IX Congreso de Ingeniería de Organización
Gijón, 8 y 9 de septiembre de 2005

Cálculo de la capacidad necesaria para obtener un nivel de servicio predeterminado*

Albert Corominas Subias[1], Amaia Lusa García[1], Norberto Muñoz Gómez[2]

[1] Instituto de Organización y Control. Dpto. de Admón. de Empresas. ETSEIB. Universitat Politècnica de Catalunya. Avda. Diagonal 647, p11, 08028 Barcelona. albert.corominas@upc.edu, amaia.lusa@upc.edu
[2] Instituto de Organización y Control. Universitat Politècnica de Catalunya. Avda. Diagonal 647, p11, 08028 Barcelona. norberto.munoz@upc.edu

Resumen

En los modelos de planificación o de programación de servicios, la capacidad ideal requerida en cada período del horizonte de programación constituye un dato esencial. Cuando no existen instrumentos para acoplar la demanda y la capacidad productiva, el valor medio de esta última en cada período tiene que ser superior a la demanda media prevista para el mismo, pero dicho valor no es observable o previsible directamente, sino que depende de la demanda prevista y del nivel de servicio que se desea alcanzar. Los procedimientos para calcularlo han sido objeto de una atención relativamente escasa en la literatura. Por otra parte, el absentismo es un fenómeno presente en todos los sistemas productivos, pese a lo cual no se suele mencionar en los estudios sobre la capacidad productiva. En este trabajo se presenta una síntesis de las propuestas publicadas hasta el momento al respecto y se formaliza un procedimiento de cálculo que supera algunas limitaciones de las mismas.

Palabras clave: planificación, programación, RRHH, colas

1. Introducción

Un dato esencial en los modelos de planificación o de programación de servicios es la capacidad ideal requerida en cada período comprendido en el horizonte de programación (Corominas et al., 2004). Cuando no existen instrumentos para acoplar la demanda y la capacidad productiva, el valor medio de esta última en cada período tiene que ser superior a la demanda media prevista para el mismo, pero dicho valor no es observable o previsible directamente, sino que depende de la demanda prevista y del nivel de servicio que se desea alcanzar. Los procedimientos para calcularlo han sido objeto de una atención relativamente escasa en la literatura.

Por otra parte, el absentismo, en un sentido amplio, es un fenómeno presente en todos los sistemas productivos, pese a lo cual no se suele mencionar en los estudios sobre la capacidad productiva.

En este trabajo se presenta una síntesis de las propuestas publicadas hasta el momento al respecto (sección 2) y se formaliza un procedimiento de cálculo que supera algunas limitaciones de las mismas (sección 3).

* Trabajo financiado por el proyecto DPI2004-05797

2. Antecedentes

En la programación del tiempo de trabajo en los servicios la capacidad productiva ideal que se requiere en cada período es uno de los datos indispensables. Para calcularlo se necesita una previsión de la demanda y una definición del nivel de servicio que se desea alcanzar (por ejemplo, que la proporción de unidades que han de esperar más de un tiempo prefijado no sea superior a un valor dado). Dicho cálculo no es trivial porque, en general, el tiempo entre demandas y el tiempo de servicio son aleatorios y la forma en que depende el nivel de servicio de estas magnitudes es compleja.

Por ello, se han propuesto procedimientos aproximados; por ejemplo, establecer como capacidad de cada período la que iguala la demanda del período multiplicada por un coeficiente mayor que la unidad.

O bien, considerar el horizonte de programación dividido en períodos breves y aplicar la teoría de colas para calcular la capacidad necesaria, en cada período, para alcanzar el nivel de servicio deseado (Thompson, 1990; Thompson, 1995; Ingolfsson et al, 2002). Este procedimiento ha sido denominado SIPP (Stationary Independent Period-by-Period) y se puede describir, con mayor detalle, como sigue:

- Establecer el nivel de servicio deseado.
- Dividir en T períodos (generalmente, de la misma duración) el horizonte de programación.
- Estimar la tasa de llegadas para cada período.
- Aplicar un modelo de teoría de colas a cada período para calcular la capacidad mínima que sería necesaria en el mismo, en régimen permanente, para alcanzar el nivel de servicio deseado, considerando cada período independientemente de los demás.

Evidentemente, algunos de estos puntos son comunes a otros procedimientos. Para definir el nivel de servicio existen muy diversas posibilidades, tanto en lo que se refiere a la naturaleza de los parámetros que lo caracterizan como en cuanto a los valores mínimos y máximos admisibles para los mismos. Se puede imponer, por ejemplo, que el número medio de unidades en el sistema no debe ser superior a un valor dado; en el caso de servicios a personas, no obstante, las definiciones del nivel de servicio que se basan exclusivamente en valores medios no suelen ser satisfactorias porque enmascaran el impacto de las situaciones extremas, que son las que dan lugar a una valoración negativa del servicio por las personas usuarias. Otras definiciones del nivel de servicio no tienen este inconveniente; por ejemplo, la probabilidad que la longitud de la cola sea superior a q no debe ser superior a p, o bien, la probabilidad que el tiempo de espera sea superior a τ no debe ser superior a p. Esta última definición del nivel de servicio es la adoptada en este trabajo.

Pero lo característico de SIPP es el supuesto de independencia entre períodos y el uso de la teoría de colas para calcular la capacidad correspondiente a cada período. Esto último presenta inconvenientes importantes. En primer lugar, la teoría de colas sólo puede aplicarse en supuestos muy restrictivos sobre las distribuciones de los tiempos (prácticamente, para un número de canales superior a uno, el único modelo robusto corresponde al supuesto de distribución exponencial de los tiempos entre llegadas y de los tiempos de servicio, lo cual, en general, está muy lejos de corresponder a la realidad). En segundo lugar, el modelo de teoría de colas proporciona resultados para el régimen permanente, pero para que el comportamiento del sistema se aproxime al correspondiente al régimen permanente debe transcurrir un tiempo

más o menos largo en el que se mantengan los valores de los parámetros; por consiguiente, cuando la tasa de llegadas no es constante esta aproximación puede ser muy inadecuada, ya que el período debe ser de breve duración (para que sea aceptable el supuesto de que la tasa de llegadas se mantiene sensiblemente constante en todo el período) por lo cual no se puede suponer que el comportamiento del sistema se aproxime al régimen permanente; es decir, el comportamiento del sistema en una período depende fuertemente de las condiciones iniciales, que resultan del estado del sistema al inicio del período anterior y de lo ocurrido en el mismo. Por consiguiente, también resulta ser poco aceptable el supuesto de independencia de los períodos.

Green et al. (2001), a partir de resultados que se presentan en Eick et al. (1993), introducen una mejora en SIPP, que consiste esencialmente en efectuar los cálculos de capacidad con una curva de tasa de demanda desplazada en el tiempo. La mejora se debe al hecho de que, cuando la demanda es sinusoidal, el instante de máxima congestión del sistema está retrasado en relación con el instante en que la tasa de llegadas es máxima. El alcance de la propuesta de Green et al. (2001) es limitado, ya que se basa en supuestos muy específicos y los resultados no son extrapolables.

Una vez calculada la capacidad deseada se ha de tener en cuenta, aunque sea un fenómeno poco mencionado en la literatura sobre la organización del tiempo de trabajo, el absentismo (Olivella, 2000). Se considera aquí, en sentido amplio, como la ausencia de una persona de su puesto de trabajo, independientemente de cuál sea la causa de la misma, en un período en que se ha programado la presencia de dicha persona.

Tener en cuenta el absentismo implica distinguir entre capacidad programada y capacidad efectiva. La probabilidad de coincidencia entre ambas es menor que uno, pues, a causa del absentismo, la capacidad efectiva puede ser menor que la programada. Dicha capacidad efectiva puede adoptar, en general, un cierto número de valores, a cada uno de los cuales se podrá asociar una probabilidad. En definitiva, se ha de tener en cuenta el absentismo al determinar la capacidad programada. Ésta ha de garantizar que el nivel de servicio deseado se alcanza con una probabilidad prefijada.

3. Propuesta

Con el fin de superar los inconvenientes o insuficiencias que se han indicado en el punto anterior, se propone un procedimiento para el cálculo de la capacidad que, en síntesis, consiste en:

a) Prever la demanda para cada uno de los períodos comprendidos en el horizonte de programación.

b) Calcular, mediante la aplicación, a cada período, de un modelo de teoría de colas un valor inicial de la capacidad para cada uno de los períodos.

c) Repetir, hasta que las capacidades obtenidas en dos iteraciones sucesivas coincidan (total o substancialmente): Simular el comportamiento del sistema para estimar el nivel de servicio en cada período y corregir las capacidades de acuerdo con el resultado de la simulación. Se obtiene así el valor de la capacidad requerida.

d) Calcular, a partir de la probabilidad de absentismo, el valor de la capacidad que se ha de programar para que la capacidad efectiva tenga una cierta probabilidad de ser igual o superior a la requerida.

En el paso b) se aplica un modelo de colas M/M/s, con la curva de demanda prevista desplazada tal como se propone en Green et al. (2001). Ello con independencia de cuáles sean las leyes que rigen el tiempo entre llegadas y los tiempos de servicio. Se trata sólo de obtener unos valores iniciales de capacidad para cada período, por lo que la coincidencia entre los supuestos que definen el modelo M/M/s y la realidad tiene relativamente poca importancia.

La especificidad del procedimiento corresponde a los pasos c) y d), de los cuales el c) es el más complejo y se describe a continuación.

El problema que se pretende resolver en este paso c) es el de determinar la capacidad en cada uno de los períodos de modo que se alcance el nivel de servicio deseado y se minimice una función de coste (que, por ejemplo, puede ser proporcional a la suma de las capacidades, para todos los períodos, o a la capacidad máxima en el horizonte de programación). La dificultad se deriva principalmente de que, dadas las capacidades, no se puede saber, salvo mediante la simulación, si son suficientes para alcanzar el nivel de servicio deseado; el problema no puede resolverse con ninguno de los algoritmos de optimización habituales, por lo que se requiere un procedimiento específico. Éste tiene carácter iterativo. En cada iteración se parte de una lista de capacidades (capacidad en cada período) y se simula, digamos N veces, el comportamiento del sistema, con el fin de estimar el nivel de servicio correspondiente a cada uno de los períodos y compararlo con el nivel de servicio deseado; de acuerdo con el resultado de esta comparación se procede a modificar las capacidades, incrementándolas o disminuyéndolas, según corresponda. Por supuesto, la clave de la eficiencia del proceso es el procedimiento de modificación de las capacidades, que debe alcanzar la convergencia en un número de iteraciones que no sea muy elevado. Cada iteración exige un tiempo no despreciable, puesto que el valor de N ha de ser suficientemente grande para que las estimaciones de los parámetros que caracterizan el nivel de servicio tengan poca dispersión (es decir, N ha de ser suficientemente grande para que dos simulaciones con la misma lista de capacidades arrojen resultados sensiblemente iguales); de otro modo, el comportamiento del algoritmo podría ser errático. Por supuesto, el valor de N sólo puede determinarse empíricamente, en cada aplicación.

Cuando con la lista de capacidades obtenida como resultado de una iteración las discrepancias entre el nivel de servicio obtenido en cada período y el nivel de servicio deseado son excesivas, deben modificarse, al alza o a la baja, las capacidades de la lista. Si las modificaciones introducidas en cada iteración son pequeñas se avanza con seguridad hacia la solución que se busca, pero la convergencia es lenta; por el contrario, cambios muy grandes pueden producir oscilaciones en los valores calculados para las capacidades. Para fijar ideas nos referiremos a la definición específica de nivel de servicio que se ha aplicado en el procedimiento (que se podría adaptar fácilmente a otras definiciones): La probabilidad de que el tiempo de espera sea superior a τ no debe ser superior a p. Por supuesto, el valor de p normalmente será bajo (por ejemplo: 0,05). Con las N simulaciones correspondientes a la iteración en curso se obtienen unas estimaciones p_t', de las probabilidades de que el tiempo de espera de una unidad, para cada período t, sea superior a τ. La modificación de la capacidad en cada período tiene dos componentes, que podemos llamar primario y secundario. El primero es función de la diferencia $p_t' - p$; concretamente, el valor absoluto de la

modificación de la capacidad (la cual tiene el mismo signo que la diferencia $p_t^{'} - p$) se calcula mediante la expresión $w_1 \cdot \ln\left(1 + 100 \cdot \min\left(p, \left|p_t^{'} - p\right|\right)\right)$; esta forma de calcular la modificación de capacidad es una entre otras posibles para tener en cuenta que las correcciones, en primer lugar, han de ser menos que proporcionales a la discrepancia entre el valor deseado y el valor obtenido y, en segundo lugar, de valor acotado superiormente para evitar "sobrecorrecciones" que podrían desestabilizar el proceso de convergencia (obsérvese que la limitación que se impone por medio de la expresión $\min\left(p, \left|p_t^{'} - p\right|\right)$ sólo tiene efecto para valores de $p_t^{'} > p$, ya que $p_t^{'} \geq 0$. El componente secundario, que se aplica a todos los períodos excepto el primero, tiene en cuenta las repercusiones sobre un período de la modificación de capacidad en el período inmediatamente anterior; consideremos, por ejemplo, dos períodos consecutivos ($t-1$ y t) tales que $p_{t-1}^{'}, p_t^{'} > p$; si se aumenta la capacidad en el período $t-1$, mejora el nivel de servicio en $t-1$, pero también en t, por lo cual el aumento de capacidad en t es menor del que sería necesario si el período se considerase aisladamente. En definitiva, las expresiones adoptadas para la modificación de la capacidad en cada período ($\delta_t;\ t = 1,...,T$) son:

$$\delta_1 = w_1 \cdot \ln\left(1 + 100 \cdot \min\left(p, \left|p_1^{'} - p\right|\right)\right)$$

$$\delta_t = w_1 \cdot \ln\left(1 + 100 \cdot \min\left(p, \left|p_1^{'} - p\right|\right)\right) - w_2 \cdot \delta_{t-1};\ t = 2,...,T$$

Un valor inicial adecuado de los pesos w_1 y w_2 debe determinarse experimentalmente, pero ambos pueden modificarse en el curso de la aplicación del proceso iterativo, de acuerdo con la magnitud de las modificaciones que se van obteniendo en el curso del mismo.

El proceso termina cuando la lista de capacidades se estabiliza.

La experiencia computacional llevada a cabo hasta ahora con este procedimiento ha permitido comprobar que el número de iteraciones necesario para alcanzar la convergencia es del mismo orden de magnitud que el número de períodos en que se divide el horizonte y que el tiempo de cálculo no excede de algunos minutos.

El resultado del paso c) descrito hasta aquí es la lista de capacidades efectivas requeridas para obtener el nivel de servicio deseado.

En el paso d) se determina qué capacidad hay que programar para que la probabilidad de que la capacidad no sea inferior a la capacidad efectiva no supere un valor dado. Según el contexto, se trata de determinar cuántas personas han de ser convocadas a sus puestos de trabajo en un período dado o cuál ha de ser la duración programada de la jornada para garantizar, con una probabilidad dada, respectivamente, un número mínimo de personas presentes o un número mínimo de horas de trabajo en el horizonte de programación. En cualquier caso la cuestión se reduce a resolver un problema de cálculo de probabilidades, que puede ser más o menos complejo en función de que las probabilidades de absentismo de las personas sean iguales o distintas, por una parte, o independientes o dependientes, por otra.

4. Perspectivas

Se prevé profundizar el análisis con el fin de mejorar y asegurar la robustez del cálculo de la lista óptima de capacidades en el paso c).

El procedimiento se aplicará para calcular la capacidad a programar como dato para modelos de planificación y programación del tiempo de trabajo y de la producción.

Referencias

Corominas, A.; Lusa, A.; Pastor, R. (2004). Planning Annualised Hours with a Finite Set of Weekly Working Hours and Joint Holidays. *Annals of Operations Research*, Vol. 128, No. 1-4, pp. 217-233.

Eick, S.G.; Massey, W.A.; Whitt, W. (1993). M_t/G/ infinity queues with sinusoidal arrival rates. *Management Science*, Vol. 39, No. 2, pp. 241-252.

Green, L.; Kolesar, P.; Soares, J. (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, Vol. 49, No. 4, pp. 549-564.

Ingolfsson, A.; Haque, M.D.A.; Umnikov, A. (2002). Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*, Vol. 139, No. 3, pp. 585-597.

Olivella, J. (2000). La organización industrial y el fenómeno del absentismo: un modelo aplicado a la empresa española. Tesis Doctoral, Universitat Politècnica de Catalunya.

Thompson, G.M. (1990). Shift scheduling in services when employees have limited availability: an L.P. approach. *Journal of Operations Management*, Vol. 9, No. 3, pp. 352-370.

Thompson, G.M. (1995). Labor scheduling using NPV estimates of the marginal benefit of additional labor capacity. *Journal of Operations Management*, Vol. 13, No. 1, pp. 67-86