¿BAJO QUÉ ESCENARIOS ES CONVENIENTE REALIZAR CITACIONES EN UNA EMPRESA DE SERVICIOS?

Susana Mondschein*
Gabriel Weintraub**
Departamento de Ingeniería Industrial,
Universidad de Chile

Resumen

Servicios públicos y privados utilizan políticas de citaciones para atender a sus clientes de manera ordenada. Si los tiempos de atención y la llegada de los clientes al sistema son determinísticos, entonces realizar citaciones es óptimo debido a que ni los clientes ni los servidores sufren de tiempos de espera. Sin embargo, si los tiempos de atención tienen variabilidad y/o los clientes llegan tarde a sus citaciones, la política de citaciones puede perder su virtud ordenadora. En este artículo se determinan los niveles críticos de variabilidad de los tiempos de atención y de la llegada de los clientes, para los cuales un sistema en que los clientes llegan libremente sin citaciones funciona mejor que un sistema con citaciones. Se encuentra que, en general, servicios con tiempos de atención cortos les conviene funcionar con llegada libre, debido a la presencia de atrasos. Más aún, si la variabilidad de los tiempos de atención es alta, con mayor razón conviene utilizar esta modalidad. Por su parte, servicios con tiempos de atención largos, en general, les conviene funcionar con citaciones. Son servicios que regularmente no tienen una alta variabilidad en los tiempos de atención ni presentan grandes atrasos en relación a los tiempos de atención. Además con el modelo fue posible encontrar excepciones a esta regla. Por ejemplo, un servicio altamente congestionado podría funcionar mejor con llegada libre a pesar de que los atrasos de los clientes y la variabilidad de los tiempos de atención sean bajos. El marco conceptual utilizado para el análisis correspondió a una empresa de servicios privada que maximiza su utilidad esperada y enfrenta una demanda sensible al tiempo de espera.

* Departamento de Ingeniería Industrial, Universidad de Chile, y Yale School of Management. susana.mondschein@yale.edu. Los autores agradecen el apoyo financiero de Fondecyt (Chile), proyecto 1010457.

** Departamento de Ingeniería Industrial, Universidad de Chile. gweintra@dii.uchile.cl

1. Introducción

En la actualidad, una gran mayoría de las empresas de servicio enfrenta mercados altamente competitivos, con clientes cada vez más exigentes. Estas empresas compiten no sólo vía precios, sino que también a través de la calidad del servicio entregado. Uno de los aspectos de la calidad que ha cobrado gran relevancia en los últimos años es la rapidez con que se entrega el servicio; es decir, los clientes no quieren perder su tiempo esperando a ser atendidos. Taylor (1994) y Katz, Larson y Larson (1991), entre otros, reconocen el impacto negativo que tiene la espera en la satisfacción global de los consumidores con respecto al servicio. El primero realiza encuestas y el segundo entrevistas personales, encontrando ambos que si el tiempo de espera aumenta, la evaluación del servicio realizada por los clientes empeora.

En cuanto a la celeridad de prestación del servicio existen dos características que deben ser consideradas: (i) la percepción que tiene el cliente acerca del tiempo de espera y (ii) el tiempo de espera en sí mismo. Es evidente que un menor tiempo de espera incide directamente en una menor percepción de éste. Sin embargo, si el tiempo de espera no puede ser controlado, entonces se debe controlar la percepción que tienen los clientes acerca de él. En ese sentido, Larson (1987) sostiene que un ambiente externo agradable (música en la sala de espera, por ejemplo) y la sensación de que el sistema de atención es «socialmente justo» (First-In-First-Out) reducen la percepción del tiempo de espera, aumentando la satisfacción de los clientes con respecto al servicio recibido.

Un factor determinante en el tiempo de espera de los clientes es la política de atención impuesta por el servidor. En este contexto, entendemos la política de atención como la forma en que llegan los clientes al sistema. Distintos servicios utilizan diferentes políticas de atención de clientes. Por ejemplo, la mayoría de los médicos realiza citaciones. Si éstas se realizan de manera muy seguida, los clientes enfrentarán largas esperas y, por el contrario, si las citaciones se realizan separadas por intervalos de tiempo largos, el tiempo de espera de los clientes será bajo. Por su parte, es común que los bancos y los restaurantes de comida rápida funcionen sin citas previas en que los clientes llegan libremente al sistema. ¿Por qué ciertos tipos de servicios funcionan con citas y otros no lo hacen? ¿Cuál es la mejor forma de controlar el tiempo de espera de los clientes utilizando eficientemente los recursos del servidor? ¿Qué factores son determinantes en la decisión?

En este artículo se estudia bajo qué condiciones es conveniente realizar citaciones y bajo cuáles, es mejor dejar que los clientes lleguen libremente al sistema. Existen servicios para los cuales esta respuesta es obvia, por ejemplo, los servicios que se requieren de manera no planificada sólo pueden funcionar con llegada libre (servicios médicos de emergencia). Sin embargo, existe una amplia gama de servicios que pueden funcionar indistintamente con llegada libre de clientes o con citaciones, para los cuales la pregunta de qué política de atención se debe usar es relevante.

La modalidad más común de citación son las citas individuales equiespaciadas (se cita un cliente a la vez a intervalos de tiempo constante), modalidad utilizada, por ejemplo, por la mayoría de las consultas médicas privadas de Chile. Pero, ¿por qué se utiliza esta modalidad? Una respuesta rápida a esta pregunta es que de esta manera se impone una forma más ordenada en la llegada de los clientes, lo cual reduce el tiempo de espera de ellos y el tiempo ocioso del servidor. Por ejemplo, si tenemos un servicio en el cual los tiempos de atención son determinísticos iguales a 20 minutos, entonces resulta obvio citar precisamente cada 20 minutos. Si los clientes son absolutamente puntuales, se observará que tanto el tiempo de espera de los clientes como el tiempo ocioso del servidor serán nulos. Sin embargo, ¿qué sucede si los tiempos de atención y la llegada de los clientes dejan de ser determinísticos? O sea, ¿qué pasa si los tiempos de atención son variables aleatorias y los clientes se atrasan en la llegada a sus respectivas citas? En este caso, el sistema de citas puede no ser el mejor. Considerando que en la mayoría de las situaciones en la práctica, los tiempos de atención son aleatorios y los clientes se presentan con atrasos a sus citaciones, es interesante contestar la pregunta, sobre qué niveles de variabilidad de los tiempos de atención y de la llegada de los clientes, las citas individuales pierden su efecto «ordenador» y simplemente es mejor no realizar citaciones, dejar que los clientes lleguen libremente y atenderlos según una disciplina FIFO. En este trabajo, se intenta dar respuesta a esta pregunta para el caso de una empresa prestadora de servicios que maximiza su utilidad esperada.

El artículo se distribuye de la siguiente manera. En la Sección 2 se hace una revisión bibliográfica en la cual se critica el modelo de análisis utilizado hasta ahora en la literatura. A partir de esa crítica, en la Sección 3 se propone un nuevo modelo de análisis para una empresa privada prestadora de servicios. Luego en la Sección 4 se explica la forma de resolución del modelo y en la Sección 5 se describen los escenarios analizados. Finalmente en la Sección 6 se presentan los resultados obtenidos y en la Sección 7 se realizan las conclusiones y recomendaciones para investigaciones futuras.

2. Revisión Bibliográfica

En la literatura se ha estudiado ampliamente el tema del diseño de políticas de citación, estudiando el siguiente problema: *dado que se debe citar* a un número determinado de clientes, cuál es la manera óptima de hacerlo. Se han estudiado diferentes políticas de citación bajo distintos escenarios (distintas distribuciones del tiempo de atención, diversos tipos de servicios y comportamiento de clientes, etc.).

Por ejemplo, Jansson (1966) estudia el comportamiento en el largo plazo de un sistema simple, en que los clientes son citados individualmente a intervalos constantes, y en que los tiempos de atención son exponenciales. Encuentra expresiones para la esperanza del tiempo ocioso del servidor y del tiempo de espera de los clientes, con lo cual determina el intervalo óptimo entre citaciones. Se nota

que si la valoración del tiempo ocioso del servidor en la función objetivo es grande con respecto a la valoración del tiempo de espera de los clientes, las citaciones se realizarán separadas por intervalos de tiempo pequeños y viceversa.

Fries y Marathe (1981) y Liu y Liu (1998a, 1998b) entre otros abordan el problema de citar a N personas en K instantes de tiempo predeterminados. Es decir, se debe decidir el número de clientes a citar en los K instantes (n_1, n_2,..., n_K), de modo que $\sum_{i=1}^{K} n_i = N$. Por su parte en Wang (1993, 1997) se levanta el supuesto que los instantes de citación están predeterminados y se consideran políticas de citación más generales, en que los instantes de citación son variables de decisión. Es decir, se resuelve el siguiente problema: se debe citar a N clientes y se debe escoger el vector X = (x_1, x_2,..., x_N), en que x_i es el tiempo de citación del cliente *i*. Haciendo ciertos supuestos sobre la distribución de los tiempos de atención y el comportamiento de los clientes (no se ausentan y son puntuales), es posible deducir expresiones matemáticas que permiten determinar la política óptima de citaciones.

Ho y Lau (1992, 1999) y Yang, Lau y Quek (1998) no consideran un conjunto de políticas de citación factibles tan general como Wang (1993, 1997), sin embargo analizan un mayor número de escenarios utilizando simulación. éstos se refieren a considerar distintas distribuciones del tiempo de atención, distintos porcentajes de ausentismo y número de clientes por atender.

Todos los trabajos mencionados hasta ahora, suponen que los clientes son homogéneos en el sentido que sus tiempos de atención son variables aleatorias i.i.d.. Weiss (1990) y Klassen y Rohleder (2000) levantan este supuesto y suponen que existe más de un segmento de clientes, dentro de los cuales los tiempos de atención se distribuyen de la misma manera, pero las funciones de distribución entre los distintos segmentos son diferentes. Este problema es muy común en la programación de las intervenciones de una sala de operaciones, en que se sabe que distintos procedimientos tienen diferentes distribuciones de tiempo de duración.

Adicionalmente, existe un conjunto de trabajos, como por ejemplo Babes y Sarma (1991) y Bennett y Worthington (1998), que entregan recomendaciones más generales y cualitativas para el buen diseño de una política de citaciones. En su gran mayoría estos trabajos han surgido al realizar aplicaciones de investigación de operaciones en clínicas médicas para mejorar la calidad de servicio.

Alguna de las principales conclusiones obtenidas en la literatura existente son:

1. La mayoría de los servicios funciona durante una jornada de trabajo acotada, por lo cual es poco realista suponer que se alcanza estado estacionario. Dado eso, la herramienta más utilizada para resolver los problemas es simulación.
2. No existe la mejor regla de citación a todo evento, sino que ésta depende del escenario estudiado (variabilidad de los tiempos de atención, valoración del tiempo de espera de los clientes en relación a la valoración del tiempo ocioso del servidor, etc.).
3. Si la variabilidad de los tiempos de atención aumenta, el desempeño de un sistema que funciona con citaciones empeora.

Una completa revisión bibliográfica del tema de políticas de citaciones se puede encontrar en Mondschein y Weintraub (2000). A pesar de la extensa literatura en el tema, todos los trabajos suponen que la política de atención a utilizar es realizar citaciones. Es decir, el problema que a continuación se estudia, correspondiente a determinar las condiciones bajo las cuales es conveniente realizar citaciones en un servicio no ha sido abordado en la literatura.

Por otro lado, la función objetivo utilizada en los artículos anteriores corresponde a una combinación lineal de la esperanza del tiempo total de espera de los clientes y del tiempo ocioso esperado del servidor (o alguna variación similar a esto). Así se introduce el «trade-off» involucrado en este problema: si la valoración del tiempo de espera de los clientes en la función objetivo es grande con respecto a la valoración del tiempo ocioso del servidor, las citaciones se realizarán separadas por intervalos de tiempo grandes, de modo de disminuir el tiempo de espera de los clientes y viceversa. Además en los trabajos anteriores se supone que la cantidad de clientes a atender es fija e independiente de la política de citación impuesta. Es decir, se supone que la demanda es exógena y no depende del tiempo de espera de los clientes, el cual es afectado directamente por la política de citación utilizada. Mondschein y Weintraub (2001) discuten estos supuestos y muestran que la función objetivo utilizada en la literatura es adecuada sólo para un servicio público que enfrenta una demanda insensible al tiempo de espera (por ejemplo, un consultorio médico en que la gente está dispuesta a esperar mucho con tal de recibir el servicio). Sin embargo, para otros escenarios, como un servicio privado que enfrenta una demanda sensible al tiempo de espera, la función objetivo es inadecuada.

En primer lugar, en un modelo más realista se debe suponer que los clientes no desean esperar para ser atendidos. Es decir, mientras más se hace esperar a la gente, menor será la cantidad demandada por el servicio. Además se debe considerar la función objetivo adecuada considerando los costos y beneficios apropiados, dependiendo de si el servidor es público o privado.

En la sección siguiente se presenta un modelo que nos permitirá comparar el desempeño de una política de llegada libre de clientes con respecto al de una política de citas individuales equiespaciadas para una empresa privada prestadora de servicios. El modelo presentado es bastante más realista que el utilizado en la literatura, incorporando una demanda aleatoria que depende del tiempo de espera y utilizando la función objetivo adecuada para un servidor privado (maximizar su utilidad esperada). Por un lado, la empresa desea atender la mayor cantidad de clientes posible, sin extender su jornada de trabajo «excesivamente» (sin sufrir de mucho tiempo ocioso). Por otra parte, debe considerar que si implementa una política de atención que provoque altos tiempos de espera, la cantidad de demanda por el servicio podría reducirse considerablemente.

3. Modelo

A continuación se presenta el modelo general de análisis, introduciendo el problema de la empresa prestadora de servicios y el comportamiento de los clientes. Luego, el modelo se describe para las dos políticas de atención a comparar: llegada libre de clientes y citas individuales equiespaciadas.

3.1 Modelo General

3.1.1 El Problema de la Empresa Prestadora de Servicios

Consideraremos una empresa que ofrece un servicio a través de un único servidor. La disciplina de atención es FIFO y los tiempos de atención de los clientes son variables aleatorias i.i.d. según alguna función de densidad conocida. La empresa debe decidir el precio unitario a cobrar por el servicio y la política de atención a utilizar. El presente trabajo se focalizará en las decisiones de política de atención de la empresa y, por lo tanto, se supone que el precio es conocido y exógeno al modelo.

El objetivo de la empresa es maximizar su beneficio diario, que se compone de los ingresos percibidos por las prestaciones brindadas menos el costo asociado al funcionamiento. Este último corresponde al costo incurrido por unidad de tiempo trabajada, el cual, en general, es más alto en aquellas horas posteriores a la jornada normal de trabajo, o sea, en sobretiempo. Suponemos adicionalmente, que el resto de los costos de operación son constantes para todas las políticas de atención factibles y, por lo tanto, no se consideran en la optimización. En consecuencia, dada una política de atención s, la utilidad diaria de la empresa es:

$$\pi(s) = p \cdot Q(s) - \nu \cdot min(t_c(s), D) - \sigma \cdot t_o(s) \ , \qquad (1)$$

donde p es el precio a cobrar, *Q(s)* es la cantidad de clientes atendidos (que depende de la política de atención s impuesta), *D* es el largo de la jornada normal de trabajo (por ejemplo de 9:00 a 18:00), $t_c(s)$ es el largo de la jornada efectiva de trabajo del servidor (desde que comienza a trabajar hasta que atiende al último cliente) y $t_o(s)$ es el sobretiempo (es igual a $\max(t_c - D, 0)$). Las constantes v y σ ($v < \sigma$) corresponden al costo unitario por hora trabajada durante la jornada normal y después de ella, respectivamente (por ejemplo, son el salario por unidad de tiempo del servidor).

A continuación demostramos la siguiente proposición, que nos será de utilidad para la formulación del problema de la empresa. Denotamos *E(X)* como la esperanza de la variable aleatoria *X*. En la proposición y en su demostración se omite la dependencia de las variables en *s* para simplificar la notación.

Proposición 1

$E(\pi) = p \cdot E(Q) - \nu \cdot E(t_c) - (\sigma - \nu) \cdot E(t_o) = p \cdot E(Q) - \nu \cdot (E(t_c) - E(t_o)) - \sigma \cdot E(t_o)$

Demostración:

Tomando esperanza sobre la expresión (1), se obtiene que:

$$E(\pi) = p \cdot E(Q) - \nu \cdot E(\min(t_c, D)) - \sigma \cdot E(t_o) \qquad (2)$$

Se tiene que:

$$\begin{aligned} E(t_o) &= E(\max(t_c - D, 0)) \\ &= E(\max(t_c - D, 0) \mid t_c \geq D) \cdot \Pr[t_c \geq D] + \\ &\quad E(\max(t_c - D, 0) \mid t_c < D) \cdot \Pr[t_c < D] \\ &= \int_D^{\infty} (t - D) f_{t_c}(t) dt \ , \end{aligned} \tag{3}$$

en que $f_{tc}(t)$ es la función de densidad de t_c.

Por otro lado, tenemos que:

$$\begin{aligned} E(\min(t_c, D)) &= E(\min(t_c, D) \mid t_c \leq D) \cdot \Pr[t_c \leq D] + \\ &\quad E(\min(t_c, D) \mid t_c > D) \cdot \Pr[t_c > D] \\ &= \int_0^{D} t f_{t_c}(t) dt + D \int_D^{\infty} f_{t_c}(t) dt \end{aligned}$$

Restando y sumando $D \int_D^{\infty} f_{t_c}(t) dt$, agrupando términos y considerando la expresión (3), se obtiene.

$$\begin{aligned} E(\min(t_c, D)) &= \int_0^{\infty} t f_{t_c}(t) dt - \int_D^{\infty} (t - D) f_{t_c}(t) dt \\ &= E(t_c) - E(t_o) \end{aligned} \tag{4}$$

Reemplazando (4) en (2), finalmente se obtiene:

$$\begin{aligned} E(\pi) &= p \cdot E(Q) - \nu \cdot (E(t_c) - E(t_o)) - \sigma \cdot E(t_o) \\ &= p \cdot E(Q) - \nu \cdot E(t_c) - (\sigma - \nu) \cdot E(t_o) \end{aligned}$$

La utilidad diaria del servidor p es una variable aleatoria y la función objetivo del servidor es maximizar la utilidad diaria esperada. En consecuencia, utilizando la proposición 1, el problema que debe resolver el servidor queda definido como:

$$\max_{s \in S} \quad E(\pi(s)) = p \cdot E(Q(s)) - \nu \cdot [E(t_c(s)) - E(t_o(s))] - \sigma \cdot E(t_o(s)) \ , \tag{5}$$

donde *S* es el conjunto de políticas de atención factibles a estudiar, que incluyen las citas individuales equiespaciadas y la llegada libre de clientes. En el caso de citas individuales, se debe decidir el número máximo de citaciones a realizar durante la sesión (*C*). Se debe notar que *C* es el cupo disponible para citaciones y es posible que no se llenen todos si es que la cantidad de demanda por el servicio es menor a *C*. Para llegada libre se decide el instante hasta el cual se deja ingresar clientes al sistema (*H*). Bajo esta modalidad ingresan clientes libremente al sistema en el intervalo [*0, H*], en que H es una decisión que toma el servidor y es menor al largo de la jornada normal de trabajo (*D*).

Se debe notar que en un modelo más general, las funciones de costos que asocia el servidor a las horas trabajadas podrían ser no lineales, punto que creemos toma mayor relevancia en el caso asociado al sobretiempo. Por ejemplo, sería

razonable pensar que al médico le molesta quedarse media hora después de las 7:00 en su consulta, pero le molesta aún más quedarse la misma media hora después de las 7:30, con lo cual la función de costos asociada al sobretiempo sería no lineal y convexa.

3.1.2 La Demanda por el Servicio

Por el lado de la demanda, existe una población de individuos que genera una demanda potencial por el servicio, la que se describe por un Proceso de Poisson de tasa λ_0 [clientes/sesión]. En un modelo más general, la tasa media de demanda potencial podría ir variando a lo largo del día, considerando así la posibilidad de un proceso no-homogéneo. De esta manera, se incorporarían al modelo servicios en los cuales existen «horas-peak» de demanda.

Se debe notar que este proceso describe la demanda diaria potencial por el servicio, es decir, es independiente de las decisiones de la empresa y en particular, independiente de la política de atención implementada (es la misma para llegada libre y citaciones individuales). Además, suponemos que los clientes sólo deciden acudir al servicio en un día particular, sin importarles a qué hora hacerlo. Por ejemplo, supongamos un escenario en que el servicio funciona con la disciplina de llegada libre de clientes y supongamos que el intervalo de tiempo en el cual se deja ingresar gente al sistema es de ocho horas. Ahora, tenemos el mismo servicio, pero sólo se deja ingresar personas durante seis horas. El supuesto anterior indica que en ambos escenarios la demanda potencial será Poisson de tasa λ_0 [clientes/sesión]. En el caso en que se deja ingresar personas por un intervalo de tiempo menor, la demanda se *reacomodará* en este intervalo más pequeño. En consecuencia, la demanda potencial por unidad de tiempo en este último caso será mayor, para así llegar a la misma demanda potencial por sesión. Una justificación a este supuesto es que los servicios considerados en este trabajo, cumplen con la propiedad que los clientes planifican de antemano su requerimiento (en efecto, pueden funcionar con citaciones), por lo cual se espera que ellos no sean tan sensibles frente a la hora en la cual deban recibirlo. Por ejemplo, si una persona debe ir al dentista, lo planifica y podría acomodar su horario a una cita en particular. No ocurre lo mismo en el caso de una fuente de soda, en que la mayoría de los consumidores, simplemente ingresa a ella en un acto espontáneo, sin demasiada planificación. En este caso, si la fuente de soda está cerrada, la demanda potencial por ese servicio se pierde y no se reacomoda en otro horario.

Cada cliente posee una utilidad asociada a recibir el servicio dada por el valor que tiene el servicio para él, o equivalentemente, su máxima disposición a pagar por recibir el servicio si el tiempo de espera fuera nulo (precio de reserva), menos el precio del servicio y el costo asociado al tiempo de espera. Denotamos la utilidad del cliente i como u_i la que se escribe como:

$$u_i = r_i - p - aw,$$

donde r_i es el precio de reserva del cliente i, p es el precio del servicio, w es el

tiempo de espera del cliente y a es el valor unitario del tiempo de espera para los clientes.

También suponemos que los clientes son neutros al riesgo y por lo tanto, al tomar sus decisiones evalúan el valor esperado de su utilidad. Es decir, un cliente demanda efectivamente el servicio si:

$$E(u_i) = r_i - p - a\bar{W} > 0 \ , \tag{6}$$

en que $\overline{W}$ es la esperanza del tiempo de espera ($E(w)$). Es decir, un cliente potencial hace efectiva su demanda por el servicio si la utilidad esperada asociada a hacerlo es positiva. Cada cliente conoce su precio de reserva, sin embargo desde el punto de vista del servidor son variables aleatorias. Sin pérdida de generalidad, suponemos que los clientes provienen de una población homogénea y, por lo tanto, el precio de reserva se describe por una única densidad de probabilidad $f_R(r)$ conocida.

La forma usada para la utilidad de los consumidores es análoga a la utilizada en algunos trabajos que estudian colas desde un punto de vista microeconómico. Ellos consideran el hecho de que cuando un consumidor ingresa al sistema genera una externalidad negativa (mayor tiempo de espera), que afecta a los consumidores que ingresan después de él. Lo que se busca en estos trabajos es encontrar mecanismos de precios que a partir de las decisiones individuales de los consumidores, se induzcan comportamientos socialmente óptimos. En esta línea podemos mencionar a Mendelson y Whang (1990) y Hassin (1995). La función de utilidad utilizada supone clientes neutros al riesgo. Un modelo más realista podría considerar clientes aversos al riesgo, en que no sólo importa la esperanza del tiempo de espera sino que también su varianza. Por ejemplo, un cliente podría preferir un servicio en el cual siempre esperara 10 minutos con seguridad, por sobre otro, en el cual tuviera que esperar 20 minutos la mitad de las veces, y nada, la otra mitad. En efecto, Larson (1987) realizando entrevistas concluye que, en general, el costo de espera para las personas es una función no lineal del tiempo de espera (los primeros minutos no parecen tan largos, si por ejemplo, se lee una revista en la sala de espera). Además, Leclerc, Schmitt y Dubé (1995) muestran a través de estudios empíricos que la tendencia de los individuos es a ser aversos al riesgo para decisiones que involucran el uso de su tiempo. Una explicación posible es que se desea evitar el riesgo en el dominio del tiempo para así poder planificar mejor.

Para facilitar la notación definimos $\hat{r}_i = r_i - p$. Por lo tanto, la utilidad esperada de los consumidores, dada una política de atención s, será:

$$E(u_i(s)) = \hat{r}_i - a\bar{W}(s) \ .$$

Notemos que el tiempo medio de espera $\overline{W}(s)$ depende de la política de atención (s) impuesta. Recordando que la demanda potencial es Poisson de tasa λ_0 y que los clientes demandan el servicio si $E(u_i(s))$ es positivo, entonces dado un valor de $\overline{W}(s)$, la demanda efectiva observada por el servidor será Poisson de tasa:

$$\lambda(\bar{W}(s)) = \lambda_0 \cdot \Pr[\hat{R} \geq a\bar{W}(s)] \ \left[\frac{clientes}{sesión}\right] \ , \tag{7}$$

en que $\widehat{R}$ es la variable aleatoria que describe realizaciones i.i.d. de $\hat{r}_i$. El resultado anterior se obtiene en virtud de la propiedad de división de procesos Poissonianos.

Una vez que un cliente decide demandar el servicio, se supone que ya no puede tomar más decisiones. Por lo tanto, una vez que el cliente arriba al sistema para recibir el servicio no se permite que lo abandone antes de hacerlo.

3.1.3 El Equilibrio de Expectativas Racionales

Los clientes no conocen con certeza cuánto tendrán que esperar para recibir el servicio, pero tienen alguna información de su distribución de probabilidad con la cual hacen una estimación del tiempo promedio de espera $\overline{W}$ (s), para así tomar sus decisiones de demanda. Por ahora supondremos que esta expectativa de tiempo de espera es única y constante a lo largo del día. Sin embargo, en un modelo más realista, podría incluirse el hecho de que a distintas horas del día, se pueden tener distintos tiempos promedio de espera.

Suponemos que la estimación que los clientes realizan acerca de $\overline{W}$ (s) es la misma para todos los clientes y es la correcta. Es decir, suponemos que los clientes poseen *expectativas racionales* (Muth (1961)). Por lo tanto, dada una política de atención s definida por el servidor, los clientes estiman un valor $\overline{W}$ (s) con el cual toman sus decisiones de demanda. Esta demanda, Poisson de tasa $\lambda(\overline{W}(s))$, genera un tiempo de espera que justamente coincide con la estimación inicial hecha por los clientes.

El suponer que los clientes poseen expectativas racionales para estimar la esperanza del tiempo de espera, significa que ellos son capaces de aprender de sus errores y en el largo plazo realizar una estimación correcta. Supongamos que la expectativa que tienen los clientes con respecto a $\overline{W}$ (s) es, en promedio, menor a lo que realmente se observa. En ese caso, existirán clientes que regularmente demandarán el servicio y luego esperarán más de lo que pensaban, obteniendo en promedio, utilidades negativas. Por el contrario, si la expectativa que tienen los clientes con respecto a $\overline{W}$ (s) es mayor a lo que realmente es, ocurrirá que clientes que deberían demandar el servicio (tienen una utilidad esperada positiva si lo hicieran) no lo hacen. Los dos casos anteriores presentan una inconsistencia entre la expectativa de los clientes y el verdadero valor del tiempo promedio de espera. Suponemos que los clientes aprenden de esta inconsistencia en el largo plazo y por lo tanto, la expectativa que se forman de $\overline{W}$ (s) es justamente el valor que toma la variable. Creemos que este supuesto es bastante realista, ya que en general, uno es capaz de estimar (por experiencia propia o de otros que le han contado) el tiempo promedio de espera en diferentes servicios.

3.2 Modelo de Llegada Libre

Bajo esta modalidad, los clientes llegan al sistema en el momento en que ellos lo

estimen conveniente, con la única restricción de que en el instante de su arribo el sistema debe estar abierto. Los clientes no necesitan una cita de manera anticipada.

Por su parte, el servidor debe decidir el instante hasta el cual dejar entrar clientes al sistema. En este punto, se debe notar que no necesariamente le conviene dejar entrar gente hasta el final de su jornada normal de trabajo D, sino que eventualmente, podría serle más conveniente cerrar la entrada en algún momento anterior, el cual denotamos como H. Este caso se puede deber a que si la entrada al sistema se mantiene abierta durante un intervalo de tiempo muy grande, entonces es probable que el servicio deba trabajar durante una jornada mayor para atender a todos los clientes. Si el costo asociado al largo de la jornada de trabajo y más aún, el costo de sobretiempo son muy grandes, entonces desde algún instante de tiempo, el beneficio obtenido por cada cliente nuevo recibido puede ser menor al costo incurrido. De esta manera si se escoge un valor pequeño de H, la demanda potencial debe reacomodarse en un intervalo menor de tiempo, generando mayor congestión en el sistema (mayores tiempo de espera), pero disminuyendo la jornada de trabajo del servidor.

Como se mencionó en la sección 3.1.2, en cuanto a que la demanda se reacomoda totalmente al largo de la sesión, la demanda potencial se describe por un Proceso de Poisson de tasa λ_0 [clientes/sesión], independiente del largo del intervalo de tiempo en el cual se deja ingresar clientes al sistema (H). Por lo tanto, si se deja ingresar personas al sistema durante un intervalo de largo H, la demanda potencial media por unidad de tiempo será $\frac{\lambda_0}{H}$[clientes/ unidad de tiempo]. En consecuencia, la demanda efectiva que observa el servidor, es decir, el proceso de llegada al sistema es Poisson de tasa λ ($\overline{W}$) [clientes/sesión] (ecuación (7)).

Por lo tanto, recordando la expresión (5), el problema del servidor en llegada

$$\begin{aligned} \max_{H} \quad & E(\pi(H)) = p \cdot \lambda(\bar{W}(H)) - \nu \cdot [E(t_c(H)) - E(t_o(H))] - \sigma \cdot E(t_o(H)) \\ \text{s.a.} \quad & 0 \leq H \leq D \end{aligned} \tag{8}$$

3.3 Modelo de Citaciones Individuales a Intervalos Constantes

Esta política corresponde al caso en que los clientes se citan individualmente separados por intervalos de tiempo constantes. Por lo tanto, la empresa prestadora del servicio debe decidir el número máximo de citaciones a realizar durante la sesión (C), de modo de maximizar su utilidad esperada diaria. Sólo se citan clientes durante la jornada normal de trabajo y por lo tanto durante el intervalo [0,D].

Una regla de citación muy utilizada en la realidad (por ejemplo, muchos médicos la utilizan) es citar individualmente a intervalos iguales al tiempo medio de atención. Esta regla es factible de implementar, si el número de citaciones es lo suficientemente pequeña como para que todas se puedan realizar durante el intervalo [0,D]. Por ejemplo, si el tiempo medio de atención es τ entonces esta regla es factible de implementar si $C\tau \leq D$. En caso contrario, si el número de citaciones es mayor, la forma usual de realizarlas es equiespaciadamente y ordenando los

cupos para citaciones de manera uniforme a lo largo del día, es decir, citando a intervalos iguales a D/C . Si el número de citaciones crece, el intervalo entre ellas se reduce, aumentando los tiempos de espera de los clientes y reduciendo la jornada laboral del servidor.

Considerando lo anterior, la regla de citaciones que utilizaremos se define de la siguiente manera:

- Si el número C de clientes a citar es tal que $C\tau \leq D$, entonces se evalúa la posibilidad de citar a tiempo medio de atención. También se evalúa la alternativa de citar equiespaciadamente a intervalos D/C , que da la posibilidad de citar a intervalos más largos y se escoge la mejor.
- Si $C > D/t$, entonces simplemente se cita equiespaciado a intervalos regulares iguales a D/C .
- A medida que las solicitudes de los clientes van llegando, los cupos disponibles se van llenando desde comienzos de la sesión hacia el final.
- La disciplina de atención es FIFO.
- Dada la existencia de atrasos, es posible que lleguen clientes después de la jornada normal de trabajo D. Si un cliente llega al sistema después del intervalo $[0,D]$ y el servidor está ocupado, entonces una vez que el servidor se desocupe el cliente entrante será atendido. En caso contrario, si al momento de la llegada del cliente, el servidor está desocupado, el cliente no recibe la atención. Este supuesto se asemeja al comportamiento observado en la realidad en cuanto a que, en general, si el servidor termina de atender al último cliente después de su jornada normal de trabajo y en ese momento no hay nadie más esperando, éste se retira del sistema aunque existan clientes que deberían llegar más tarde.

La regla así definida simplifica la resolución del problema. Notemos que la regla de citación individual equiespaciada más general es del tipo (C,β), en que β es el intervalo entre citaciones y también es una variable de decisión. Nosotros hemos omitido la decisión β en la formulación para reducir la complejidad del problema: para un valor de C dado, la regla que definimos nos indica el intervalo entre citaciones. Si β también fuera una variable de decisión, la optimización sería sobre dos variables, aumentando el tamaño del problema. Sin embargo, la regla que utilizaremos es bastante usada en la realidad y por lo tanto, el modelo así definido nos entregará resultados que de todas formas serán muy útiles para entregar recomendaciones de gestión a distintos tipos de servicios.

Frente a una política de citación (C), la demanda efectiva que observará la empresa servidora es un proceso de Poisson de tasa $\lambda(\overline{W}(C))$ al cual denotamos $G(C)$. En este contexto, la demanda posee una connotación distinta que en llegada libre, donde correspondía al proceso de llegada de personas al sistema. En este caso, la demanda efectiva es el proceso de conteo correspondiente a las solicitudes por el servicio para el día analizado. Podríamos decir que es el proceso estocástico que corresponde a todas las llamadas telefónicas solicitando una cita para la sesión en estudio.

Notemos que C es una decisión que toma el servidor y corresponde al cupo dispuesto para citaciones durante la sesión. Esto no significa que necesariamente existirá demanda como para llenar todos estos cupos. Eso dependerá del valor que tome la demanda efectiva $G(C)$. Por lo tanto, el número de clientes que atiende el servidor corresponde a $Q(C) = \min(G(C),C)$. Además, se supone que no hay ningún costo asociado a la demanda insatisfecha (caso en que la cantidad demandada por el servicio es mayor que C).

Utilizando la expresión (1) podemos escribir la utilidad diaria del servidor:

$$\begin{aligned}\pi(C) &= p \cdot Q(C) - \nu \cdot \min(t_c(C), D) - \sigma \cdot t_o(C) \\ &= p \cdot \min(G(C),\, C) - \nu \cdot \min(t_c(C), D) - \sigma \cdot t_o(C)\end{aligned}$$

Tomando esperanza y utilizando la proposición (1), el problema de optimización que resuelve el servidor en el caso de citaciones individuales a intervalos constante es:

$$\begin{aligned}\max_C \quad E(\pi(C)) \quad &= \quad p \cdot E(Q(C)) - \nu \cdot [E(t_c(C)) - E(t_o(C))] - \sigma \cdot E(t_o(C)) \\ \text{s.a.} \quad &\quad C = 0, 1, 2, ...\end{aligned} \tag{9}$$

3.4 La Comparación

Dado un escenario determinado (parámetros del modelo fijos, cierta distribución de los tiempos de atención y cierta distribución de los tiempos de atrasos en citaciones), interesa comparar la máxima utilidad esperada que se puede obtener con llegada libre y con una política de citaciones individuales equiespaciadas, para determinar cuál de las dos es más conveniente. Es decir se compara:

$$\max_H \; E(\pi(H)) \qquad \text{con} \qquad \max_C \; E(\pi(C)) \;,$$

en que el primer término es la máxima utilidad esperada diaria que se puede obtener con llegada libre (expresión (8)) y, el segundo, con citas individuales (expresión (9)). La comparación puede corregirse por una constante, si es que los costos fijos de operación son mayores para citas individuales (secretaria, teléfono, etc.), como es de esperar sobretodo en servicios en los cuales se atiende a muchos clientes diariamente.

4. Resolución del Modelo

4.1 Llegada Libre

Recordando la expresión (8), el problema del servidor en llegada libre es:

$$\begin{aligned}\max_H \quad E(\pi(H)) &= p \cdot \lambda(\bar{W}(H)) - \nu \cdot [E(t_c(H)) - E(t_o(H))] - \sigma \cdot E(t_o(H)) \\ \text{s.a.} \quad &\quad 0 \le H \le D \;,\end{aligned}$$

en que $\lambda(\overline{W}(H)) = \lambda_0 \Pr[\hat{R} \ge a\, \overline{W}(H)]$.

A continuación, se describe el algoritmo de resolución para el caso de llegada libre, en el cual se debe maximizar $E(\pi(H))$ sobre los posibles valores de H. Para hacerlo, se utiliza el algoritmo de optimización no lineal Golden Section Method, el cual converge al óptimo si la función objetivo es estrictamente unimodal (Bazaraa, Sherali y Shetty (1993)), propiedad que se verifica experimentalmente para $E(\pi(H))$ en la gran mayoría de los escenarios analizados. En algunos escenarios la función objetivo no es estrictamente unimodal sino que estrictamente creciente. Sin embargo, en esos casos el algoritmo llega a una solución muy cercana al óptimo (a menos del 0,5 %), con lo cual se puede garantizar que el algoritmo de optimización utilizado entrega la solución óptima (o una solución a menos del 0,5 % en algunos casos) en los escenarios estudiados.

El algoritmo de resolución es el siguiente:

1. **Paso 1**: Se escoge un valor de H según como el algoritmo de optimización lo indique (en la primera iteración se deben escoger dos). Para un valor de H dado, se encuentra $\overline{W}^*(H)$ correspondiente al equilibrio de expectativas racionales, que corresponde a la solución de la siguiente ecuación de punto fijo:

$$\bar{W}(H) = T(\lambda(\bar{W}(H)), H) \ ,$$

donde $T(\lambda(\overline{W}(H)),H)$ = esperanza del tiempo de espera dado $\lambda(\overline{W}(H)$,y dado H. La expresión $\lambda(\overline{W}(H))$ está dada por la ecuación (7). Para encontrar el punto fijo, se resuelve numéricamente la ecuación (10) utilizando el método de Van Wijngaarden, Dekker y Brent (Press, Flannery, Teukolsky y Vetterling (1990)). Para ello se debe evaluar la función $T(\lambda(\overline{W},H))$ en distintos valores de $\overline{W}$, lo cual se realiza utilizando simulación (Law y Kelton (2000)). Es decir, dado H, para un valor de $\overline{W}$, se calcula $\lambda(\overline{W})$, con el cual se genera un proceso de Poisson de llegada de clientes al servicio. Dado ese proceso de llegada y los tiempos de atención correspondientes, se calcula el tiempo promedio de espera utilizando simulación. La precisión relativa utilizada en estas simulaciones es del 0,5 % a un nivel de confianza del 99 %.

2. Paso 2: Una vez que se determina $\overline{W}^*(H)$ de equilibrio, se determina la tasa de llegada al sistema $\lambda(\overline{W}^*(H))$ de manera directa, con la cual se calcula $E(t_c(H))$ y $E(t_o(H))$ mediante simulación. De esta manera, se calcula la utilidad esperada diaria del servidor, $E(\pi(H))$, para el valor de H correspondiente.

3. Paso 3: Se verifica si se cumple la condición de optimalidad dada por el algoritmo de optimización. Si es así, se para y se obtiene el óptimo H^* y $E(\pi(H^*))$. Si no, se vuelve al paso 1.

4.2 Citaciones Individuales

El problema del servidor es decidir el número C de clientes a citar en el día. Por lo tanto, de la expresión (9), el problema del servidor es:

$$\max_{C} \quad E(\pi(C)) = p \cdot E(Q(C)) - \nu \cdot [E(t_c(C)) - E(t_o(C))] - \sigma \cdot E(t_o(C))$$

$$\text{s.a.} \qquad C = 0, 1, 2, \ldots \ ,$$

en que $Q(C)$ es el mínimo entre la demanda efectiva que observa el servidor ($G(C)$) y el número de cupos disponibles para citaciones (C).

La forma de resolución para citas individuales es análoga al caso de llegada libre salvo por algunas particularidades. En primer lugar, la optimización se realiza sobre un conjunto discreto (los posibles valores de C). En este caso también se aprovecha la unimodalidad de $E(\pi(C))$, propiedad observada en todos los experimentos computacionales realizados, y se utiliza un algoritmo en el cual se va descartando el espacio factible de a mitades.

Además, otra diferencia con la resolución de llegada libre es que la llegada de los clientes es diferente. En este caso, existen citaciones más atrasos. También, se debe considerar que no necesariamente se llenan los C cupos predeterminados para las citas (si la demanda es menor) y que la cantidad máxima de clientes atendidos en una sesión es precisamente C.

5. Escenarios Estudiados

Se consideran distintos escenarios con el objetivo de cubrir diferentes tipos de servicios. En primer lugar, se utilizan diferentes valores para la jornada normal de trabajo del servidor, D. Supondremos que el tiempo medio de atención es siempre igual a 1 [ud. de tiempo] y por lo tanto, un valor mayor de D implica que durante la sesión caben más atenciones en promedio. En consecuencia, distintos valores de D, pueden caracterizar diferentes tipos de servicios. Por ejemplo, el caso D igual a 10 [uds. de tiempo] puede representar a un médico cuyo tiempo medio de atención es 30 mins. y el largo de la sesión es 5 hrs. o en que el tiempo medio de atención es 1 hora y el largo de la sesión es 10 hrs. En cambio, el caso D igual a 100 [uds. de tiempo] podría representar un banco, cuyo tiempo medio de atención es 3 mins. y el largo de la sesión es 5 hrs. En general, escenarios con un valor pequeño de D representan servicios en que caben pocas atenciones diarias en promedio (tiempo promedio de atención es alto) y viceversa. En este trabajo, se considera el caso en que el largo de la jornada normal de trabajo (D) es 10, 50 y 100 [uds. de tiempo].

Además los parámetros que caracterizan al servicio se relacionan con la demanda y el servidor. Con respecto a la demanda, existen servicios que presentan desde una demanda potencial alta y una alta valoración de los clientes (estarán altamente congestionados), hasta servicios con baja demanda potencial y baja valoración de los clientes (tendrán baja congestión). Desde el punto de vista del servidor, debemos caracterizar servicios con diferentes estructuras de costos. Por ejemplo, el costo de funcionamiento (podría ser su costo de oportunidad) de un médico prestigioso es bastante alto en relación al costo de funcionamiento de una peluquería común y corriente.

Para cada tipo de servicio, interesa investigar el efecto del coeficiente de variación de los tiempos de atención y de los atrasos de los clientes con respecto a

sus citaciones, sobre el desempeño del sistema. Es decir, interesa determinar sobre que niveles de variabilidad de los tiempos de atención y de los atrasos con respecto a las citaciones se hace más conveniente dejar llegar libremente a los clientes al sistema.

A continuación, se definen los parámetros del modelo, las distribuciones del tiempo de atención y las distribuciones de los atrasos a utilizar.

5.1 Parámetros del Modelo

Se consideran distintos parámetros del modelo, en lo referente a la demanda y al servicio.

A) Demanda

Recordemos que existe una demanda potencial que se describe por un proceso de Poisson de tasa λ_0 [clientes/ sesión]. Consideramos tres valores para la media de la demanda potencial:

- Baja: $\lambda_0 = D/2$
- Media: $\lambda_0 = D$
- Alta: $\lambda_0 = 2 \cdot D$

La demanda efectiva que observa el servidor es Poisson de tasa λ_0 Pr[$\hat{R} \geq a \overline{W}$]. Suponemos que $\hat{R}$ se distribuye Weibull(α,β). Por lo tanto,

Dado que mover α y a es equivalente, se fija $a = 0,5$ y se varía α. Además, el valor de β se fija igual a 2, con lo cual queda determinado el coeficiente de variación de R igual a 0,5. Los valores de α se varían de modo de obtener diferentes medias de $\hat{R}$ (a medida que α crece, $E(\hat{R})$ decrece). Se consideran escenarios en que los clientes tiene una valoración promedio baja por el servicio y escenarios en que la valoración es alta.

B) Servidor

Recordemos que la utilidad esperada del servidor es:

$$E(\pi) = \pi\, E(Q) - v\,(E(t_c) - E(t_c)) - \sigma \cdot E(t_0)\ ,$$

donde v y σ son los costos por unidad de tiempo trabajada durante la jornada normal de trabajo y en sobretiempo, respectivamente.

Sin pérdida de generalidad suponemos que el precio es igual a 1. Se consideran diferentes estructuras de costo:

- Costos bajos: v = 0; 2 y σ = 0; 5
- Costos altos: v = 0; 5 y σ = 1; 2

5.2 Atraso de Clientes Con Respecto a Sus Citaciones

Se supone que los clientes se atrasan con respecto a sus citaciones y que los atrasos son variables aleatorias i.i.d. exponencialmente distribuidas. Para estudiar su efecto, se compara utilizando los mismos parámetros del modelo en ambos casos, el escenario con citaciones con diferentes niveles de atrasos con respecto al escenario con llegada libre, obviamente sin atrasos (no existen).

Se consideran diferentes tiempos medios de atraso, que dependen del valor de D. Por ejemplo, para D = 10 se consideran atrasos medios que van desde 0 (clientes puntuales) hasta 2 veces el tiempo medio de atención (atrasos medios del orden de una hora). Para el caso D = 100, se consideran atrasos medios que pueden llegar hasta 7 veces el tiempo medio de atención, que podrían corresponder a atrasos medios de 15 a 20 minutos.

5.3 Coeficiente de Variación de los Tiempos de Atención

Se supone que los tiempos de atención son variables aleatorias i.i.d. Weibull de media igual a 1[ud. de tiempo]. Se quiere estudiar el efecto del coeficiente de variación de los tiempos de atención, para lo cual se consideran los siguientes valores para el coeficiente de variación: 0,5, 1, 2 y 3.

Un coeficiente de variación de 0,5 es relativamente bajo y podría representar el caso de un médico, en que él tiene cierto control sobre el largo de las atenciones, disminuyendo la variabilidad de éstas (O'Keefe (1985)). Un coeficiente de variación de 2 es muy alto y puede representar el caso de un servicio con dos o más segmentos de clientes con diferentes distribuciones de los tiempos de atención de medias bastante distintas. Por ejemplo, supongamos un banco que atiende a dos tipos de clientes: clientes comunes, que son el 80%, cuyo tiempo medio de atención es 1,5 minutos; clientes especiales (administrativos de oficinas, por ejemplo), cuyo tiempo medio de atención es 9 minutos. En este caso, la esperanza del tiempo de atención de un cliente cualquiera es 3 minutos y el coeficiente de variación es cercano a 2.

Para cada set de parámetros definidos en 5.1, se mueven los coeficientes de variación de los tiempos de atención y las medias de los tiempos de atrasos de los clientes en citas individuales por separado, para estudiar los efectos aisladamente. Es decir, sin considerar atrasos se determina sobre que nivel del coeficiente de variación de los tiempos de atención, llegada libre se hace mejor a citas individuales. Luego, dada una distribución de tiempos de atención fija (Weibull de media igual a 1 y coeficiente de variación igual a 0,5), se estudia el efecto de los atrasos en citaciones para determinar sobre que nivel es preferible dejar llegar libremente los clientes al sistema.

6. Resultados Obtenidos

6.1 Atraso de Clientes Con Respecto a Sus Citaciones

Cuando los tiempos medios de atrasos son pequeños en relación al tiempo medio de atención, la utilidad esperada que entrega la utilización de citas individuales es, en gran parte de los casos, mayor a la de llegada libre. Es decir, citas individuales produce un efecto ordenador que permite atender a los clientes de mejor manera, entregando mayores utilidades para el servidor. Sin embargo, si la magnitud de los atrasos crece, este efecto ordenador se pierde y las diferencias de utilidades recibidas con llegada libre y citas individuales se reducen. Más aún, sobre cierto nivel de atrasos la utilidad esperada recibida con llegada libre de clientes es mayor a la obtenida con citas individuales. Como se observa en la figura 1, para el caso D = 10 se requieren tiempos medios de atrasos del orden de 0,5 a 2 veces el tiempo medio de atención para que llegada libre entregue una utilidad esperada mayor que citas individuales. Para los casos D = 50 y D = 100 los atrasos requeridos son del orden de 2 a 5 y 3 a 7 veces el tiempo medio de atención respectivamente.

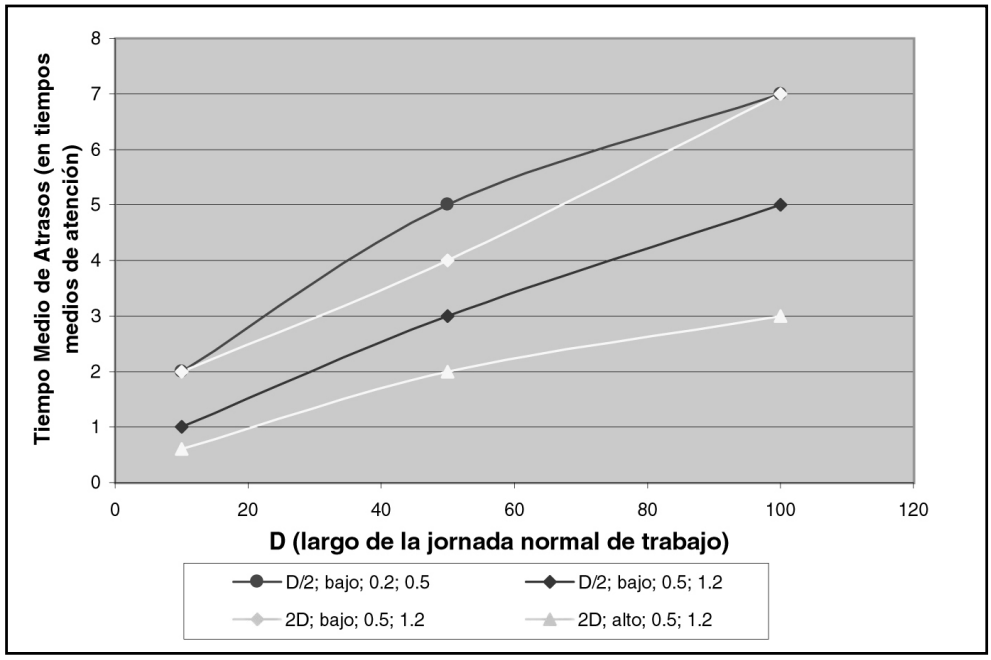

Figura 1: Tiempos medios de atrasos requeridos para que la utilidad esperada entregada por llegada libre sea mayor a la obtenida con citas individuales, para diferentes valores de D. Cada curva representa un set de parámetros determinado, correspondiente a λ_0 (demanda potencial media); E($\hat{R}$) (valoración media que tienen los clientes por el servicio); v, σ (costos del servidor).

A pesar de que cuando D crece, se requieren tiempo medios de atrasos mayores en relación al tiempo medio de atención para que convenga utilizar llegada

libre, es justamente en estos últimos escenarios en que es más probable la conveniencia de no realizar citaciones. Para el caso en que en una sesión caben 10 atenciones en promedio, el tiempo medio de atención es del orden de 30 minutos. En consecuencia, los atrasos medios requeridos para que la utilización de llegada libre entregue una utilidad esperada mayor que citas individuales, que son del orden de una a dos veces el tiempo medio de atención, pueden ser muy altos y poco realistas. Sin embargo, en servicios en los cuales caben más atenciones en promedio durante una sesión, el tiempo medio de atención corresponde a un intervalo menor (2 a 5 minutos, por ejemplo), con lo cual es perfectamente posible que existan atrasos suficientemente grandes como para no utilizar citaciones. En efecto, es realista suponer que existan atrasos medios del orden de 5 veces el tiempo medio de atención (10 minutos, por ejemplo). De esta manera, podríamos explicar que muchos servicios con tiempos de atención relativamente altos, como los médicos, funcionan con citaciones. Por el contrario, servicios que tienen tiempos de atención cortos (bancos, por ejemplo), funcionan con llegada libre de clientes. En estos casos, pequeños atrasos, los cuales se observan frecuentemente, tienen un impacto negativo importante en el desempeño de una política de citaciones.

Además, los resultados dependen del resto de los parámetros del problema, los cuales dependiendo de sus valores representan distintos tipos de servicio. En nuestro modelo escenarios en que la valoración media que tienen los clientes por el servicio es baja son servicios que presentan un bajo nivel de congestión. En general, estos son servicios que operan en un mercado bastante competitivo, son suntuarios y la gente no está dispuesta a esperar demasiado por recibirlos. En ellos se requieren atrasos bastante mayores para que convenga utilizar llegada libre (escenario [D=2; bajo; 0.2; 0.5] en figura 1).

Si la valoración que tienen los clientes por el servicio crece se requieren tiempos medios de atrasos menores para que llegada libre sea preferible a citar individualmente (comparar escenarios [2D; bajo; 0.5; 1.2] y [2D; alto; 0.5; 1.2] en la figura 1). Lo mismo ocurre si los costos del servidor aumentan (comparar escenarios [D/2; bajo; 0.2; 0.5] y [D/2; bajo; 0.5; 1.2] en la figura 1). Adicionalmente en los escenarios en que la valoración de los clientes por el servicio, los costos del servidor y la demanda potencial son altos se requieren los menores atrasos para que llegada libre entregue una utilidad esperada mayor a citas individuales. Básicamente estos son escenarios en que el servicio se observa muy congestionado, debido a la alta valoración que tienen los clientes por el servicio y a los altos costos de operación del servidor, lo que lleva a que las citaciones se realicen de manera muy seguida. En general, estos son servicios muy especiales que tienen su nicho de mercado, son de primera necesidad (salud pública, por ejemplo) o la gente, por diferentes motivos, está dispuesta a esperar mucho por recibirlos (es gente joven con un costo de oportunidad del tiempo menor, por ejemplo). En ellos se requieren niveles de atrasos bastante bajos para que sea conveniente dejar de realizar citaciones. En efecto, llegada libre puede entregar una utilidad esperada mayor incluso para tiempos medios de atrasos muy menores (escenario [2D; alto; 0.5; 1.2] en Figura 1).

En la figura 2, se resumen los resultados hasta aquí expuestos. En el eje y se coloca el tiempo medio de atención del servicio, distinguiendo en alto y bajo. En

el eje x se coloca el nivel de congestión del servicio, entendiéndolo de la manera anteriormente mencionada y nuevamente distinguiendo en alto y bajo. En cada cuadro se entrega la política de atención que es conveniente utilizar.

Tiempo Medio de Atención			
	ALTO	Citas Individuales (atrasos observados no son suficientemente grandes como para dejar de utilizarlas). Ej.: Médicos en general.	En general citaciones, salvo que atrasos sean suficientemente grandes y el servicio presente altísimos niveles de congestión, en cuyo caso podría convenir llegada libre.
	BAJO	En general llegada libre, salvo que atrasos en citas individuales sean muy pequeños.	Llegada libre Ej.: Banco
		BAJO	**ALTO**
		Nivel de Congestión del Servicio	

Figura 2: Política de atención recomendada a utilizar para servicios con distintos tiempos medios de atención y niveles de congestión.

Los resultados obtenidos concuerdan con lo observado en la realidad, con respecto a que la mayoría de los servicios suntuarios pero elegantes funcionan con citaciones (restaurantes elegantes, peluquerías finas, etc.). En este tipo de servicios la gente no está dispuesta a esperar mucho por ser atendido. En nuestro modelo esto significa que la valoración que tienen los clientes por el servicio es baja y como vimos, en estos escenarios se requieren atrasos demasiado grandes y que no se observan en la realidad como para que sea conveniente dejar de utilizar citaciones. Por el contrario, en servicios más populares la gente está dispuesta a esperar más por ser atendidos (mayor nivel de congestión). Efectivamente, muchos de ellos funcionan con llegada libre.

Al momento de decidir la política de atención se debe tener en cuenta la diferencia en los costos administrativos entre ambas políticas si es que la hubiere. En particular, es esperable que los costos administrativos asociados a una política de citaciones sean mayores a los de llegada libre y, por lo tanto, si citas entrega una utilidad sólo un poco mayor que llegada libre (por ejemplo, una diferencia porcentual de 5 %), es probable que la política a implementar sea esta última. Estos costos son mayores en la medida que se realiza un mayor número de citaciones (se requiere más secretarias, líneas telefónicas, etc.), es decir para valores mayores de D. Es justamente en estos servicios donde se observa con mayor frecuencia el uso de llegada libre.

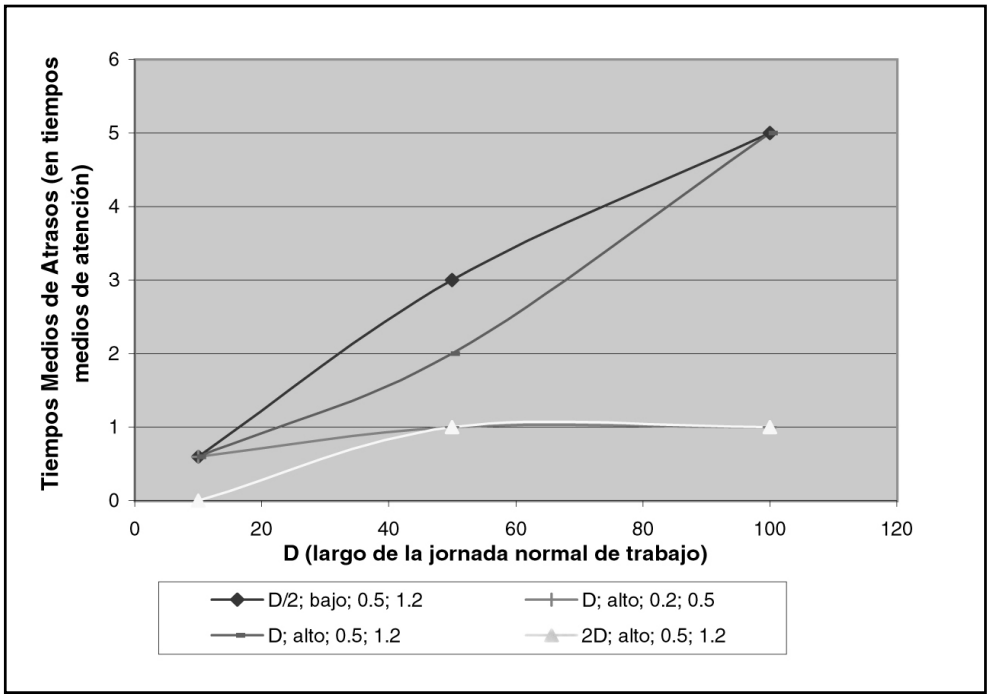

Figura 3: Tiempos medios de atrasos requeridos para que la diferencia porcentual entre la utilidad esperada entregada por llegada libre con respecto a la obtenida con citas individuales sea menor al 5 %, para diferentes valores de D. Cada curva representa un set de parámetros determinado, correspondiente a λ_0 ; $E(\hat{R})$;v;σ.

En efecto, como se observa en la figura 3, los tiempos medios de atrasos requeridos para que la diferencia porcentual entre la utilidad esperada que entrega citas individuales y llegada libre sea menor al 5 % son del orden de 0 a 1 veces el tiempo medio de atención para el caso D = 10; 1 a 3 veces el tiempo medio de atención para el caso D = 50 y 1 a 5 veces el tiempo medio de atención para el caso D = 100. Estos niveles de atrasos pueden ser perfectamente realistas en diferentes servicios, sobretodo para valores grandes de D.

6.2 Coeficiente de Variación de los Tiempos de Atención

Con respecto al efecto de la variabilidad de los tiempos de atención, se observa que si ésta es baja (coeficiente de variación entre 0,5 y 1) la utilidad que entrega el uso de citas individuales es, en la mayoría de los casos, bastante superior a la de llegada libre (diferencias porcentuales mayores al 10%). Si el coeficiente de variación aumenta, tanto el desempeño de citaciones como el de llegada libre empeora. Sin embargo, la utilidad obtenida con citaciones disminuye de manera más pronunciada, reduciendo así su diferencia con respecto a llegada libre. A pesar de que en estos casos, la utilidad obtenida con citas individuales es siempre mayor a la obtenida con llegada libre, en general, se observa que para valores del coeficiente de variación de los tiempos de atención del orden de 2, las diferencias

porcentuales entre ellas son bastante pequeñas (5 a 10 %). En la figura 4 se ejemplifica lo anterior para el escenario D = 50. En él se observa el efecto del aumento de la variabilidad de los tiempos de atención con respecto a la diferencia porcentual entre la utilidad esperada que entrega citas individuales y llegada libre.

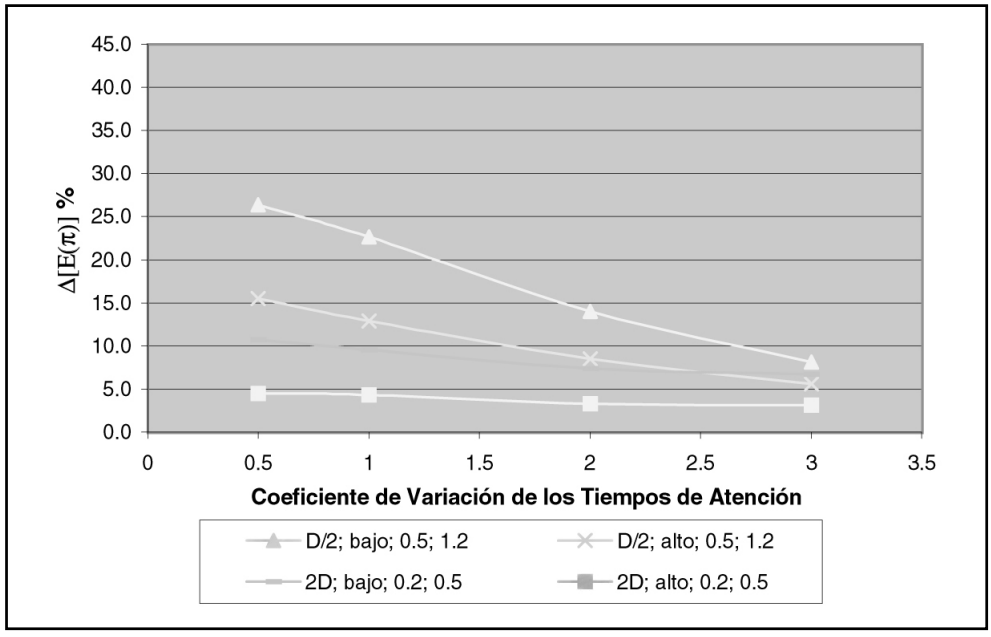

Figura 4: Efecto del coeficiente de variación de los tiempos de atención sobre las diferencias porcentuales de la utilidad esperada entregada por llegada libre con respecto a la obtenida con citas individuales para D = 50. $\Delta[E(\pi)]$% corresponde a 100 $\frac{E(\pi)CI - E(\pi)LL}{E(\pi)CI}$ en que LL indica llegada libre y CI indica citas individuales. Cada curva representa un set de parámetros determinado, correspondiente a $\lambda_0; E(\hat{R}); v; \sigma$.

Como se dijo en la Sección 5 algunos servicios que presentan alta variabilidad en los tiempos de atención son los bancos, la oficina de timbraje del Servicio de Impuestos Interno y oficinas administrativas en general. Estos son servicios que, en promedio, atienden muchos clientes diariamente (valor alto de D). Además la gran mayoría de los clientes tienen atenciones cortas (clientes «comunes»), sin embargo, una minoría tiene atenciones bastante largas (administrativos de oficinas con muchos trámites a realizar) que hacen aumentar la variabilidad de los tiempos de atención. Justamente, se observa que muchos de estos servicios funcionan con llegada libre de clientes. En efecto, si se considera que funcionar con citas individuales conlleva un costo operacional que para estos casos (muchas atenciones diarias) puede ser bastante alto (varias líneas telefónica, secretaria), es probable que lo conveniente sea funcionar con llegada libre.

Si la valoración que tienen los clientes por el servicio es alta las diferencias de desempeño entre llegada libre y citas individuales son aún menores (comparar escenarios [D/2; bajo; 0.5; 1.2] y [D/2; alto; 0.5; 1.2] en la figura 4). Adicionalmente si la demanda potencial es alta, lo que resultaría en un escenario con un alto nivel

de congestión (escenario [2D; alto; 0.2; 0.5] de la figura 4), incluso para coeficientes de variación muy bajos (0,5), las diferencias de desempeño entre llegada libre y citas individuales son muy pequeñas (menores al 7 %). En general, en escenarios con un nivel de congestión mayor se requiere de variabilidades menores de los tiempos de atención para que las diferencias de utilidades percibidas con citaciones y llegada libre sean pequeñas. Los resultados anteriormente descritos se resumen en la figura 5.

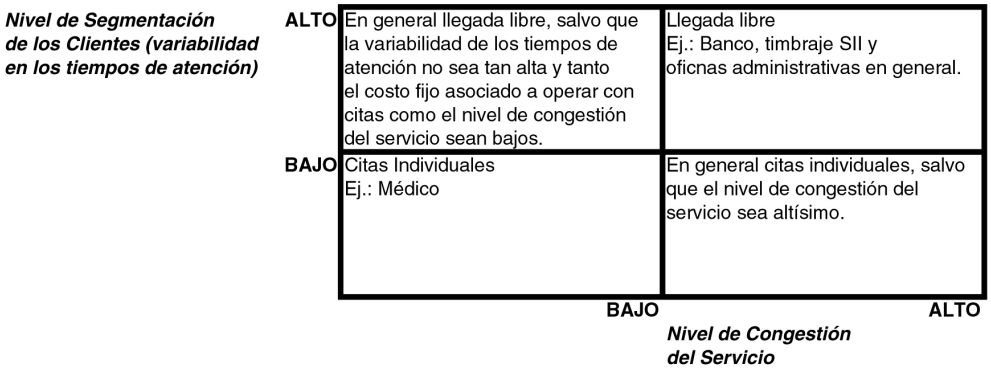

Figura 5: Política de atención recomendada a utilizar para servicios con distintos niveles de segmentación de los clientes en relación a los tiempos de atención (si el nivel de segmentación es mayor, la variabilidad de los tiempos de atención crece) y diferentes niveles de congestión.

7. Conclusiones y Recomendaciones para Investigación Futura

Con respecto a los atrasos de los clientes en citaciones, se observa que mientras más grandes sean, menor es la utilidad obtenida al utilizar esta política. De esta manera, existe un nivel de atrasos sobre el cual es conveniente no realizar citaciones y dejar llegar libremente los clientes al sistema.

En general, servicios cuyos tiempos de medios de atención son relativamente largos, no presentan atrasos suficientemente grandes en relación a ellos como para dejar de realizar citaciones. Por el contrario, si los tiempos de atención son pequeños es factible esperar niveles de atrasos suficientes como para que convenga dejar a los clientes llegar libremente al sistema, sobretodo si se considera el costo fijo operacional asociado a realizar citaciones.

Los resultados anteriores son bastante intuitivos y es la norma en el funcionamiento de muchos servicios: si las atenciones son relativamente cortas funcionan con llegada libre y si son largas con citaciones. Sin embargo, con nuestro modelo además pudimos encontrar escenarios en que esta regla puede fallar. Es el caso de un servicio con tiempos de atención altos pero sumamente congestionado (muy apetecido por los clientes y con altos costos de operación para el servidor). En este caso, basta que existan atrasos promedio muy pequeños (5 minutos,

por ejemplo), para que la utilidad entregada con llegada libre sea similar a la obtenida con citas individuales. Probablemente en un escenario como el recién descrito lo conveniente sería funcionar con llegada libre.

Asimismo pueden haber servicios con tiempos de atención bastante cortos en que a-priori parecería poco conveniente realizar citaciones, pero en que podría llegar a serlo. Estos serían servicios con niveles de congestión, atrasos y costos fijos relacionados a operar con citaciones muy bajos (por ejemplo, de todas formas hay secretarias disponibles).

Con respecto al efecto de la variabilidad de los tiempos de atención, se observa que si ésta es baja la utilidad que entrega el uso de citas individuales es, en la mayoría de los casos, bastante superior a la de llegada libre. Si el coeficiente de variación de los tiempos de atención aumenta, las diferencias de desempeño se reducen considerablemente.

Servicios cuyos tiempos de atención poseen una alta variabilidad son servicios que atienden a diferentes tipos de clientes. A ellos podría convenirles funcionar con llegada libre de clientes, sobretodo si se considera el costo administrativo que conlleva funcionar con citaciones, que en estos casos es alto debido a que generalmente estos servicios atienden a muchas personas diariamente. Este fenómeno también es bastante intuitivo y se observa a menudo en distintos servicios.

Sin embargo, con el modelo encontramos escenarios en que la simple regla: si la variabilidad de los tiempos de atención es muy alta use llegada libre y si no realice citaciones, puede no ser la correcta. Es posible que un servicio que atiende a clientes con un alto nivel de segmentación (alta variabilidad en los tiempos de atención), de todas formas le convenga realizar citaciones. Es el caso de un servicio con un bajo nivel de congestión y en que el costo fijo asociado a operar con citaciones es muy bajo. Por su parte, un servicio muy congestionado, podría convenirle funcionar con llegada libre aunque tenga poca variabilidad en sus tiempos de atención, sobretodo si los costos fijos asociados a funcionar con citas son altos.

En este trabajo se analizaron por separado el efecto de los atrasos de los clientes y de la variabilidad de los tiempos de atención. Sin embargo, en muchos casos reales ambos pueden estar presentes a la vez. Por ejemplo, un servicio puede tener tiempos de atención con coeficiente de variación igual a 1,5 y los clientes atrasarse 2 veces el tiempo medio de atención en promedio. En estos casos, la suma de estos dos efectos se potencia empeorando el desempeño de un sistema de citaciones.

En resumen, generalmente servicios con tiempos de atención cortos les conviene funcionar con llegada libre, debido a la presencia de atrasos. Más aún, si la variabilidad de los tiempos de atención es alta, con mayor razón conviene utilizar llegada libre. Por su parte, servicios con tiempos de atención largos, en general, les conviene funcionar con citaciones. Son servicios que regularmente no tienen una alta variabilidad en los tiempos de atención ni presentan grandes atrasos en relación a los tiempos de atención. Se debe notar que con el modelo fue posible encontrar excepciones a esta regla. Por ejemplo, un servicio altamente congestionado

podría funcionar mejor con llegada libre a pesar de que los atrasos de los clientes y la variabilidad de los tiempos de atención sean bajos.

Además de considerar el costo fijo asociado a operar con un sistema de citaciones, existe un factor más cualitativo, difícil de introducir en el modelo, pero relevante en desmedro de las citaciones. Los clientes le exigen más a un sistema de citaciones que a un sistema de llegada libre, es decir, la insatisfacción producida por esperar es mayor en un sistema que funciona con citaciones, que supuestamente debe funcionar ordenadamente, en relación a uno que no lo hace. Alguien que llega puntual a su cita y debe esperar debido a los atrasos del resto de los clientes y más aún observa un sistema desordenado, saldrá muy descontento con el servicio, probablemente mucho más descontento que alguien que espera una cantidad de tiempo similar en un sistema que funciona con llegada libre. Obviamente este efecto tendrá un impacto negativo en la satisfacción del cliente con el servicio, mermando las utilidades de la empresa. En la mayoría de los escenarios estudiados en que citas entrega una utilidad sólo un poco mayor que llegada libre, la diferencia en los tiempos de espera promedio generados por cada política también son muy pequeñas. Por lo tanto, considerando el costo fijo asociado a realizar citaciones y este factor cualitativo, en estos casos convendría utilizar llegada libre de clientes.

Adicionalmente pueden existir otros aspectos cualitativos que se deben tener en cuenta. Por ejemplo, existen bancos cuya publicidad es estar a cualquier hora disponible en todas partes. En este caso, implementar una política de citaciones puede ser contraproducente con el mensaje del banco.

Como corolario se puede decir que los servicios que presentan una alta variabilidad en los tiempos de atención, podrían alcanzar importantes mejoras si es que a priori pudieran distinguir y separar a sus clientes en dos segmentos: los de bajo y los de alto tiempo de atención. Al segmento con bajos tiempos de atención (por ejemplo, 2 a 3 minutos en promedio) sería conveniente atenderlos con llegada libre, ya que como vimos, la magnitud de los atrasos sería suficiente como para no realizar citaciones. Por el contrario, al grupo con atenciones largas (por ejemplo, 10 minutos en promedio), sería conveniente atenderlos con citaciones. Además, al atender por separado a los diferentes tipos de clientes, se disminuye la variabilidad de los tiempos de atención para ambos grupos, mejorando el desempeño tanto de llegada libre para el segmento con bajos tiempos de atención, como el de citas individuales para el segmento con altos tiempos de atención.

En varias simulaciones de ejemplos simples realizadas se observó que al atender por separado a los distintos tipos de clientes de la manera anteriormente descrita se pueden reducir de manera considerable los tiempos de espera. En general, el tiempo promedio de espera se puede reducir más o menos a la mitad, sin alargar la jornada de trabajo. Las reducciones son mayores mientras mayor es la diferencia entre el tiempo medio de atención del segmento de clientes con atenciones cortas con respecto al tiempo medio de atención del segmento de clientes con atenciones largas.

Por ejemplo, supongamos un servicio que atiende a 100 personas diariamente, en que el 80 % tiene tiempos de atención Weibull de media un minuto y

medio y coeficiente de variación 0,5. Por su parte, el resto tiene tiempos de atención Weibull de media 9 minutos y coeficiente de variación igual a 1. Si se atiende a todos los clientes juntos sin distinguir su tipo, se tiene que los tiempos de atención son variables aleatorias i.i.d. de media 3 minutos y coeficiente de variación igual a 1,7. En este caso y utilizando llegada libre de clientes, el tiempo promedio de espera de los clientes es 25 minutos aproximadamente. Si se modifica el funcionamiento del sistema, atendiendo por separado a los clientes con atenciones cortas (con llegada libre) y a los con atenciones largas (con citas individuales), el tiempo medio de espera se reduce a 9 minutos aproximadamente, sin alargar la jornada de trabajo.

Un ejemplo de esta recomendación es la modificación que se realizó hace algún tiempo en las oficinas de timbraje del Servicio de Impuestos Internos. En principio, se atendía a todos los clientes con una política de llegada libre. Bajo esta modalidad, se observaba que existían clientes con altísimos tiempos de atención, con muchos documentos por timbrar, que retrasaban significativamente la atención de todos los clientes detrás de él. Para remediar tal situación, se determinaron los siguientes cambios: en la mañana se atiende con llegada libre a los «pequeños timbradores» y en la tarde, se atiende con citaciones a los «grandes timbradores» (empresas, contadores, etc.). Nuestros resultados confirman que lo realizado por el SII va en la dirección correcta y además sirven de evidencia como para que más servicios con segmentos de clientes distinguibles, adopten medidas similares.

Finalmente, es posible que existan reglas de citación individuales equiespaciadas que, para escenarios determinados, sean mejores a la utilizada en este trabajo. En consecuencia, si con la regla aquí definida se encuentra que usar llegada libre entrega una utilidad mayor que al realizar citaciones, es posible que exista una regla de citación mejor, que revierta esta desigualdad. Incluso se podrían estudiar otras reglas de citación que pueden entregar mejores desempeños en determinadas circunstancias. Por ejemplo, en un servicio en que los tiempos de atención son cortos podría ser conveniente realizar citaciones en bloques más que individuales. Estos aspectos se estudiarán en el futuro.

En el futuro también se desea investigar otros factores que pueden ser determinantes al momento de decidir si es conveniente implementar una política de citaciones. Creemos que estos factores están siempre relacionados con la variabilidad de la llegada de los clientes y de las atenciones. Con respecto al primer punto, algunos aspectos interesantes que podrían ser estudiados, es el ausentismo de los clientes con respecto a las citaciones y la existencia de «walk-ins».

Adicionalmente se podrían extender los resultados obtenidos para un modelo más general. En esta dirección, se podrían considerar clientes aversos al riesgo y servicios con horas «peak» de demanda. También se podría abordar el problema más general, en que la empresa debe decidir conjuntamente su política de atención y el precio a cobrar por el servicio.

Referencias

1 Babes, M. and G. Sarma (1991), «Out-patient Queues at the Ibn-Rochd Health Care,» Journal of the Operational Research Society, Vol. 42, No. 10, 845-855.

2 Bazaraa, M. S., H. D. Sherali y C. M. Shetty (1993), «Nonlinear Programming, Theory and Algorithms,» John Wiley & Sons Inc.

3 Bennett, J. and D. J. Worthington (1998), «An Example of a Good but Partially Succesful OR Engagement: Improving Outpatient Clinic Operations,» Interfaces, Vol. 28, No. 5 , 56-69.

4 Fries, B. E. and V. P. Marathe (1981), «Determination of Optimal Variable-Sized Multiple- Block Appointment Systems,» Operations Research, Vol. 29, No. 2, 324-345.

5 Hassin, R. (1995), «Decentralized Regulation of a Queue,» Management Science, Vol. 41, No.1,163-173.

6 Ho, C. and H. Lau (1992), «Minimizing Total Cost in Scheduling Outpatient Appointments,» Management Science, Vol. 38, No. 12, 1750-1764.

7 Ho, C. and H. Lau (1999), «Evaluating the Impact of Operating Conditions on the Performance of Appointment Scheduling Rules in Service Systems,» European Journal of Operational Research, Vol. 112, 542-553.

8 Jansson, B. (1966), «Choosing a Good Appointment System - A Study of Queues of the Type (D, M, 1),» Operations Research, Vol. 14, 292-312.

9 Katz, K., B. Larson and R. Larson (1991), «Prescription for the Waiting-in-Line Blues: Entertain, Enlighten and Engage,» Sloan Management Review, Winter, 44-53.

10 Klassen, K. and T. Rohleder (1996), «Scheduling Outpatient Apointments in a Dynamic Environment,» Journal of Operations Management, Vol. 14, 83-101.

11 Larson, R. C. (1987), «Perspectives on Queues: Social Justice and the Phsychology of Queueing,» Operations Research, Vol. 35, No. 6, 895-904.

12 Law, A. y D. Kelton (2000), «Simulation Modeling and Analysis,» McGraw-Hill.

13 Leclerc, F., B. Schmitt and L. Dubé (1995), «Waiting Time and Decision Making: Is Time Like Money?,» Journal of Consumer Research, Vol. 22, No. 1, 110-119.

14 Liu, L. and X. Liu (1998a), «Dynamic and Static Job Allocation for Multi-Server Systems,» IIE Transactions, Vol. 30, 845-854.

15 Liu, L. and X. Liu (1998b), «Block Appoinment Systems for Outpatient Clinics with Multiple Doctors,» Journal of the Operational Research Society, Vol. 49, No. 12, 1254-1259.

16 Mendelson, H. and S. Whang (1990), «Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue,» Operations Research, Vol. 38, No. 5, 870-883.

17 Mondschein, S. y G. Weintraub (2000), «Appointment Policies in Service Operations: A Critical Survey,» MIMEO.

18 Mondschein, S. y G. Weintraub (2001), «Appointment Policies in Service Operations: A Critical Analysis of the Economic Framework,» aceptado bajo modificaciones en Production and Operations Management.

19 Muth, J. (1961), «Rational Expectations and the Theory of Price Movements,» Econometrica, Vol. 29, No. 3, 315-335.

20 O'Keefe, R. M. (1985), «Investigating Outpatient Departments: Implementable Policies and Qualitative Approaches,» Journal of the Operational Research Society, Vol. 36, No. 8, 705-712.

21 Press, W., B. Flannery, S. Teukolsky and W. Vetterling (1990), «Numerical Recipes in C, The Art of Scientific Computing,» Cambridge University Press.

22 Rohleder, T. and K. Klassen (2000), «Using Client-Variance Information to Improve Dynamic Appointment Scheduling Performance,» Omega, Vol. 28, 293-302.

23 Taylor, S. (1994), «Waiting for Service: the Relationship Between Delays and the Evaluation of Service,» Journal of Marketing, Vol. 58, 56-69.

24 Wang, P. (1993), «Static and Dynamic Scheduling of Costumer Arrivals to a Single-Server System,» Naval Research Logistics, Vol. 40, No. 3, 345-360.

25 Wang, P. (1997), «Optimally Scheduling N Client Arrival Times for a Single-Server System,» Computers and Operations Research, Vol. 24, No. 8, 703-716.

26 Weiss, E.N. (1990), «Models for Determining Estimated Start Times and Case Ordering in Hospital Operating-Rooms,» IIE Transactions, Vol. 22, No. 2, 143-150.

27 Yang, K.K., M. L. Lau and S. A. Quek (1998), «A New Appointment Rule for a Single-Server, Multiple-Client Service System,» Naval Research Logistics, Vol. 45, No. 3, 313-326.