifica un conjunto de indicadores del desempeño de eficacia y eficiencia. La Sección III muestra la simulación de un caso real de cómo interactúan estos parámetros. La Sección IV caracteriza el comportamiento aleatorio de los clientes. La Sección V vincula los indicadores de desempeño en términos del comportamiento de los clientes. Finalmente, la Sección VI presenta un resumen de los principales conceptos expuestos.

I. EL IMPACTO DE LA ESPERA EN EL SERVICIO

La opinión que se forman los clientes acerca del servicio que se les entrega depende de diversos aspectos subjetivos, tales como la capacidad técnica de quienes atienden, la amabilidad del trato, la presentación y la limpieza. Su evaluación depende de la ejecución del servicio y de sus expectativas (Maister, 1985). Sólo si la ejecución supera las expectativas los clientes quedan conformes. El mal servicio perjudica la reputación de la

firma mucho más que el buen servicio la favorece. Mientras los clientes satisfechos le informan a un promedio de cinco personas acerca de su experiencia positiva, los clientes molestos le informan su desagrado a un promedio de 19 personas (Ittner, 1996).

Un aspecto determinante para la calidad del servicio es el tiempo que se debe esperar para obtenerlo. El tiempo se divide en dos componentes: el lapso de servicio y el tiempo de espera. En general se prefiere tiempos de atención breves, si bien algunos servicios (consulta médica, peluquería) demandan un lapso mínimo. Casi siempre el costo psicológico del tiempo de espera es mucho mayor que el del lapso de servicio. Este aspecto fue pasado por alto en el diseño del Transantiago, el actual sistema de transporte público de la ciudad de Santiago. Para reducir la congestión, y por ende los tiempos de viaje, se optó por aumentar la capacidad de los autobuses que circulan en las principales vías (troncales) desde 65-80 pasajeros a 160 pasajeros por autobús (articulado). Sin embargo, dividir por dos el número de autobuses implica bajar a la mitad su frecuencia. Si los horarios en que pasan por los paraderos no son precisos, el pasajero duplica el tiempo de espera, lo cual menoscaba significativamente su percepción del servicio.

Quizás el principal motivo por el cual el costo psicológico de esperar es tan alto, es que usualmente los clientes no ocupan su tiempo mientras esperan (Mobach, 2007)[1]. Para evitar esta pérdida de tiempo, en algunos casos se adelanta el inicio de la atención. Algunos restaurantes hacen pasar a sus clientes a la barra mientras esperan por una mesa. Muchos bancos colocan pantallas de televisión o habilitan exhibiciones de arte. Como modo de distracción para quienes están esperando el ascensor, los edificios suelen instalar en sus accesos espejos de cuerpo entero[2].

Una manera de aliviar la molestia del cliente es respetar ciertas normas de justicia, como por ejemplo, que se atienda primero a quienes llevan más tiempo esperando. En el caso de los sistemas de una sola cola, esto implica utilizar la política FIFO (*first in first out*), es decir, atender en orden de llegada. Para evitar que nadie se adelante, se puede ordenar físicamente a los clientes o se puede utilizar dispensadores que asignan números de atención. Cuando el sistema no garantiza esta norma, los clientes pueden incurrir en altos costos para evitar ser adelantados. Ése es el caso de algunos centros de distribución que atienden a los camiones por orden de llegada. Supongamos que un camión A y un camión B se dirigen al centro por una misma ruta. Si el conductor del camión B observa que el camión A le lleva una pequeña delantera, podría acelerar con el objeto de adelantarlo. Ante tal amenaza, el conductor del camión A aumentará su velocidad por encima de un límite prudente, lo que redundará en mayores costos de combustible y riesgos de accidentes. Cualquiera sea el resultado de esta carrera, ambos camiones llegarán a su destino casi al mismo tiempo, lo que obligará al segundo a esperar en cola más de lo necesario.

En los sistemas de varias colas, como los que ocupan los restaurantes de comida rápida o los supermercados, no se puede garantizar la norma de atender primero a quienes llevan más tiempo esperando, porque cualesquiera de las colas puede atrasarse fortuitamente. Acciones destinadas a reducir el tiempo de espera pueden, en ocasiones, producir un efecto de injusticia. Cuando en un supermercado se abren cajas adicionales, quienes están más adelante en la cola, y por ende han invertido más tiempo de espera, usualmente son

[1] Casualmente este párrafo lo escribí estando en una gran sala de espera de un centro médico. Comprobé personalmente este hecho. Los pacientes resignadamente esperaban; casi nadie leía, hablaba por celular o leía un libro, mucho menos trabajaba en su *laptop*.

[2] Para muchas personas el mirarse, y mirar a otros, es muy entretenido.

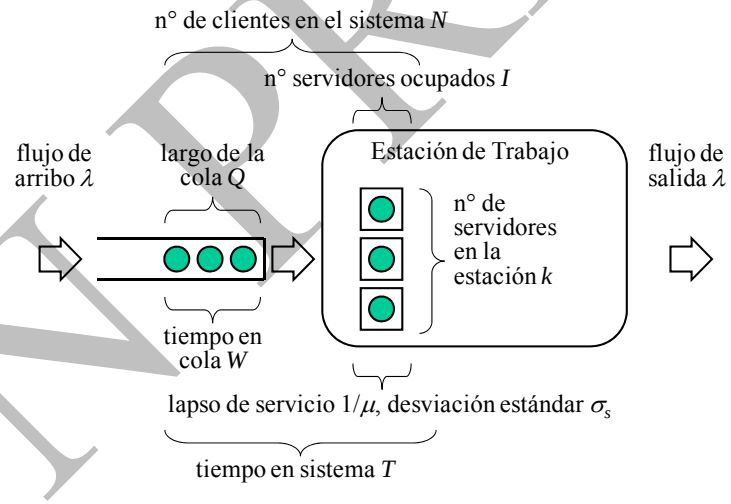
adelantados por quienes acaban de llegar. Para evitar esta situación, algunos supermercados utilizan una cola única. Sin embargo, la mayoría de ellos dispone mostradores de productos en el espacio de cola frente a cada caja, lo cual no podría hacerse si se ocupa una cola única.

II. INDICADORES DE GESTIÓN DE SISTEMAS DE ESPERA

El modelo de la Figura 1 muestra una estación de trabajo con una línea de espera previa. En una fábrica, esta línea se interpreta como un inventario intermedio (*work in process* o WIP en inglés). En las empresas de servicio, la línea de espera corresponde a la cola de clientes. En ambos casos este inventario o cola se denomina *buffer* ("amortiguador" en inglés), pues absorbe la variabilidad de la llegada de clientes a la estación de trabajo y la variabilidad del lapso de servicio. En lo que sigue, supondremos que se sigue la política FIFO de atender en orden de llegada, si bien algunas propiedades se dan para cualquier política de atención. También supondremos que no hay abandono, es decir, todos los clientes que llegan son atendidos.

FIGURA 1
MODELO DE SISTEMA DE ESPERA

Consiste en un número de servidores y en una cola de clientes. Los valores N, λ y Q, entre otros, definen las características del sistema.



A partir del modelo podemos identificar los indicadores de gestión más relevantes. Los clasificaremos de acuerdo al *Balanced Scorecard* (Kaplan y Norton, 1996; Nagar y Rajan, 2005), una de las herramientas de control de gestión de mayor difusión en la actualidad. Los indicadores clave de desempeño (*key performance indicators* o KPI en inglés) son ordenados según cuatro perspectivas:

- relacionados con los procesos internos, incluidos en la Sección *A*;
- de interés de los accionistas o dueño de la empresa, incluidos en la Sección *B*
- de interés de los clientes, incluidos en la Sección *C*
- relacionados con la proyección de futuro, no incluidos en este análisis.

A. Parámetros de Diseño (Procesos Internos)

Definimos los siguientes parámetros de diseño del sistema de espera:

- λ [u/h]: flujo promedio o tasa de recepción de órdenes de atención. Para fijar ideas, lo mediremos en unidades por hora [u/h].
  Es inversamente proporcional al lapso entre el arribo de dos unidades de flujo consecutivas a la estación de trabajo, por lo que a mayor flujo de recepción, menor el lapso entre dos llegadas consecutivas:
  $$\lambda \text{ [u/h]} = 1/ \text{ lapso entre llegadas consecutivas [h/u]}.$$
  En régimen regular, la cola no puede crecer ni disminuir ilimitadamente, por lo que el flujo de llegada es idéntico al de salida. Este flujo (*throughput* en inglés) de entrada-salida mide la "potencia" de atención del sistema, y por ende es uno de los indicadores más relevantes.
- μ [u/h]: flujo promedio o tasa de atención de cada servidor cuando opera a máxima capacidad, medido en unidades por hora.
  El flujo promedio es inversamente proporcional al lapso entre dos atenciones consecutivas de un servidor dado si está constantemente ocupado, por lo que a mayor tasa de atención, menor es el tiempo que requiere el servidor para ejecutar su trabajo:
  $$\mu \text{ [u/h]} = 1/ \text{ lapso de servicio [h/u]}.$$
- σ_s [h/u]: desviación estándar del lapso de servicio. Por definición es la raíz cuadrada de la varianza, que es igual al valor esperado de las diferencias al cuadrado entre cada una de las observaciones y su promedio.
  La desviación estándar puede interpretarse de la siguiente manera: si en la práctica el lapso de servicio es de "3 más-menos 2 horas" entonces $1/\mu = 3$ [h/u] y $\sigma_s = 2$ [h/u] aproximadamente[3].
- k: número de servidores o "recursos" de la estación de trabajo. Por lo tanto, k es una medida de capacidad de atención.
  Para que la cola no se vuelva infinita, debe ocurrir que el flujo de atención máximo debe ser superior al flujo de recepción, es decir:
  $$\lambda \text{ [u/h]} < k\ \mu \text{ [u/h]}.$$

B. Indicadores de Interés del Administrador (de Eficiencia)

Los indicadores a continuación relacionan la utilización de los recursos invertidos con la cantidad de clientes que están en proceso de atención.

- I: número de servidores ocupados en el sistema.[4] También se puede interpretar como el inventario o el número de órdenes siendo procesadas en la estación de trabajo de la Figura 1. El valor promedio o esperado de I se representa como $\bar{I}$.[5]

---

[3] Esto no debe confundirse con los "límites de control" que podrían definir que en promedio una variable debe tomar el valor 3 y no debe superar 3 + 2 = 5 ni ser menor que 3 – 2 = 1. En el ejemplo, la desviación estándar 2 es la "diferencia promedio" entre la variable y su valor medio, pero puede ser mayor o menor que 2.

[4] I es una variable aleatoria, por lo tanto puede tomar diversos valores, cada uno con una cierta probabilidad. Si la variable es continua, y por ende puede tomar un número infinito de valores, entonces la probabilidad de cada valor es infinitesimal, la que es medida por una llamada función de densidad.

Este inventario promedio se relaciona con el flujo de salida y con el lapso de servicio a través de la *Ley de Little* (1961), posiblemente la fórmula más importante de la teoría de colas:

Inventario promedio = flujo de salida · lapso servicio.

Para ejemplificar la Ley de Little, supongamos que una cajera de supermercado atiende a un flujo promedio de dos personas por minuto (2 [p/m]). Si el cliente promedio enfrenta una cola (incluyendo a quien se está atendiendo) de 10 personas (10 [p]), entonces el lapso de servicio esperado en minutos [m] es (Inventario promedio / flujo de salida) = 10 [p] / 2 [p/m] = 5 [m].

En el caso que estamos estudiando, la Ley de Little se traduce en que:

$$\bar{I} = \lambda \frac{1}{\mu} = \frac{\lambda}{\mu}.$$

- ρ: *factor de utilización*. Es igual a la razón entre el número promedio de servidores ocupados y el número total de servidores:

$$\rho = \frac{\bar{I}}{k} = \frac{\lambda}{k\,\mu}.$$

Tanto I como ρ miden la utilización de recursos, la que conviene maximizar desde el punto de vista de la eficiencia. Por ejemplo, centros de llamadas altamente productivos logran factores de utilización cercanos a 90% o 95% (Koole y Mandelbaum, 2002).

Los siguientes indicadores se relacionan con la cantidad de clientes en el sistema de espera.

- Q [u]: tamaño de la cola de espera medido en unidades [u]. Su valor promedio es $\overline{Q}$.
- N [u]: población de clientes en el sistema medido en unidades [u], es igual a:

$$N = Q + I.$$

El valor promedio de N es $\overline{N} = \overline{Q} + \bar{I}$.

Los indicadores Q y N miden el volumen de clientes o de inventario en proceso, lo que determina el tamaño de la infraestructura necesaria para alojarlos. Una estación de servicio donde no caben más de diez automóviles está sujeta a la restricción $N \leq 10$, lo que la hace perder clientes cuando la demanda crece. Una línea de producción cuyo buffer entre dos estaciones de trabajo tiene una capacidad máxima de cinco toneladas obliga a detener la producción "aguas arriba" cuando se repleta el buffer.

C. Indicadores de Interés del Cliente (de Eficacia)

Los indicadores a continuación se relacionan con la evaluación objetiva del servicio entregado a los clientes.

- P_0: probabilidad de que el sistema esté vacío. Si sólo hay un servidor P_0 también es el porcentaje de clientes que son atendidos de inmediato.

[5] Consideramos la distribución de probabilidad (o la función de densidad) de I (y de las variables aleatorias que definiremos más abajo) en estado estacionario, es decir, cuando la cola está en régimen.

- P_n: probabilidad de que hayan exactamente n clientes en el sistema.
- $P(N > n)$: probabilidad de que hayan n o más clientes en el sistema. Esta probabilidad mide cuán probable es que el sistema esté sobrecargado de clientes, obligando a muchos de ellos a esperar más de lo tolerable. En algunos casos un número excesivo de clientes los obliga a abandonar el sistema, como lo explicamos para el caso de una estación de servicio. En otros casos, la espera por la atención perjudica irreparablemente al cliente. En Inglaterra, los servicios de ambulancia están diseñados para evitar que menos del 5% de los pacientes esperen más de 14 minutos. Esto es, $P(N > n) \leq 5\%$, donde el n-ésimo paciente en la cola espera menos que 14 minutos y el n+1-ésimo paciente espera 14 minutos o más.
- W [h]: tiempo de espera de cada cliente antes de ser atendido por la estación de trabajo, medido en horas.
  En tanto λ [u/h] < $k\ \mu$ [u/h] la cola no crece indefinidamente, así es que la tasa de llegada a la cola λ [u/h] es igual a la tasa promedio de salida. Por la Ley de Little:
  $$\overline{W} = \frac{\overline{Q}}{\lambda}.$$
  El tiempo promedio de cola $\overline{W}$ depende de la configuración del sistema. La Sección V estudia algunos de los casos más importantes.
- $P(W > 0)$: probabilidad de espera, es decir, probabilidad de que un cierto cliente deba esperar en cola.
  Esta probabilidad evalúa la primera impresión que se lleva el cliente al llegar al sistema. La primera impresión es crucial, pues predispone positiva o negativamente al cliente respecto de toda su experiencia de servicio (Maister, 1985). Según Koole y Mandelbaum (2002), $P(W > 0)$ en los centros de llamadas es aproximadamente 50%, así es que la mitad de las personas son atendidas de inmediato. Goldberg y Szidarovszky (1991) y Borras y Pastor (2002) explican la importancia de este indicador para los sistemas de ambulancias. En la medida en que $P(W > 0)$ crece, se restringe la asignación de una emergencia a una base cualquiera. Perder flexibilidad impide atender la emergencia desde la base más cercana, lo cual aumenta el lapso de servicio.
- γ: grado de servicio del sistema. Mientras mayor es su valor, menor es $P(W > 0)$ y menor es $\overline{W}$. En general debe interpretarse como una constante en las fórmulas que miden el desempeño del sistema. Con tales fórmulas se puede despejar cómo cambia un cierto parámetro cuando los otros se modifican.
- $F_W(t)$: probabilidad de que la espera en cola de un cierto cliente sea menor o igual al lapso t. Esta medida es importante cuando existe un umbral por encima del cual esperar es insoportable para el cliente. Por ejemplo, es muy improbable arrestar a un criminal si la policía arriba al sitio del suceso más de diez minutos después de haberse reportado el crimen (Larson, 1987). En tal caso F_W (10 minutos) representa el porcentaje de crímenes para los cuales probablemente se producirá un arresto inmediatamente después de la llamada a la policía.
- $E(W / W > 0)$: tiempo promedio de espera cuando llega un cierto cliente, condicionando a que dicho cliente deba esperar. Esta medida es relevante cuando no obstante la espera promedio es baja, por ejemplo porque existe un gran número de servidores, si llega a ocurrir una espera porque están ocupados todos los servidores, ésta podría ser muy perjudicial.

- $E(Q \mid Q > 0)$: número esperado de clientes en cola cuando llega un cierto cliente, condicionando a que existan clientes en cola cuando él llega.
- T [h]: permanencia total en el sistema de un cliente, igual a la suma del tiempo de espera W más el lapso de servicio. Recordando que $N = Q + I$ es el inventario total del sistema, por la Ley de Little:

$$\overline{T} = \frac{\overline{N}}{\lambda} = \frac{\overline{Q}}{\lambda} + \frac{\overline{I}}{\lambda} = \overline{W} + \frac{1}{\mu}.$$

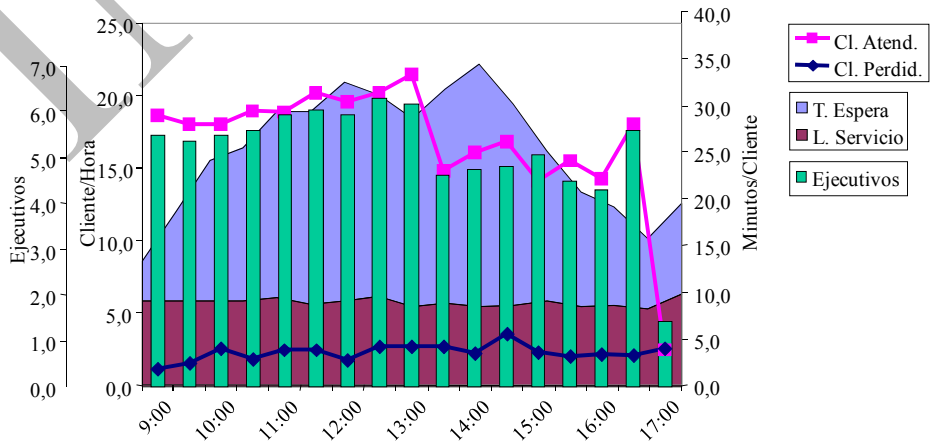
- $F_T(t)$: probabilidad de que la espera en el sistema completo de un cierto cliente sea menor o igual al lapso t.

III. RELACIONES DE TRANSACCIÓN EMPÍRICAS-SIIMULADAS

A continuación presentamos un ejemplo de cómo interactúan los parámetros de diseño, los indicadores del administrador y los indicadores del cliente a partir de los datos de una empresa chilena que maneja 27 sucursales de atención a clientes. La Figura 2 muestra el comportamiento de una de estas sucursales. La línea de cuadrados y la línea de rombos muestran el flujo de clientes que son atendidos y el flujo de los clientes que se pierden por abandono respectivamente. El flujo de atención en la mañana es de unos 20 clientes por hora y en la tarde es de unos 15 clientes por hora. El flujo de clientes perdidos es alrededor de 2 clientes por hora durante todo el día. El flujo λ de entrada-salida del sistema es la suma de ambos flujos. La ilustración también muestra el lapso de servicio $1/\mu$ que es algo menor que 10 minutos por cliente durante todo el día. El tiempo de espera W es muy variable, creciendo a más de 25 minutos por cliente a las 14:00. A esa hora la permanencia total T es de 35 minutos por cliente. Las barras muestran el número de ejecutivos k, que en la mañana pasan de 5,5 a 6 y en la tarde caen a menos de 5.

FIGURA 2
COMPORTAMIENTO DE UNA SUCURSAL

Las barras muestran el número de ejecutivos. Las líneas punteadas muestran los clientes atendidos y perdidos. Las áreas muestran los tiempos de espera y de servicio.

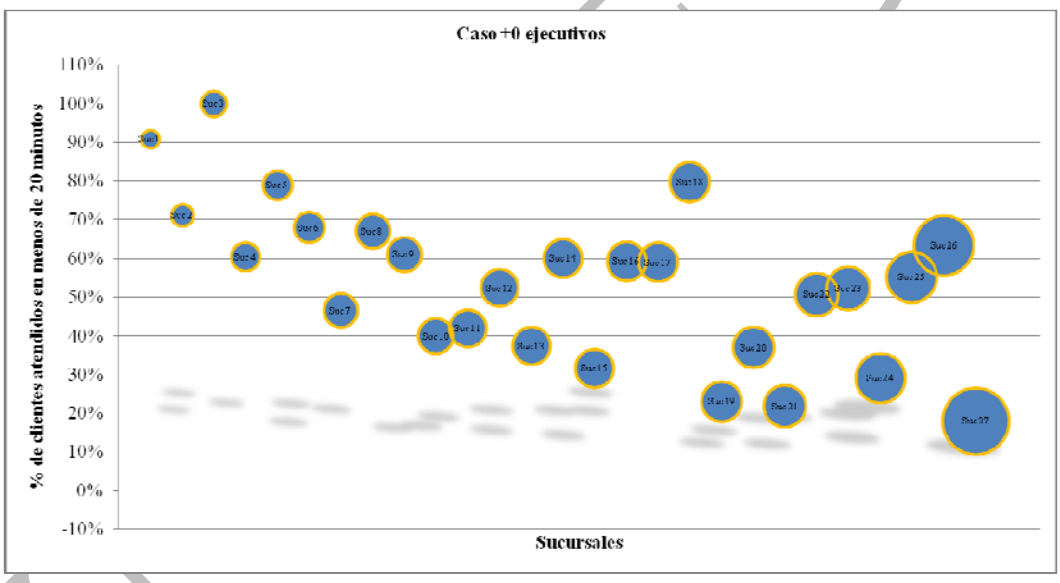


Al momento del estudio, la empresa era la segunda en participación de mercado, pero cambios en el escenario competitivo la estaban exponiendo a una fuga masiva de clientes. Como estrategia de defensa optó por reforzar la calidad del servicio, imponiéndose un estándar de que la permanencia total *T* fuera de a lo más 20 minutos por cliente.

La Figura 3 muestra el desempeño de la empresa por cada una de sus 27 sucursales. El tamaño de cada circunferencia representa el número relativo de clientes atendidos y su altura muestra el porcentaje de ellos a los que se les cumple el estándar de permanencia. Se observa una fuerte heterogeneidad de calidad de servicio: mientras la Sucursal 3 cumple para el 100% de sus clientes, la sucursal 27 cumple sólo para el 18%. En general, mientras mayor es el número de clientes de la sucursal peor es su nivel de cumplimiento.

FIGURA 3
CASO BASE DE RAPIDEZ DE ATENCIÓN

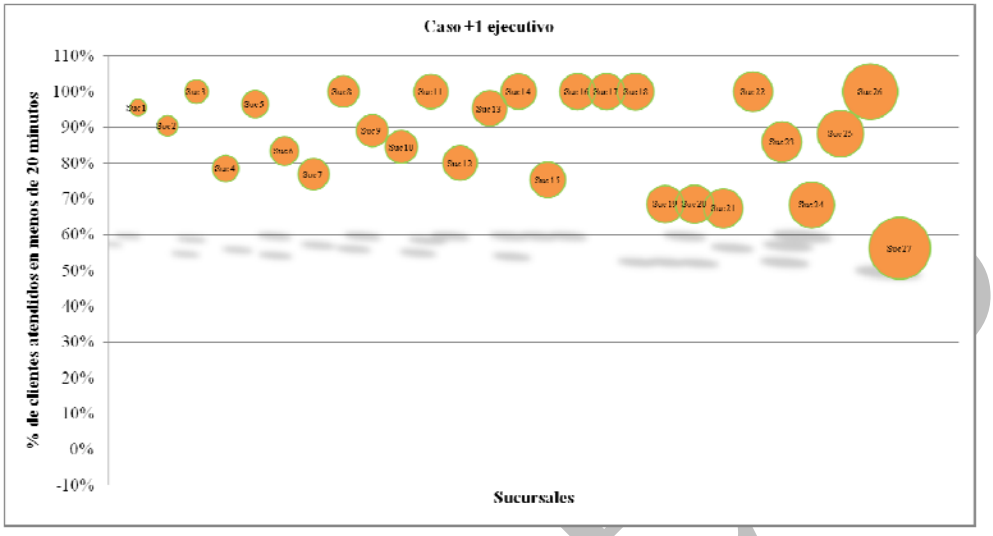
El tamaño de cada circunferencia representa el número relativo de clientes y su altura muestra el porcentaje a los que se les cumple el estándar de permanencia.



Evaluamos los efectos de cambios en los parámetros, por ejemplo del número *k* de ejecutivos, mediante simulación computacional implementada en el software Extend. Calibramos la simulación de manera de que sus resultados coincidan con los datos empíricos disponibles. La Figura 4 muestra el efecto de agregar un ejecutivo en cada una de las sucursales. En algunas el efecto es insignificante; en varias otras más que duplica el porcentaje de clientes a los cuales se les cumple el estándar.

FIGURA 4
CASO CON UN EJECUTIVO ADICIONAL POR SUCURSAL

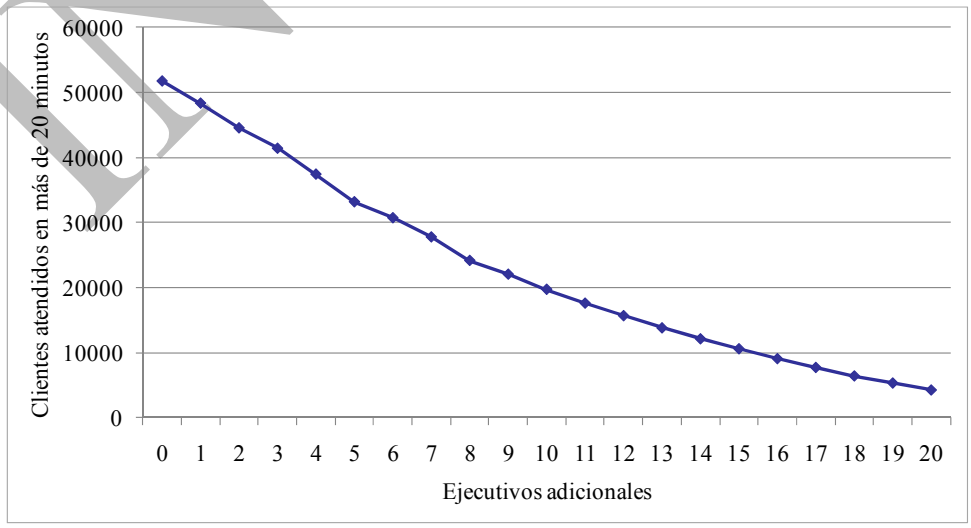
En algunas sucursales el efecto es insignificante; en otras se duplica el cumplimiento del estándar de servicio.



La Figura 5 muestra el efecto de agregar ejecutivos a las sucursales, siguiendo una secuencia que prioriza el aumento de dotación en las sucursales que muestran un efecto más significativo. La curva es convexa, es decir, la mejora marginal tiende a caer en la medida en que se agregan más ejecutivos. Con esta curva la empresa puede evaluar el costo y beneficio nuevas contrataciones.

FIGURA 5
EFECTO DE LA ADICIÓN DE MÁS EJECUTIVOS

El efecto se muestra priorizando las sucursales que muestran un efecto más significativo en la mejora del servicio por cada ejecutivo adicional.



Si bien la simulación entrega resultados confiables, es compleja de implementar. Primero se debe programar el software, luego hay que obtener los datos empíricos y finalmente utilizarlos para calibrar la simulación. Como alternativa, la teoría de colas puede entregar resultados aproximados pero con herramientas más expeditas. A continuación presentamos los elementos de la teoría, partiendo por cómo se modela la demanda de servicio por parte de los clientes.

IV. PROCESO EXPONENCIAL Y DISTRIBUCIÓN DE POISSON

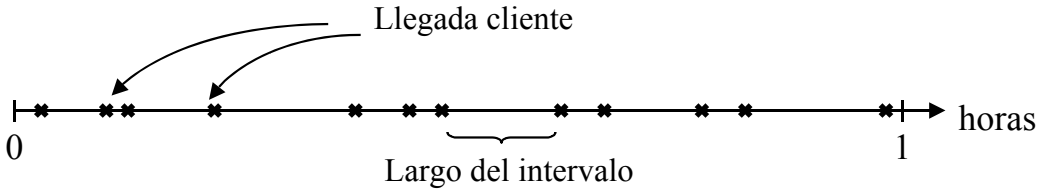
Tanto los indicadores del administrador como los del cliente dependen del comportamiento de dos tipos de eventos inciertos: la llegada de un cliente y la ejecución de una atención. La variable aleatoria que mide el intervalo entre la ocurrencia de dos eventos consecutivos tiene una distribución de probabilidad *exponencial* cuando se dan las siguientes condiciones:

- El número de eventos que ocurren es proporcional al intervalo de tiempo que se considera. Por ejemplo, las ocurrencias durante dos semanas duplican a las ocurrencias durante una semana.
- No pueden ocurrir dos o más eventos de manera simultánea.
- La ocurrencia de un evento no influencia la ocurrencia de un evento posterior.

Muchos sistemas de servicio (puestos de peaje, cajeros automáticos, plantas telefónicas, servicios de emergencia) muestran un proceso de llegada de clientes aproximadamente exponencial, similar al de la Figura 6. En todos ellos la probabilidad de que llegue un cliente en un cierto instante es constante. Por ejemplo, si la probabilidad de que llegue un cliente entre el segundo 0:02:34 y el segundo 0:02:35 es 0,001 entonces la probabilidad de llegada entre el segundo 0:48:12 y el segundo 0:48:13 también es 0,001. Dado que esta probabilidad no se ve modificada si inmediatamente antes ocurrió una llegada, los procesos exponenciales se denominan "sin memoria". Los procesos que muestran estas características son los que cumplen las siguientes condiciones. Primero, los clientes llegan uno por uno, no en grupos. Segundo, no se influencian unos a otros; cada llegada es independiente de la otra. Tercero, el número de llegadas en un lapso es, en promedio, proporcional al tamaño del lapso.

FIGURA 6
PATRÓN DE LLEGADA EXPONENCIAL

La probabilidad de que llegue un cliente en un cierto instante es constante, es decir, una llegada no influencia la ocurrencia de otras.

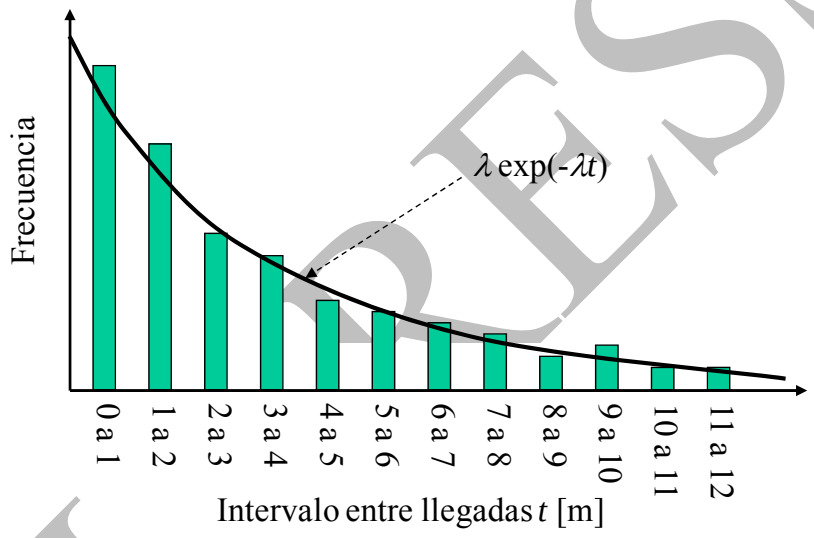


La tasa de llegada λ mide la cantidad promedio de eventos por intervalo de tiempo. En la Figura 6, λ es aproximadamente 12 clientes por hora [cl/h]. El lapso promedio entre llegadas consecutivas es $1/\lambda$, que en el ejemplo es igual a 1/12 [h/cl].

Una manera de comprobar que el proceso de llegada tiene una distribución de tipo exponencial es realizando un histograma de los intervalos entre arribos como el de la Figura 6. El histograma debería ser similar a la función de densidad de la distribución exponencial: $\lambda \exp(-\lambda t)$, donde t es el largo del intervalo. La bondad del ajuste para cada intervalo se estima usando el estadígrafo χ^2 (Singer y Donoso, 2008).

FIGURA 7
HISTOGRAMA DE TIEMPO ENTRE ARRIBOS CONSECUTIVOS

Si el proceso de llagada es exponencial, las barras de frecuencia deberían seguir un patrón definido.



Para obtener la probabilidad de que un intervalo se encuentre dentro de un cierto rango, se integra la expresión $\lambda \exp(-\lambda t)$ entre los márgenes del rango. Por ejemplo:

- Probabilidad de que un lapso sea menor que t: $\int_0^t \lambda e^{-\lambda t} dt = 1 - \exp(-\lambda t)$.
- Probabilidad de que un lapso esté entre t y u: $\int_t^u \lambda e^{-\lambda t} dt = \exp(-\lambda t) - \exp(-\lambda u)$.

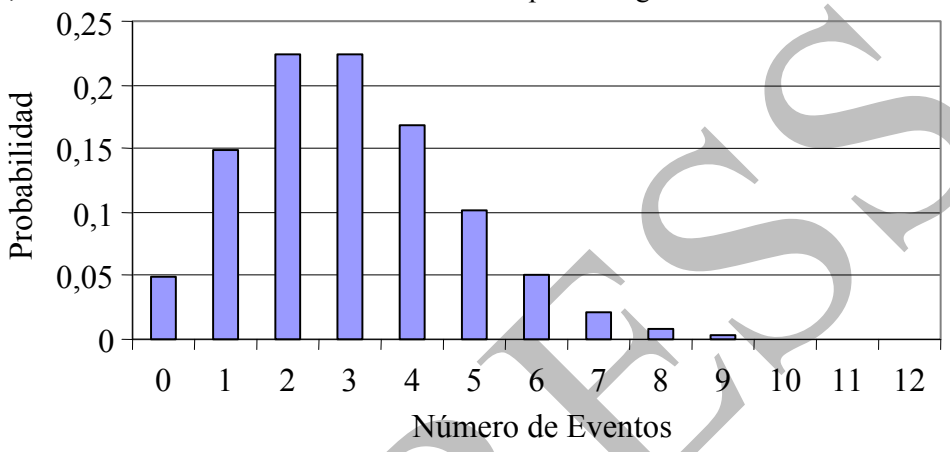
Supongamos que se tienen registros históricos de las llamadas telefónicas a una central que muestran que $\lambda = 3$ [llamadas/hora]. Si observamos el proceso durante una hora, no existe garantía de que ocurran exactamente tres llamadas: es posible que ocurran dos o cuatro o más. Si los intervalos entre llamadas tienen una distribución exponencial, el número de llamadas n durante un cierto intervalo t tiene una distribución de probabilidad llamada de *Poisson*, dada por:

$$\text{Probabilidad que en un lapso de tamaño } t \text{ hayan } n \text{ eventos} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

La Figura 8 grafica la probabilidad de ocurrencia de distintos números de eventos, suponiendo que $\lambda = 3$ [llamadas/hora]. Por ejemplo, la probabilidad de que ocurran dos llamadas durante una hora es 0,224.

FIGURA 8
PROBABILIDAD DEL NÚMERO DE LLAMADAS EN UNA HORA

En general, la distribución de las barras forma una campana alargada hacia la derecha.



Una de las particularidades de los procesos exponenciales es que la desviación estándar σ del lapso de llegada es idéntica al promedio de dicho lapso igual a $1/\lambda$. En algunos sistemas de espera, como por ejemplo los de accesos a servidores de Internet, los intervalos de servicio muestran una distribución exponencial. Si se graficara la frecuencia en que el servidor estuvo trabajando entre 0 y 1 minuto, entre 1 y 2 minutos, entre 2 y 3 minutos y así sucesivamente, se obtendría un histograma similar al de la Figura 7. Si se grafica el ritmo de trabajo en términos de órdenes por minuto procesadas, se obtendría un gráfico similar al de la Figura 8.

V. MODELOS DE ESPERA ARQUETÍPICOS

En 1953 David Kendall clasificó los sistemas de espera mediante la nomenclatura *A*/*B*/*k*, donde

- *A*: tipo de distribución de probabilidad de tiempo entre arribos consecutivos;
- *B*: tipo de distribución de probabilidad del tiempo de servicio o atención;
- *k*: número de servidores de la estación de trabajo.

Las distribuciones de probabilidad más estudiadas son

- *M*: distribución exponencial o sin memoria (la "M" viene del inglés "*memory-less*");
- *D*: tiempos determinísticos o constantes;
- *G*: denota cualquier tipo de distribución.

Por ejemplo, un sistema *M*/*D*/3 indica que el sistema tiene una llegada exponencial de clientes, que cada servidor tiene un tiempo de servicio constante y que cuenta con 3 servidores.

La nomenclatura de Kendall se extiende a *A*/*B*/*k*/*C*/*N*/*D*, donde:

- *C*: capacidad total del sistema, con $C \geq k$. Al igual que en el ejemplo de la estación de servicio, una capacidad limitada puede hacer que se pierdan clientes;
- *N*: tamaño de la población desde la que se obtienen los clientes;
- *D*: disciplina o política de prioridad con la que se atienden los clientes, la que determina diversos aspectos de calidad del servicio.

Omitir "/*C*/*N*/*D*" significa que $C = \infty$, $N = \infty$ y *D* es FIFO.

Hacer $C = \infty$ equivale a suponer que los clientes siempre se ponen a la cola, no importa qué tan larga sea ésta. En la práctica, por otro lado, los clientes estiman la espera de la cola y deciden quedarse dependiendo de un cierto valor umbral (Pazgal y Radas, 2008). Una vez que están en la cola, es raro que la abandonen si no han cambiado las condiciones de servicio. Zohar, Mandelbaum y Shimkin (2002) estudian las colas de los centros de llamadas y demuestran, teórica y empíricamente, que la tasa de abandono es proporcional al tiempo de espera $\overline{W}$.

La disciplina FIFO cumple con la norma de justicia de atender primero a quien lleva esperando más tiempo, aunque no necesariamente es la disciplina más aconsejable. Para la policía y para otros servicios de emergencia puede ser más efectiva la disciplina LIFO (*last in first out*), que atiende primero al último llamado (Larson, 1987). Tal como señalamos anteriormente, la probabilidad de realizar un arresto es casi nula si se llega al sitio del crimen después de diez minutos de recibida la denuncia. Salvar a quien sufre un infarto al corazón es posible sólo si se lo auxilia durante los primeros minutos de la emergencia. En ambos casos atender los llamados de acuerdo a FIFO podría significar llegar tarde casi siempre si el sistema está congestionado. Otras disciplinas usadas en la práctica son SIRO (*service in random order*), que consiste en atender en orden aleatorio y *Round-Robin*, que atiende a cada cliente un cierto período y lo devuelve a la cola si el servicio no ha sido completado.

A continuación describimos los modelos de espera más comunes, que permiten vincular los indicadores de desempeño internos, del administrador y del cliente del sistema.

A. Sistema M/M/1

Un sistema *M*/*M*/1 tiene una llegada de clientes de tipo exponencial con tasa λ y un tiempo promedio de servicio igual a $1/\mu$. Este sistema cumple las siguientes ecuaciones (Pazos, Suárez & Díaz, 2003):

- $\overline{Q} = \frac{\rho^2}{1-\rho}$, donde $\rho = \frac{\lambda}{\mu}$ muestra el grado de ocupación del servidor.
- $\overline{N} = \frac{\rho}{1-\rho}$.
- $\overline{W} = \frac{\overline{Q}}{\lambda} = \frac{\rho}{\mu(1-\rho)}$.

- $\overline{T} = \dfrac{\overline{N}}{\lambda} = \dfrac{1}{\mu - \lambda}$.
- $P_0 = (1 - \rho)$.
- $P_n = (1 - \rho)\, \rho^n$.
- $P(N > n) = \rho^n$.
- $E(W / W > 0) = \dfrac{\overline{W}}{\rho}$.
- $E(Q / Q > 0) = \dfrac{\overline{Q}}{\rho^2}$.

Las ecuaciones muestran que a menos que ρ esté por debajo de 1, el sistema colapsa y el servicio se deteriora irremediablemente. Lo mismo ocurre con los otros sistemas estudiados más adelante. Por lo tanto, siempre se requiere algo de ociosidad, cuyo valor promedio óptimo depende de la calidad del servicio que se desea entregar y del costo de operación.
Suponiendo que la disciplina de atención es FIFO, se cumple además:

- $F_W(t) = 1 - \rho \exp(-(1 - \rho)\, \mu\, t)$, con $t \geq 0$.
- $F_T(t) = 1 - \exp(-(1 - \rho)\, \mu\, t)$, con $t \geq 0$.

B. Sistema M/M/k

Un sistema $M/M/k$ tiene una llegada de clientes de tipo exponencial con tasa λ y un tiempo de servicio también exponencial de tasa μ para cada uno de sus k servidores. Con $k = 1$ corresponde a $M/M/1$ descrito en la sección anterior. Recordando que $\bar{I} = \lambda/\mu$ y que $\rho = \bar{I} / k$, para este sistema se cumplen las siguientes ecuaciones:

- $P_0 = \left(\sum_{n=0}^{k-1} \dfrac{\bar{I}^n}{n!} + \dfrac{\bar{I}^k}{k!} \cdot \dfrac{1}{1-\rho} \right)^{-1}$
- $P_n = \dfrac{\bar{I}^n}{n!}\, P_0, \qquad \forall\ n \in [1,k].$
- $P_n = \rho^{n-k} P_k = \dfrac{\bar{I}^n}{k^{n-k}\, n!}\, P_0, \quad \forall\ n \geq k.$
- $\overline{Q} = \dfrac{\bar{I}^k\ \rho}{k!\ (1-\rho)^2} P_0.$
- $\overline{W} = \dfrac{\overline{Q}}{\lambda} = \dfrac{\bar{I}^k}{\mu\ k\ k! (1-\rho)^2} \left(\sum_{n=0}^{k-1} \dfrac{\bar{I}^n}{n!} + \dfrac{\bar{I}^k}{k!} \cdot \dfrac{1}{1-\rho} \right)^{-1}.$

Suponiendo que la disciplina de atención es FIFO, se cumple:

- $F_W(t) = 1 - \dfrac{P_k}{1-\rho} \exp(-k\,(1 - \rho)\, \mu\, t)$, con $t \geq 0$.
- $E(W / W > 0) = \dfrac{1}{\gamma \cdot \sqrt{2}}$.

Las siguientes expresiones son aproximaciones, las que son más precisas con valores bajos de γ, es decir, cuando el sistema está relativamente congestionado.

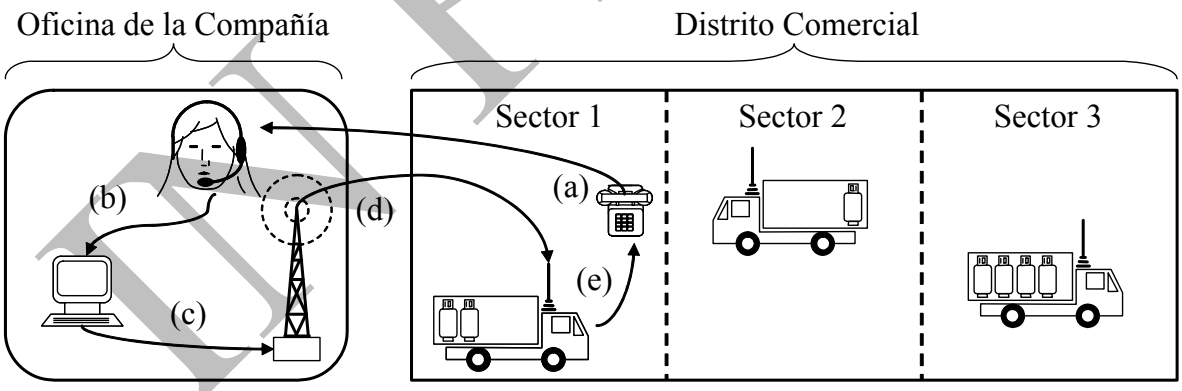
- $P(W=0) = \dfrac{P_k}{1-\rho} \cong \gamma = (1-\rho)\ \sqrt{k}$.
- $\overline{W} \cong \dfrac{1-\gamma}{\gamma\sqrt{k}}$.

Las expresiones muestran "economías de escala" respecto del número de servidores. No obstante un servidor de capacidad 4μ induce el mismo factor de utilización $\rho = \lambda / (k\ \mu)$ que cuatro servidores de capacidad μ, los cuatro servidores obtienen un grado de servicio γ dos veces mejor. Por muy rápido que sea un único servidor, éste puede ser bloqueado por un cliente complicado y perjudicar con ello a quienes están en la cola. Al haber cuatro servidores, el bloqueo de uno de ellos no tiene un efecto tan significativo.

La economía de escala de combinar varios servidores recomienda consolidar colas de espera cuando ello es posible. Singer, Donoso & Jara (2002) estudian la configuración de una flota de reparto de gas licuado en cilindros, cuyos camiones reciben las órdenes de entrega en línea mientras están en ruta. La compañía agrupa a sus clientes en distritos comerciales, cada uno atendido por un centro de distribución. El distrito se subdivide en sectores que son asignados por un cierto camión, tal como lo sugiere la Figura 9. Primero, los clientes llaman (a) a un centro de llamados que ingresa la información (b) a un sistema que determina, de acuerdo con la localización del cliente, cuál camión debiera atender ese pedido (c). La orden de visita es transmitida al camión (d) que finalmente entrega el producto al cliente (e).

FIGURA 9
ESQUEMA DE ATENCIÓN DE PEDIDOS

Los clientes llaman, un sistema designa un camión, la orden es transmitida al camión que entrega el producto.



La configuración descrita genera una cola de pedidos en cada sector, la que es atendida por su respectivo camión. Como alternativa, se podría crear una cola única de pedidos para cada distrito comercial. Si bien cada camión se concentraría en su propio sector, podría traspasarse a un sector contiguo si el camión correspondiente está detenido o bloqueado por algún cliente. Utilizando simulación computacional, se comprueba que consolidar colas de tres camiones reduce el tiempo promedio de espera de los clientes $\overline{W}$ en tres minutos. Suponiendo que los clientes abortan el pedido luego de un cierto tiempo de espera, la cola

única reduce el número de pedidos perdidos en 85%, lo que representa un incremento en las ventas del 3%.

C. Sistema M/G/1

Un sistema *M*/*G*/1 tiene una llegada de tipo exponencial de clientes de tasa λ, la atención de cada uno de ellos toma un tiempo de servicio con media $1/\mu$ y varianza σ_s^2 y cuenta con sólo un servidor. Cuando la distribución *G* del lapso de servicio es constante, es decir el tiempo de atención es fijo, entonces $\sigma_s^2 = 0$. Para este sistema se cumplen las siguientes ecuaciones:

- $\rho = \lambda \dfrac{1}{\mu} = 1 - P_0$. Esto es válido para cualquier *G*/*G*/1.
- $\overline{Q} = \dfrac{\lambda^2 \sigma_s^2 + (\lambda/\mu)^2}{2(1-\lambda/\mu)}$.
- $\overline{N} = \overline{Q} + \rho$.
- $\overline{W} = \dfrac{\overline{Q}}{\lambda}$.
- $\overline{T} = \dfrac{\overline{Q} + \rho}{\lambda}$.

La fórmula para $\overline{Q}$ pone en evidencia la importancia de σ_s y por ende la utilidad de estandarizar el tiempo de proceso. Muchos supermercados definen cajas "Express" especializadas en clientes rápidos. Suponiendo que existen dos tipos de clientes, los que se atienden rápido y los que se atienden lento, agrupar a cada tipo en su caja respectiva reduce σ_s^2 respecto de la situación en que los clientes se asignan de manera arbitraria. Una segunda ventaja de esta técnica es que evita que un cliente demandante obligue a muchos clientes rápidos a esperar por éste.

Lo anterior explica por qué la gestión de la calidad de los procesos prescribe estandarizar procedimientos (Deming 1986 p. 321; Manz y Stewart, 1997). Algunas metodologías, tales como Seis Sigma o el Control Estadístico de Procesos (*Statistical Process Control* o SPC en inglés), son especialmente enfáticas en detectar y corregir la variabilidad de los procesos.

D. Sistema M/G/k

Un sistema *M*/*G*/*k* tiene un llegada de clientes de tipo exponencial con tasa λ, la atención de cada uno de ellos toma un tiempo de servicio con media $1/\mu$ y varianza σ_s^2 y cuenta con *k* servidores. Para este sistema se cumple (Nozaki & Ross, 1978):

- $$\overline{W} \cong \frac{\left(1/\mu^2 + \sigma_s^2\right) \lambda \left(\lambda/\mu\right)^{k-1}}{2\,(k-1)!\,\left(k-\lambda/\mu\right)^2} \cdot \left(\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{(k-1)!\,\left(k-\lambda/\mu\right)} \right)^{-1}$$
$$= \frac{1+\sigma_s^2\,\mu^2}{2} \cdot \frac{\bar{I}^k}{\mu\, k\, k!\,(1-\rho)^2} \cdot \left(\sum_{n=0}^{k-1} \frac{\bar{I}^n}{n!} + \frac{\bar{I}^k}{k!} \cdot \frac{1}{1-\rho} \right)^{-1}.$$

Esta última expresión es idéntica a $\overline{W}$ para el modelo *M*/*M*/*k*, multiplicada por el factor $(1 + \sigma_s^2\mu^2)/2$. Recordando de la Sección II que si la distribución del lapso de servicio es exponencial entonces la desviación estándar σ_s es $1/\mu$, la estimación de $\overline{W}$ para *M*/*G*/*k* es exacta para *M*/*M*/*k*. Si el tiempo de servicio es constante, por lo que $\sigma_s^2 = 0$, la espera en cola se reduce a la mitad que la generada por un sistema *M*/*M*/*k*.

En la mayoría de los sistemas de espera el número de servidores es el principal ítem de costo (entre un 60% y 70% en los centros de llamadas, según Koole & Mandelbaum, 2002). Una de las fórmulas más prácticas para dimensionar el número necesario de estaciones de trabajo es (Puhalskii & Reiman, 2000):

- $k = \frac{\lambda}{\mu} + \gamma\sqrt{\frac{\lambda}{\mu}}$.

Recordando que $\lambda/\mu = \bar{I}$ es el número promedio de servidores ocupados en el sistema, la expresión $\gamma\sqrt{\lambda/\mu}$ se denomina "dotación de seguridad", es decir, la dotación que se mantiene ociosa en promedio, con el objeto de sostener un grado de servicio γ.

E. Sistemas G/M/k, G/D/k y G/G/k

En los sistemas cuyo arribo de clientes muestra una distribución general de probabilidad, definimos σ_a como la desviación estándar del lapso entre dos arribos consecutivos. Para un sistema *G*/*M*/*k* relativamente congestionado, es decir, cuyo γ es relativamente bajo, se cumple que:

- $P(W = 0) \cong \gamma = \frac{(1-\rho)\sqrt{k}}{\sqrt{(\lambda_a^2\sigma_a^2 + 1)/2}}$.

Para los sistemas *G*/*D*/*k* relativamente congestionados se cumple que:

- $P(W = 0) \cong \gamma = \frac{(1-\rho)\sqrt{k}}{\lambda_a\sigma_a}$.

Con valores altos de γ y de *k*, se cumple la siguiente aproximación para *G*/*G*/*k* (Whitt, 1992):

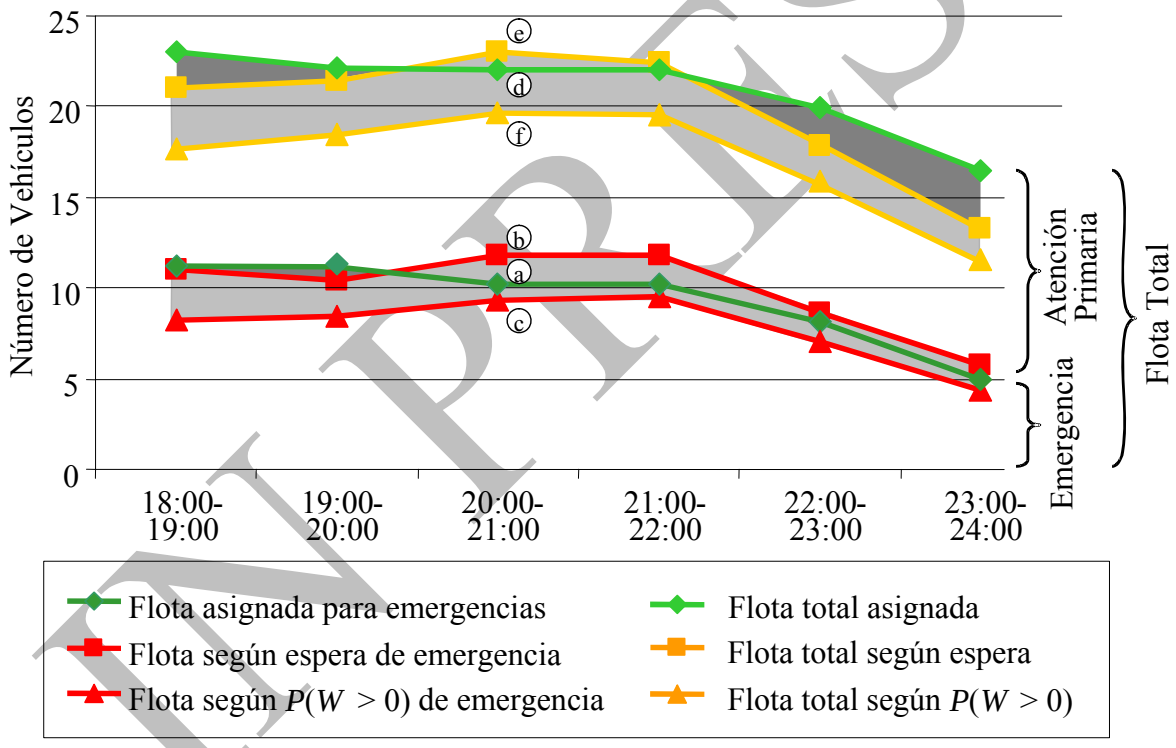
- $P(W = 0) \cong \gamma = (1-\rho)\sqrt{k}$.

Singer & Donoso (2008) utilizan estas fórmulas para evaluar si un servicio de emergencia alcanza los estándares de desempeño que le corresponde al número de ambulancias que están de turno. Tal como lo explican Green & Kolesar (2004), con frecuencia un vehículo que se supone disponible, en realidad está cargando combustible, en reparación o fuera de servicio porque la tripulación se está alimentando. En algunas ciudades se pierde hasta el 60% del tiempo de patrullaje de policía en estas y otras actividades. Tampoco se cumple con rigurosidad el inicio y el término de los turnos de trabajo. Lo anterior obliga a determinar empíricamente cuántos vehículos están realmente disponibles para atender llamados, en función de indicadores de interés del cliente, tales como la espera promedio $\overline{W}$ y la proporción $P(W > 0)$ efectiva de llamados obligados a esperar.

La Figura 10 compara la flota programada por la empresa con las estimaciones de *k* derivadas del desempeño observado durante el período en estudio. El punto (a) representa una flota total de diez ambulancias programadas para llamados de emergencia entre las 20:00 y 21:00. El punto (b) muestra que para lograr el tiempo de espera $\overline{W}$ observado, en teoría hubo doce ambulancias, así es que la flota logra un sobre-desempeño. El punto (c) muestra que para lograr la probabilidad $P(W > 0)$ observada, en teoría hubo nueve ambulancias, así es que la flota muestra un sub-desempeño. Los puntos (d), (e) y (f) son análogos a (a), (b) y (c), pero considerando la flota total, esto es, la destinada a emergencia más la destinada a atención primaria (llamados de baja prioridad).

FIGURA 10
ESTIMACIÓN DE LA FLOTA EFECTIVA

Se compara la flota programada para emergencias y para atención primaria con las atenciones de *k* de acuerdo a modelos de espera alimentados por el desempeño observado.



Los datos revelan que para los llamados de emergencia se obtiene un desempeño dentro de los márgenes esperados. Para el total de llamados, se produce una pérdida de recursos efectivos del orden de dos vehículos de un total de 23 en el rango de las 18:00 a 19:00, dos vehículos de 20 entre las 22:00 y 23:00 y tres de 17 entre las 23:00 y 24:00.

VI. CONCLUSIONES

La teoría de colas apoya la gestión de las empresas y organizaciones que atienden público, cuantificando la manera en que se combinan los indicadores de efectividad (calidad del servicio), de eficiencia (uso de recursos) e internos (de diseño del sistema). Dependiendo de las características específicas de cada sistema, las fórmulas muestran de manera estilizada las relaciones de transacción entre estos indicadores. Algunas fórmulas muestran cómo aumenta la probabilidad de que el cliente sea atendido de inmediato cuando se habilitan nuevas estaciones de trabajo. Otras indican cuántas estaciones deben estar ociosas en promedio para ofrecer un determinado tiempo de espera promedio a los clientes. Todas estas fórmulas son fácilmente definibles en una planilla de cálculo, lo que permite tomar decisiones rápidas con relativamente pocos datos.

Si bien la teoría de colas es muy usada para evaluar los sistemas de servicio, existen alternativas tales como la simulación computacional o la prueba y error. En la Sección III mostramos el ejemplo del uso de la simulación para evaluar el efecto de cambios en la dotación de ejecutivos de una empresa que maneja sucursales de atención a clientes. Alternativamente, la empresa podría apelar a la prueba y error: aumentar o disminuir ejecutivos en cada sucursal y posteriormente medir el resultado en el servicio. Aunque tanto la simulación como la prueba y error son más realistas que un análisis basado en fórmulas, ambas técnicas son complejas de implementar, complejidad que desdibuja las relaciones de causa y efecto que vinculan las decisiones con sus resultados. Por el contrario, la teoría de colas hace explícitas las relaciones de causalidad, lo que permite ganar una mayor comprensión de los sistemas estudiados.

La comprensión cabal redunda en decisiones más efectivas para lograr mejoras en la calidad del servicio. Por ejemplo, los sistemas de atención muestran economías de escala respecto del número de servidores, así es que es más conveniente disponer de n servidores de capacidad unitaria que un solo servidor de capacidad n. Concentrar los requerimientos en una cola única también puede mejorar la calidad del servicio, porque evita que el bloqueo de una cierta estación perjudique a los clientes que están esperando en su cola respectiva. Cualquiera sea el diseño del sistema, los tiempos de espera ocurren por la variabilidad de la llegada y del lapso de servicio, lo cual implica que regularizar ambos procesos redunda en un mejor servicio.

Además de la habilitación de recursos adicionales, existen diversas maneras de mejorar la percepción subjetiva del servicio. Por ejemplo, la disciplina FIFO resguarda que se atienda primero a quien lleva más tiempo esperando. También es posible adelantar la entrega del servicio a quienes están en la cola, de manera de transformar el tiempo de espera (de alto costo psicológico) en lapso de servicio (de menor costo).

Decisiones correctas en el diseño de los sistemas permiten entregar un mejor servicio sin necesariamente habilitar recursos adicionales. Por lo tanto, la teoría de colas puede ser una herramienta competitiva que le permite a la organización entender y por ende acceder a su frontera de posibilidades.

REFERENCIAS

Borras, F. and Pastor, J.T. (2002). The ex-post evaluation of the minimum local reliability level: An enhanced probabilistic location set covering model, *Annals of Operations Research*, 111 (1-4), 51-74.

Deming, W.E. (1986) *Out of the crisis*. Cambridge: MIT Press.

Goldberg, J.B. and Szidarovszky, F. (1991). Methods for solving nonlinear equations used in evaluating emergency vehicle busy probabilities, *Operations Research*, 39 (6), 903-916.

Green, L.V. and Kolesar P.J. (2004). Improving Emergency Responsiveness with Management Science, *Management Science*, 50 (8), 1001–1014.

Ittner, C.D. (1996). Exploratory evidence on the behavior of quality costs, *Operations Research*, 44 (1), 114-130.

Kaplan, R. and Norton, D. (1996) *The Balanced Scorecard*. Boston: Harvard Business School Press.

Koole, G. and Mandelbaum, A. (2002). Queuing models of call centers: An introduction. *Annals of Operations Research*, 113 (1-4), 41-59.

Larson, R.C. (1987). Perspectives on queues social justice and the psychology of queuing. *Operations Research*, 35 (6), 895-905.

Little J.D.C. (1961). A proof of the queuing formula $L = \lambda W$, *Operations Research*, 9 (3), 383-387

Maister, D.A. (1985) "The psychology of waiting lines" en Czepiel, J.A., Soloman, M.R. and Surprenant, C.F. (eds.) *The Service Encounter: Managing Employee/Customer Interaction in Service Business*. Lexington Books, Lexington, MA

Manz, C.C. and Stewart, G.L. (1997). Attaining flexible stability by integrating total quality management and socio-technical systems theory, *Organization Science*, 8 (1), 59-70.

Mobach, M.P. (2007). Consumer behavior in the waiting area, *Pharmacy World and Science*, 29 (1), 3-6.

Nagar, V. and Rajan, M.V. (2005). Measuring customer relationships: The case of the retail banking industry, *Management Science*, 51 (6), 904-919.

Nozaki, S.A. and Ross, S.M. (1978). Approximations in finite capacity multi-server queues with Poisson arrivals, *Journal of Applied Probability*, 15 (4), 826-834.

Pazgal, A.I. and Radas, S. (2008). Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study, *Computers & Operations Research*, 35 (8), 2537-2548.

Pazos, J.J., Suárez, A. and Díaz, R.P. (2003) *Teoría de Colas y Simulación de Eventos Discretos*. Madrid: Pearson Educación

Puhalskii A.A. and Reiman, M.I. (2000). The multiclass GI/PH/N queue in the Halfin–Whitt regime, *Advances in Applied Probability*, 32 (2), 564-595.

Singer, M. and Donoso, P. (2008). Assessing an Ambulance Service with Queuing Theory, *Computers and Operations Research*, 35 (8), 2549-2560.

Singer, M., Donoso, P. and Jara S. (2002). leet Configuration Subject to Stochastic Demand: An Application in the Distribution of Liquefied Petroleum Gas, *Journal of the Operations Research Society*, 53 (9), 961-971.

Whitt, W. (1992). Understanding the efficiency of multi-server service systems, *Management Science*, 38 (5), 708-723.

Zohar, E., Mandelbaum A. and Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support, *Management Science*,48 (4), 566-583.