¿Es posible **reducir el tiempo**

Óscar Córdoba Rodríguez, Marco A. de la Lama Zubirán,

de espera **en las colas?**

Marcelo del Castillo Mussot y Alfredo de la Lama García

Los ciudadanos son cada vez más conscientes del valor del tiempo (trabajo, transporte, recreación, etcétera), por lo que les irrita verse obligados a esperar para obtener un producto o un servicio, sin importar si se trata de una oficina gubernamental o privada, el cine, el estadio, el supermercado, la sucursal bancaria o la ventanilla universitaria. Todos los clientes son reticentes a hacer filas o colas.

¿Sería posible diseñar un sistema que mida el tiempo de atención al usuario y que además le permitiera predecir el comportamiento de la demanda y, en consecuencia, poder actuar sobre la oferta antes de que los tiempos de espera se acrecienten?, ¿sería posible desarrollar uno sencillo y barato que pudiera se usado por cualquier empresa dispuesta a ofrecer un mejor servicio a sus clientes, sin importar su tamaño?

Para poder encontrar la respuesta a estas preguntas es crucial saber si es posible ofrecer al usuario o cliente de un servicio un tiempo promedio de atención que corresponda a los niveles de calidad previamente establecidos. Para ello, primero necesitamos saber si es posible medir el tiempo promedio de espera de los clientes, y después ver qué variables se pueden manipular y así poder disminuir este tiempo de espera. Los sistemas de colas son muy comunes en la sociedad. La adecuación de estos sistemas puede tener un efecto importante sobre la calidad de vida y la productividad. Para estudiarlos la teoría de colas formula modelos matemáticos que tratan de predecir la forma en que los clientes llegan a los establecimientos, representar su operación, para después usar estos modelos y obtener medidas de desempeño.

La aparición de nuevas teorías de la organización que enfocan su atención en los deseos e intereses de los clientes o usuarios del servicio que se ofrece no es nueva, y está ligada a las teorías del control total de calidad, conocidas como TQC por sus siglas en inglés.

La teoría de colas es el estudio matemático de las filas o colas, y se relaciona, en una primera aproximación, con el área de la física de la dinámica de fluidos, con la que po-

demos ligar los correspondientes tiempos de espera. Imaginamos así a las personas como un medio continuo que se desplaza en cierta dirección al interior de algún ducto, donde la fila hace las veces del ducto y el fluido son las personas.

El proceso básico supuesto por la mayor parte de los modelos de colas es el siguiente. Los clientes que requieren un servicio aparecen a lo largo del tiempo en una fase de entrada, ingresan al sistema y se unen a una cola. En determinado momento, mediante alguna regla conocida como disciplina de servicio, se selecciona a un miembro de la cola para proporcionarle el servicio. Luego se lleva a cabo el servicio requerido por el cliente por medio de un mecanismo de servicio, después de lo cual el cliente sale del sistema de cola. Con tales conceptos se elaboran los modelos matemáticos, que tratan de predecir desde la llegada de los clientes, con qué distribución lo hacen y cuál es la de su salida, pasando por el mecanismo de servicio. Sin embargo no siempre son adecuados los modelos o se vuelven muy complicados. Nosotros consideramos que, con base en los conceptos básicos de número de clientes, operadores, tiempo de espera y número de operaciones por clientes, hay dos formas sencillas de medir los tiempos de espera en las filas.

En este problema se puede aplicar, al igual que en un tubo de corriente de algún fluido, la ecuación de continuidad:

$$\nabla \cdot (?\overline{u}) = -\frac{\partial ?}{\partial t}$$

donde $?$ es la densidad del fluido, t el tiempo y $\overline{u}$ la velocidad.

Esta expresión nos dice que el cambio en el tiempo de la densidad de un fluido $?$ es igual al cambio en el flujo de la densidad de la corriente. Pero como para un flujo permanente de una tubería la masa del fluido que pasa por cualquier sección de la tubería es constante, tenemos que:

$$\frac{\partial}{\partial t} = 0 \text{ y por tanto, } \nabla \cdot (?\overline{u}) = 0 .$$

El siguiente ejemplo nos ayuda a visualizar la semejanza del problema con el fenómeno de fluidos. Supongamos una serie de pelotas, las cuales caen en un recipiente que tiene una salida en el fondo del mismo. En esta salida hay un diablito (en alusión a Maxwell, quien diseñó un modelo similar para estudiar el comportamiento de las moléculas de un gas) que sólo permite la salida de las bolas a una determinada velocidad. Dadas estas condiciones, el tiempo de permanencia de las pelotas en el cubo estaría dado por la relación entre el número de pelotas que entran y la velocidad con que el diablito las deja salir.

Existirían tres situaciones: a) si el número de pelotas que caen es igual al número de pelotas que el diablito deja salir, entonces tendremos un tiempo de espera constante, alrededor del cual las pelotas tienden a permanecer dentro del cubo; b) si aumenta el número de bolas que caen dentro del recipiente y el diablito mantiene constante la velocidad de salida, que es menor a la entrada de las pelotas, entonces las pelotas se acumularán en el recipiente, su tiempo de permanencia aumentará y su opinión sobre la calidad del servicio decaerá; c) existe una cantidad de pelotas previas en el recipiente, si el número de pelotas que entra al recipiente disminuye y el diablito mantiene constante la velocidad a la que las deja salir, que es mayor a la velocidad a que entran, entonces el tiempo de espera de las pelotas dentro del recipiente tenderá a disminuir.

Como se puede apreciar por las tres situaciones, cuando el tiempo de salida es constante, el tiempo de espera estará determinado por el número de pelotas que entran al recipiente. Entre mayor sea el número de bolas que entra al recipiente, mayor será su tiempo de espera y viceversa.

Pasemos a una situación diferente, donde el diablito se interesa en lograr que el tiempo de salida de las pelotas varíe para lograr un tiempo de permanencia constante de las bolas dentro del cubo. Supongamos que el diablito conoce de antemano el número de pelotas que caen y además tiene libertad para adecuar la velocidad de salida de las bolas, entonces, dependiendo del número de pelotas que caigan, podrá controlar mediante la aceleración o el retardo de la salida de las pelotas el tiempo que las pelotas permanecen en el cubo.

De lo anterior se desprenden los siguientes comportamientos: d) si el número de pelotas que entra disminuye, el diablito a su vez disminuirá la velocidad de salida; e) a la inversa, al aumentar el número de pelotas que entra al cubo, el diablito apresura la salida de las bolas, de tal manera que el tiempo de permanencia pueda mantenerse más o menos constante. En otras palabras, si el sistema es capaz de controlar la salida de las pelotas y además conoce previamente el número que entrará, entonces será capaz de adecuar el sistema para que haya una velocidad constante de salida; de esta manera se podrá mantener constante el tiempo de permanencia de las pelotas dentro de la cubeta.

Si este sencillo modelo se adecúa al comportamiento real, aunque simplificado, de una ventanilla o caja de pagos, entonces podría ser la base para combinar sabiamente costos y calidad del servicio. La clave, por tanto, consiste en determinar el tiempo promedio de espera del cliente.

Métodos de optimización

Tomando en cuenta el modelo anterior y la ecuación de continuidad, proponemos dos métodos para medir el tiempo promedio de espera del cliente. El primero tiene como variables el número de clientes que llegan a la sucursal y el número de cajeros que atienden; el segundo asocia las velocidades de llegada y salida.

Método de oferta y demanda. Determinaremos el tiempo promedio de espera del cliente asociándolo directamente con el número clientes que llegan (número de pelotas en el recipiente) y el comportamiento de los operadores (número de pelotas que deja salir el diablito). Si el fenómeno de atención al público se comporta de manera estable, entonces si se sabe a cuántos clientes atiende de cada uno de los cajeros, es posible establecer un tiempo promedio de atención por usuario sin medirlos uno por uno.

Supongamos que hay D clientes y O cajeros, entonces el tiempo promedio de espera T del cliente está dado por la siguiente ecuación:

$$T = \frac{\sum_{i=0}^{n} t_i \lambda_i}{n} \quad \text{..... (1)}$$

Donde t_i es el tiempo de espera del cliente i, λ_i las operaciones del cliente i, $n = \left[\frac{D}{O}\right]$ y $[\ ]$ denota la función mayor entero (mayor número entero inferior o igual a la fracción). Es decir, simplemente se suman los tiem-

pos de espera de todos los clientes y se dividide entre el número de clientes.

Notemos que la persona que está el principio de la fila realmente no tiene que esperar casi nada de tiempo, mientras que las personas que vienen detrás tienen que esperar más tiempo. Es decir, las condiciones iniciales son cruciales.

Suponiendo que el tiempo de atención por operación de los cajeros es el mismo (t) y que en promedio los clientes hacen λ operaciones, podemos pensar en lo siguiente: la primera persona no espera para ser atendida, el tiempo que le lleva realizar su operación y esperar es $t\lambda$. La segunda persona tarda el tiempo en que la primera persona sea atendida ($t\lambda$) más el tiempo que tarden en atenderla ($t\ \lambda$), por lo que el tiempo total de espera de las segunda persona es $2t\lambda$, para la tercer persona será $3t\lambda$ y así sucesivamente. La suma del tiempo de las n primeras personas es:

$$t\lambda + 2t\lambda + 3t\lambda + \ldots + nt\lambda = t\lambda(1+2+3\ldots+n) = t\lambda\,[n(n+1)/2].$$

Por tanto, de la ecuación (1) obtenemos:

$$T = \frac{\frac{n(n+1)}{2}}{n} t? = \frac{(n+1)}{2} t? \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

De esta ecuación podemos deducir que hay tres maneras de reducir el tiempo de espera: a) disminuyendo el tiempo de operación; o b) el número de transacciones por cliente; o bien c) aumentando el número de cajeros.

Obviamente los mecanismos a) y b) son más difíciles de cambiar; una reducción en el tiempo de operación del cajero podría llevarlo a cometer errores, y no es factible condicionar a los clientes en el número de operaciones que pueden realizar. Así, la opción viable es disminuir n, es decir aumentar el número de operadores.

Este método se aplica muy bien cuando hay interrupciones en el flujo (cuando es no estacionario), porque tenemos una fracción importante de clientes que está al principio de cada nuevo débito, como cuando se acciona un gotero después de llenarlo.

Método de velocidades de flujo. Este método supone que se conoce la velocidad de entrada de los clientes y la de salida, de tal modo que a partir de estas velocidades podemos saber el tiempo que pasó una persona en la fila.

Sea N el número de personas que llega a la fila a una velocidad V_0 y que salen de la fila a una velocidad V_s. El tiempo de espera de cada persona dependerá del tiempo que tarde en entrar y del tiempo que tarden en salir, siendo la diferencia de estos dos tiempos el tiempo de espera para cada persona. El tiempo en que la persona en el lugar j tar-

dará en entrar (t_{ej}) está dado por las que ya entraron antes y su velocidad de entrada:

$$t_{ej} = \frac{P_j}{V_0} \quad \text{.....(3)}$$

Donde P_j es la posición de la persona j en la fila de espera.

El tiempo en que la persona en el lugar j tardará en salir (t_{sj}) está dado por las pelotas que le precedieron en salir y la velocidad de su salida.

$$t_{sj} = \frac{P_j}{V_s} \quad \text{.....(4)}$$

Así, el tiempo que una persona en el lugar j tardará en ser atendida (T_j), está dado por la diferencia de estos dos tiempos

$$T_j = t_{ej} - t_{sj} \quad \text{.....(5)}$$

De tal modo que si lo que queremos es encontrar el tiempo promedio de espera, sólo necesitamos encontrar el promedio (T) de los tiempos T_j:

$$T_p = \frac{\sum_{j=i}^{n} t_{ej} - t_{sj}}{n} = \frac{\sum_{j=1}^{n} T_j}{n} \quad \text{.....(6)}$$

Como estamos pensando que la velocidad de entrada y la de salida es la misma para todas las personas, dejaremos el tiempo promedio de espera en términos de la velocidad de entrada, la velocidad de salida y el numero de clientes; para eso, necesitamos modificar la ecuación (6) de la siguiente manera:

$$T_p = \frac{\sum_{j=1}^{n} t_{sj} - t_{ej}}{n} = \frac{\sum_{j=1}^{n} t_{sj} - \sum_{j=1}^{n} t_{ej}}{n} \quad \text{.....(7)}$$

Si desarrollamos la primera suma de esta ecuación tenemos que la primera persona tarda en salir $1/V_s$, la segunda persona tarda en salir $2/V_s$, la tercera persona tarda en salir $3/V_s$, y la persona en el lugar n tarda en salir n/V_s. Tenemos entonces que:

$$\sum_{j=1}^{n} t_{sj} = \frac{1}{V_s} + \frac{2}{V_s} + \frac{3}{V_s} + + \frac{n}{V_s} = \frac{1+2+3++n}{V_s} =$$

$$\frac{\frac{n(n+1)}{2}}{V_s} = \frac{n(n+1)}{2V_s} \quad \text{..... (8)}$$

En la ecuación (7), donde ahora la primera persona tardará en entrar $1/V_0$, la segunda persona tarda $2/V_0$, etcétera.

$$\sum_{J=1}^{n} t_{ej} = \frac{1}{V_0} + \frac{2}{V_0} + \frac{3}{V_0} \cdot + \frac{n}{V_0} = \frac{1+2+3++n}{V_0} =$$

$$\frac{\frac{n(n+1)}{2}}{V_0} = \frac{n(n+1)}{2V_0} \text{ ..... (9)}$$

De las tres ecuaciones anteriores resulta entonces:

$$T_p = \frac{\frac{n(n+1)}{2V_s} - \frac{n(n+1)}{2V_0}}{n} = \frac{\frac{2n(n+1)(V_0 - V_s)}{4V_s V_0}}{n} =$$

$$\frac{(n+1)(V_0 - V_s)}{2(V_s V_0)} \text{ .....(10)}$$

En esta última ecuación se aprecia que para reducir el tiempo promedio de espera necesitamos disminuir el número de personas, la velocidad de entrada o aumentar la velocidad de salida. Pero al igual que en el modelo pasado, no podemos restringir el número de personas que llegan, tampoco la velocidad a la que lo hacen, por lo que para reducir el tiempo promedio de espera debemos aumentarle velocidad de salida.

Conclusiones

Para aplicar estos modelos es necesario contar con datos precisos sobre el número de clientes que llega a los establecimientos, el número de operaciones por cliente y el número de operadores. Teniendo una base de datos estadísticos de diferentes días y horarios se puede tomar las decisiones necesarias para dar una mejor atención a los usuarios. Es importante establecer las horas pico para los diferentes días de la semana. Por ejemplo no es la misma cantidad de clientes que llega a un banco en día de quincena que en un día normal de la semana. Además, en un día de quincena

existen horas específicas de mayor afluencia. Con la tecnología actual se puede obtener fácilmente datos de desempeño pertinentes mediante el desarrollo de un programa de computadora para simular la operación del sistema.

Los modelos presentados aquí proporcionan una útil herramienta de cómo medir los tiempos de espera en todo establecimiento en el cual se originan filas. El primero de ellos nos muestra una forma de medir el tiempo promedio de espera a partir del número de operadores que se encuentran en el establecimiento, mientras que el segundo lo hace a partir de la velocidad de salida del cliente. En el primer caso los clientes llegan de forma discontinua al establecimiento, como en paquetes, en un flujo no continuo. En el segundo modelo los clientes llegan de forma continua, con cierta velocidad constante, como un flujo continuo. Si bien el gotero y la manguera son buenas analogías, en ambos casos la opción más práctica o viable es adecuar el flujo de clientes con el número de cajeros u operadores.

Óscar Córdoba Rodríguez
Marcelo Del Castillo Mussot
Instituto de Física,
Universidad Nacional Autónoma de México.
Marco A. de la Lama Zubirán,
Universidad Autónoma Metropolitana, Xochimilco.
Alfredo de la Lama García
Universidad Autónoma Metropolitana, Iztapalapa.

REFERENCIAS BIBLIOGRÁFICAS

Bose, S. 2002. *An Introduction to Queueing Systems*. Kluwer/Plenum Publishers.

Edwards, W. 1989. *Calidad, productividad y competitividad. La salida de la crisis*. Díaz de Santos México.

Everitt C. W. F. 1995. "La creatividad de Maxwell", en *Resortes de la creatividad científica*, Aris Rutherford *et al.* (comp.), FCE, México.

Gross, D. y C. Harris. 1998. *Fundamentals of Queueing Theory*. Ed. Wiley Series in Probability and Statistics.

Ishikawa, K. 1985. *¿Qué es el control total de calidad? La modalidad japonesa*. Norma, México.

Juran, J. M. 1990. *Juran y el liderazgo para la calidad. Un Manual para directivos*. Ed. Díaz de Santos, México.

Moskowitz, H. y G. P. Wright. 1991. *Investigación de Operaciones*. Prentice_Hall Hispanoamericana S.A.

Tijms, H. C. 2003. *First Course in Stochastic Models*. Chichester.

IMÁGENES

Pp. 52-53: Margaret Bourke, Louisville, K. Y., 1937. P. 54: Haciendo cola para la matinal de cine de Plaistow, Londres, 1937; Niños de Bradford haciendo cola para tomar un medicamento; 1939, Lancashire; Samuel Gottscho, **The Young Readers**: 1941. P. 55: C. Bresson, Francia, 1945-1946. P. 56: William C. Shrout N. Y., 1945; James Burke, Estambul, 1960; Agustín Centelles, **Cola de votantes durante las elecciones de febrero, de 1936**, España; Wallace Kirkland, México, 1954. Pp. 56-57: A Haunting Portrait, **ca.** 1900. Ian Smith, U. K., 1945; William Vandivert, Berlin, 1945; Alfred Eisenstaedt, 1943. P. 58: C. Bresson, **Distribución de oro en los últimos días del Kumintang**, Shangai, 1949. P. 59: Alfred Eisenstaedt, 1943.

OPTIMIZING QUEUE WAIT TIMES

Palabras clave: fluidos, optimización, sistemas granulares, ciencia y sociedad, sistemas de control.

Key words: Fluids, Optimization, Granular Systems, Science and Society, Control Systems.

Resumen: Se muestran dos formas de medir los tiempos de espera en locales de servicio en donde se generan filas o colas. El primer método mide los tiempos de espera a partir de la oferta, la demanda y el número de operaciones. El segundo método mide el tiempo de espera a partir de las velocidades de entrada y salida.

Abstract: The article shows two ways of measuring waiting times at service facilities where queues form. The first method measures waiting times based on supply, demand, and number of operations. The second method measures waiting time based on entry and departure rates.

Marco Alfredo de la Lama Zubirán es ingeniero mecánico con maestría en Ingeniería mecánica -área de fluidos. Trabaja de consultor y es profesor de matemáticas en la Universidad Autónoma Metropolitana Iztapalapa.

Alfredo de la Lama García es economista, con Doctorado en Sociología. Profesor investigador titular en la Universidad Autónoma Metropolitana Iztapalapa.

Oscar Córdoba Rodríguez es pasante de física en la Facultad de Ciencias de la UNAM, y estudiante asociado al Instituto de Física. También está cursando la carrera de matemáticas en la misma facultad.

Marcelo del Castillo-Mussot, Doctor en Física por la Universidad de California , campus San Diego, es investigador Titular C de Tiempo Completo del Instituto de Física de la UNAM. Trabaja en teoría de Fisica del Estado Sólido y en Sistemas Complejos.

Recibido el 2 de febrero de 2010, aceptado el 17 de marzo de 2010.