

PIA Inv de Op

TEORÍA DE COLAS

Aldo Fernando Hernández Tamez
Abraham Rogelio López García

INTRODUCCIÓN

La teoría de colas es una rama de las matemáticas que estudia el comportamiento de líneas de espera, buscando optimizar la eficiencia en los sistemas de servicio, en ella se analizan elementos como las tasas de llegada de usuarios, las tasas de servicio, el número de servidores y el orden de atención, entre otros factores.

Este estudio se basa en un sistema de colas en serie (o tandem), donde distintos flujos de pacientes deben ser canalizados hacia un único punto de atención especializado, también conocido como un sistema de colas de múltiples entradas y una sola salida.



NIVELES DE ATENCIÓN. IMSS

Dentro del sistema de salud, la capacidad hospitalaria se enfoca en atender las necesidades de los pacientes a través de la organización, gestión del personal y uso adecuado de los recursos físicos y tecnológicos en función de las características y demanda del establecimiento de salud.

Así nació la necesidad de clasificar las unidades por niveles de atención para coordinar adecuadamente el flujo de pacientes y evitar la saturación, habiendo así tres fases o niveles de atención. A pesar de este sistema, la demanda de servicios de salud por unidad sigue siendo bastante alta.



Es por esto que en 2017, el IMSS implementó el sistema UNIFILA para atender a pacientes sin cita y aunque la idea es prometedora, en la práctica ha sido contraproducente, ya que muchos pacientes sin cita saturan los consultorios ocupados por quienes ya tienen cita.

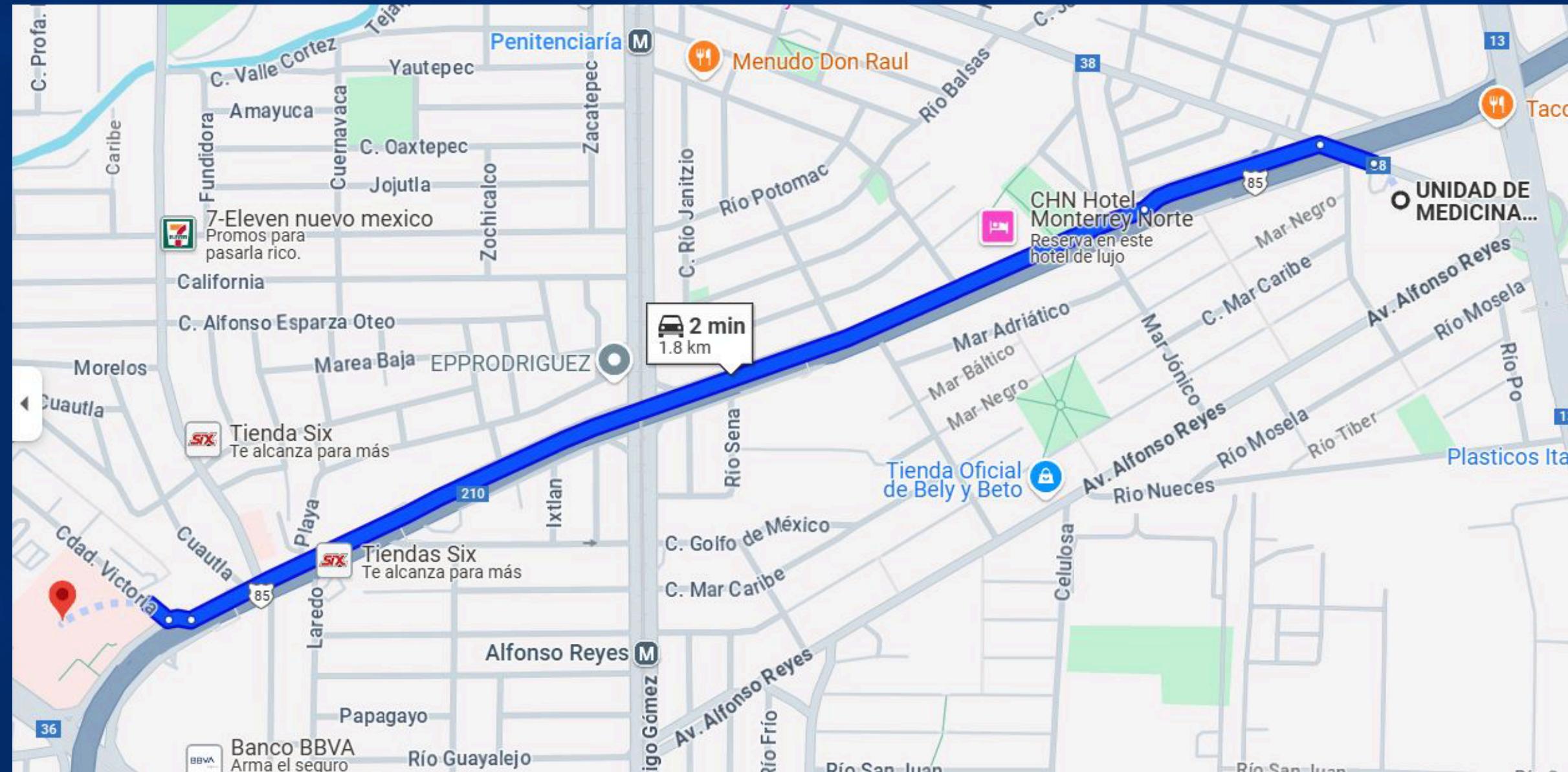
El problema que abordaremos se centra en el seguimiento de pacientes del primer nivel de atención hacia el tercer nivel dentro del sistema de salud del IMSS (Instituto Mexicano del Seguro Social), para ello, utilizaremos dos conjuntos de datos de consultas: uno sobre unidades de primer nivel y otro sobre las de tercer nivel. Estos datos contienen información de atenciones médicas realizadas en septiembre de 2017 mediante la encuesta EnSat [3] llevada a cabo al menos una vez al año por el IMSS.



CONTEXTO

En el sistema de salud del IMSS, los pacientes deben acudir primero a unidades de primer nivel, donde se brinda atención ambulatoria general o especializada, enfocada en prevención, promoción de la salud y detección temprana, estas son la vía de entrada al sistema de atención y, en caso necesario, son referidos a hospitales de tercer nivel para atención especializada y de urgencias.

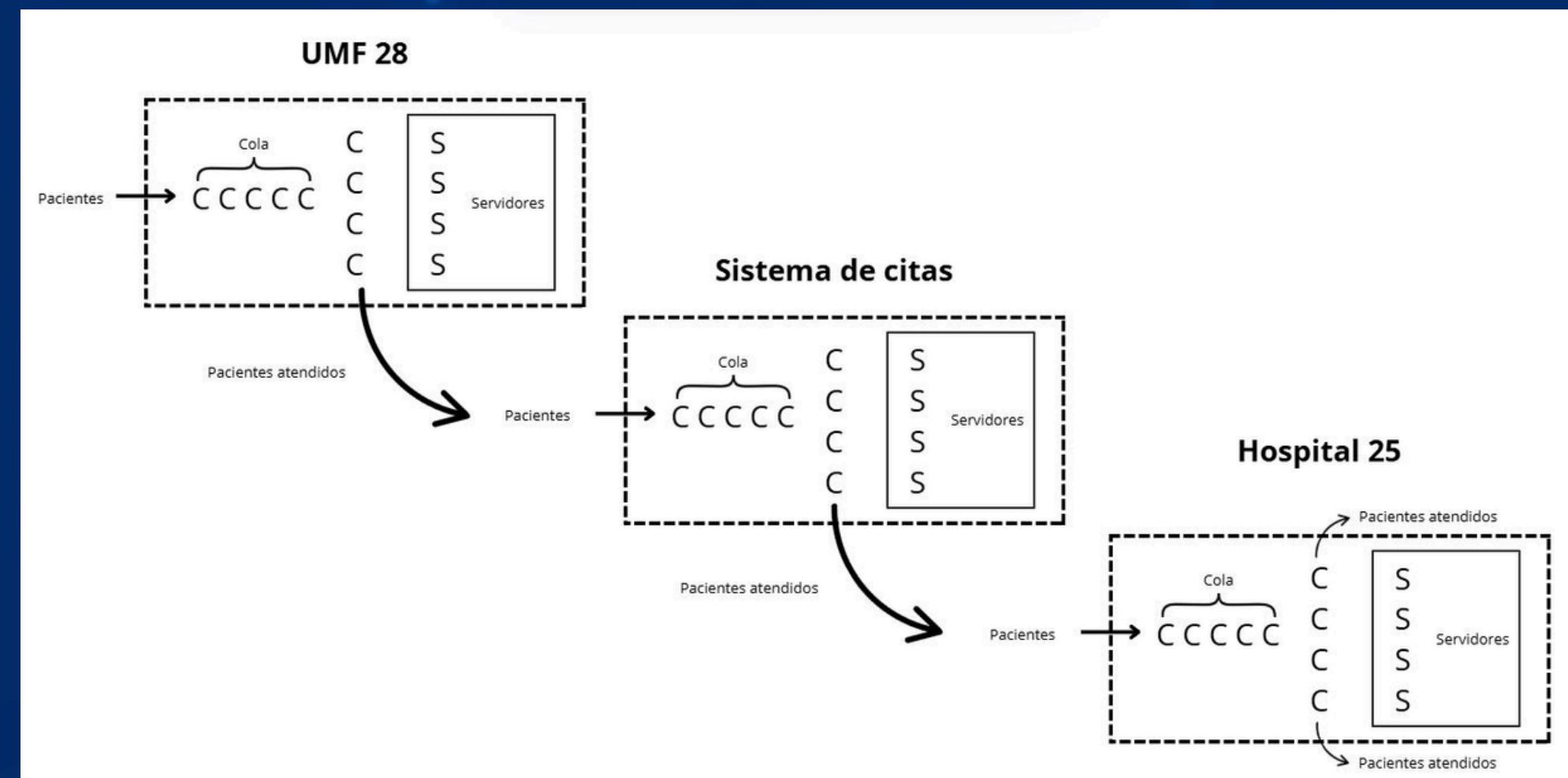
Este proyecto se centrará específicamente en la Unidad Médica Familiar No. 28 (UMF 28) de Monterrey, Nuevo León, como punto de primer nivel, y en el Hospital de Especialidades No. 25 de Monterrey como punto de tercer nivel, ambos centros de atención se encuentran geográficamente cercanos, lo cual facilita el análisis de la transferencia de pacientes entre ellos.



El flujo de atención será modelado en tres etapas:

1. **Primera etapa (UMF 28):** Los pacientes acuden a consulta general, donde son atendidos por múltiples médicos de primer contacto. Esta situación se modela mediante un sistema de colas M/M/s.
2. **Segunda etapa (Proceso de referencia):** Los pacientes que requieren atención especializada son referidos a través de un proceso administrativo, modelado como un sistema de colas M/M/1.
3. **Tercera etapa (Hospital 25):** Los pacientes son atendidos por múltiples especialistas de tercer nivel, modelado nuevamente como un sistema de colas M/M/s.

Una vez que el paciente recibe la atención especializada, existen tres posibles desenlaces: el paciente se cura y no requiere más atención (salida del sistema), el paciente necesita seguimiento y debe regresar al hospital, o el paciente abandona el tratamiento por voluntad propia. Cabe señalar que este estudio se centrará en el caso de la salida del sistema, asumiendo un escenario ideal en el que el paciente busca recuperarse.



DATOS Y PROBLEMA

Antes de proceder con el análisis, será necesario realizar un proceso de limpieza de los conjuntos de datos. Se ha decidido no hacer distinciones en cuanto a edad, sexo o tipo de paciente, a pesar de que existan columnas que contengan dicha información. Por lo tanto, se eliminarán los registros innecesarios y se seleccionarán únicamente aquellos campos relevantes para el propósito de este estudio, como los tiempos de espera y el servicio otorgado.

Además, es importante mencionar que se asumió una distribución de probabilidad exponencial tanto para los tiempos entre llegadas como para los tiempos de servicio en cada una de las colas. En las siguientes subsecciones se explicarán los procesos llevados a cabo en cada una de ellas.

PROBLEMA GENERAL

Una persona con seguro social en el IMSS comienza a presentar síntomas de alguna enfermedad, por lo que acude a la Unidad de Medicina Familiar 28 para una consulta de primer nivel. Al llegar al centro de salud, debe esperar en una sala hasta que se libere un consultorio para ser atendido. Una vez que el médico en turno lo revisa, determina que su caso requiere una evaluación por parte de un especialista, por lo que es referido al Hospital de Especialidades 25.

Sin embargo, para poder ser atendido en el hospital, el paciente debe tramitar una cita previa, cuya espera puede tomar varios días. Finalmente, el día de su cita, acude al hospital y nuevamente debe esperar en una sala hasta ser llamado a consulta. Tras la evaluación, el paciente inicia su tratamiento y, eventualmente, se le da de alta.

PRIMERA COLA (M/M/S1)

Los pacientes se generan a partir de una **fuente de entrada finita**, pero lo suficientemente grande como para no preocuparnos por la población ya que es prácticamente imposible que todos vayan a la misma UMF, por lo que se tomará como fuente ilimitada. Se asumirá una **distribución de Poisson** para las llegadas de pacientes, bajo el supuesto de que estas ocurren aleatoriamente con una tasa media fija, lo que implica que el tiempo entre llegadas sigue una **distribución exponencial**. Se ignorarán los casos en los que los pacientes deciden no ingresar al sistema, ya que en un escenario ideal se prioriza la salud.

La cola se considerará de capacidad infinita, aun cuando tenga un límite superior en la práctica, ya que incorporar esta restricción complicaría el análisis. Se utilizará una disciplina de atención **FCFS**, es decir, los pacientes serán atendidos en el orden en que lleguen a la sala de espera.

La estación de servicio contará con s_1 servidores activos durante toda la jornada de 12 horas y los tiempos de servicio seguirán una distribución exponencial para todos los pacientes. Cabe señalar que el valor de s_1 aún no está definido, por lo que se estudiarán distintos escenarios para encontrar un equilibrio entre costos y tiempos de espera.

Finalmente, se medirán los siguientes parámetros para esta cola:

$$\left\{ \begin{array}{l} \frac{1}{\lambda} = \frac{720 \text{ minutos}}{1599 \text{ pacientes}} = 0.45 \text{ minutos por paciente,} \\ \lambda = 2.22 \text{ pacientes llegan en promedio por minuto,} \\ \frac{1}{\mu} = 15 \text{ minutos por paciente,} \\ \mu = 0.066 \text{ pacientes por minuto atendidos en promedio,} \\ W_q = 18.64 \text{ minutos en promedio en la cola,} \\ W = 33.64 \text{ minutos en promedio en el sistema,} \\ [L] = 75 \text{ pacientes en promedio en el sistema,} \\ [L_q] = 41 \text{ pacientes en promedio en la cola} \end{array} \right.$$

Para obtener esta información bastó con revisar la información del conjunto de datos [3] para conocer los pacientes diarios en promedio en la UMF 28 Monterrey en 2016 y asumir un tiempo de servicio de 15 minutos. Además, se consultó el conjunto de datos [3] para obtener el tiempo de espera promedio en la cola mediante una media de datos agrupados según los datos que se muestran en la figura 3. Para los valores de L y L_q se optó por usar la función piso o techo según si los decimales eran despreciables o no, respectivamente, ya que tienen que ser números enteros.

Además, podemos deducir a partir de L y L_q que hay en promedio 34 pacientes atendidos, lo que indica que para que el sistema sea estable entonces $s_1 \geq 34$.

Esto también se demuestra con la siguiente desigualdad:

$$\rho = \frac{\lambda}{s_1 \mu} = \frac{2.22}{s_1(0.066)} < 1 \therefore s_1 > 33.64$$

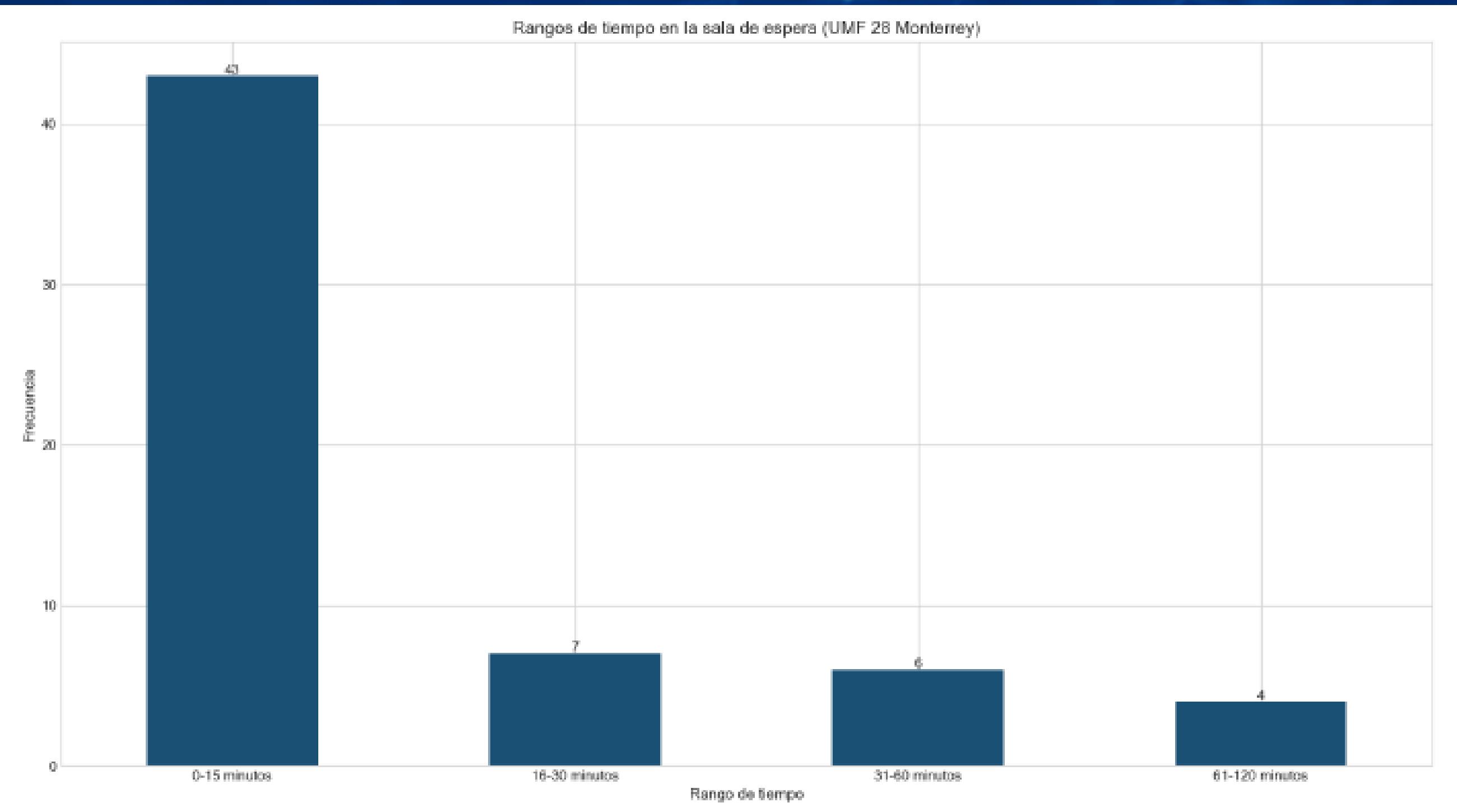


Figure 3. Gráfica de frecuencias para los tiempos de espera en la UMF 28.

SEGUNDA COLA (M/M/1)

Los pacientes se generan en una **fuente de entrada finita** lo suficientemente grande como para no preocuparnos por la población, debido a que en el flujo del sistema estos pacientes son aquellos que, al salir de su consulta en alguna unidad de primer nivel de atención, requieren una cita posterior en una unidad de tercer nivel; por lo tanto, se tomará como una fuente ilimitada. Además, se asumirá una **distribución Poisson** para las llegadas de pacientes bajo el supuesto de que llegan aleatoriamente con una cierta tasa media fija, por lo que el tiempo entre llegadas sigue una **distribución exponencial**.

Se ignorarán los casos en los que los pacientes se rehúsan a entrar al sistema (debido a que en un caso ideal su salud debería ser la prioridad), junto con la distinción de grupos entre los pacientes (por edad, enfermedad, entre otros).

La cola se considerará infinita a pesar de tener alguna cota superior, ya que tomarla en cuenta complicaría el análisis. Esta cola tendrá una disciplina FCFS, de manera que los pacientes serán atendidos conforme vayan solicitando su cita en el IMSS.

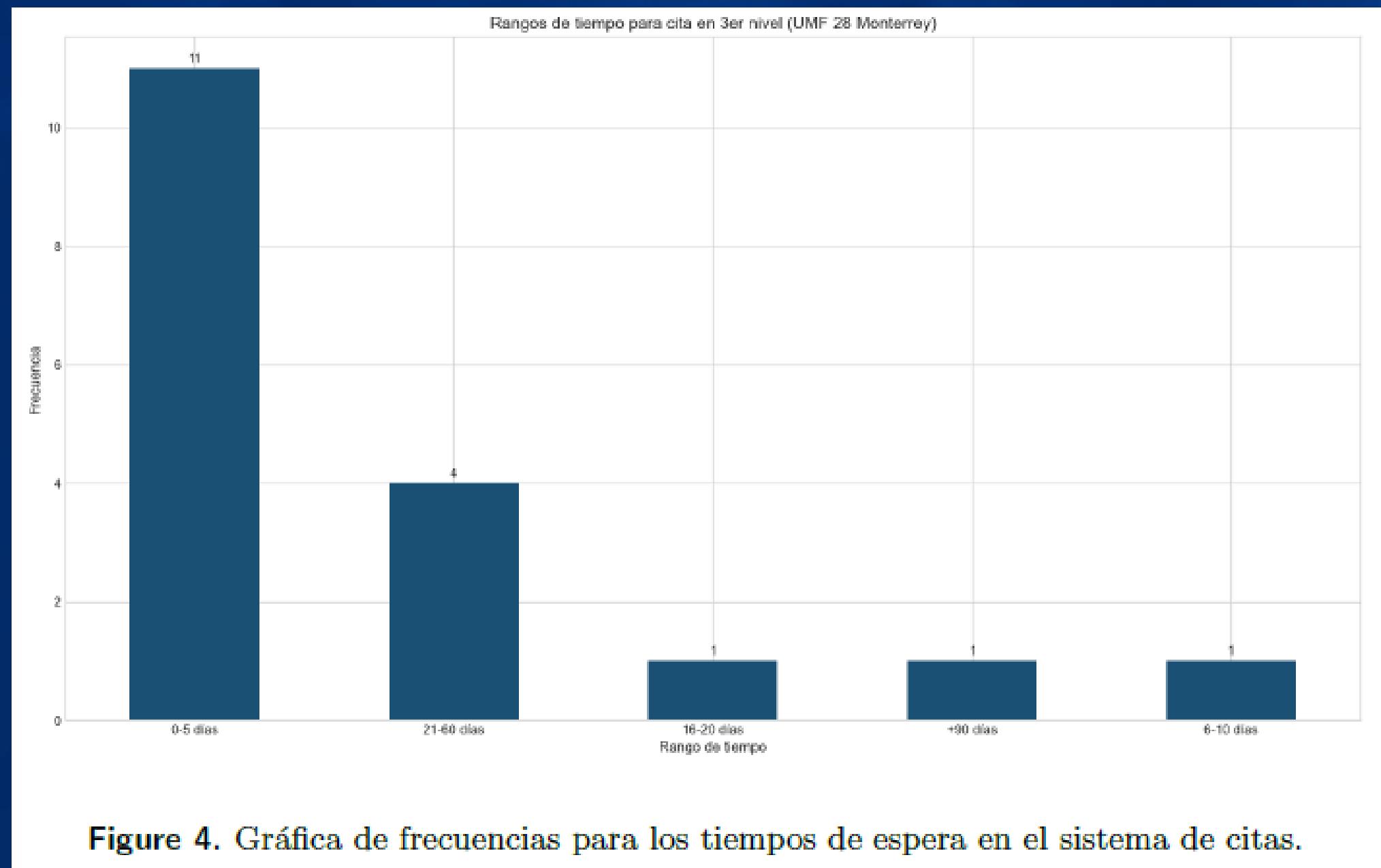
La estación de servicio contiene un único servidor activo (el "sistema") las 24 horas del día, y sus tiempos de servicio siguen una distribución exponencial para todos los pacientes.

Finalmente, medimos los siguientes parámetros para esta cola:

$$\left\{ \begin{array}{l} \frac{1}{\lambda} = \frac{1 \text{ día}}{734880 \text{ pacientes}} = 0.00000136076 \text{ días por paciente (casi 9 pacientes por segundo)}, \\ \lambda = 734880 \text{ pacientes llegan en promedio por día}, \\ \frac{1}{\mu} = 0.00000136076 \text{ día por paciente}, \\ \mu = 734880 \text{ pacientes por día atendidos en promedio}, \\ W_q = 17.51999 \text{ días en promedio en la cola}, \\ W = 17.52 \text{ días en promedio en el sistema}, \\ [L] = 12,875,098 \text{ pacientes en promedio en el sistema}, \\ [L_q] = 12,875,097 \text{ pacientes en promedio en la cola} \end{array} \right.$$

Para obtener esta información se consultó el conjunto de datos con el fin de calcular el tiempo de espera promedio en la cola, utilizando una media de datos agrupados según los datos mostrados en la figura 4. También se asumió el valor de $1/\lambda$ de manera que:

4 pacientes/hora \times 12 horas/día \times 10 servidores/UMF \times 1531 UMF = 734,880 pacientes



TERCERA COLA (M/M/S2)

Los pacientes se generan en una **fuente de entrada finita**, lo suficientemente grande como para no preocuparnos por la población ya que es prácticamente imposible que todos vayan al mismo hospital de especialidades, por lo que se tomará como fuente ilimitada. Además, se asumirá una **distribución Poisson** para las llegadas de pacientes bajo el supuesto de que llegan aleatoriamente con una cierta tasa media fija dictada por el sistema de citas, por lo que el tiempo entre llegadas sigue una **distribución exponencial**.

Se ignorarán los casos en los que los pacientes se rehúsan a entrar al sistema (debido a que en un caso ideal su salud debería ser la prioridad), junto con la distinción de grupos entre los pacientes (por edad, enfermedad, entre otros).

La cola se considerará **infinita** a pesar de tener alguna cota superior, ya que tomarla en cuenta complicaría el análisis. Esta cola tendrá una disciplina **FCFS**. La estación de servicio contiene s_2 servidores activos la jornada completa de 12 horas, cuyos tiempos de servicio siguen una distribución exponencial para todos los pacientes. Nótese que s_2 todavía no está definida, por lo que estudiaremos distintos valores para dicha variable

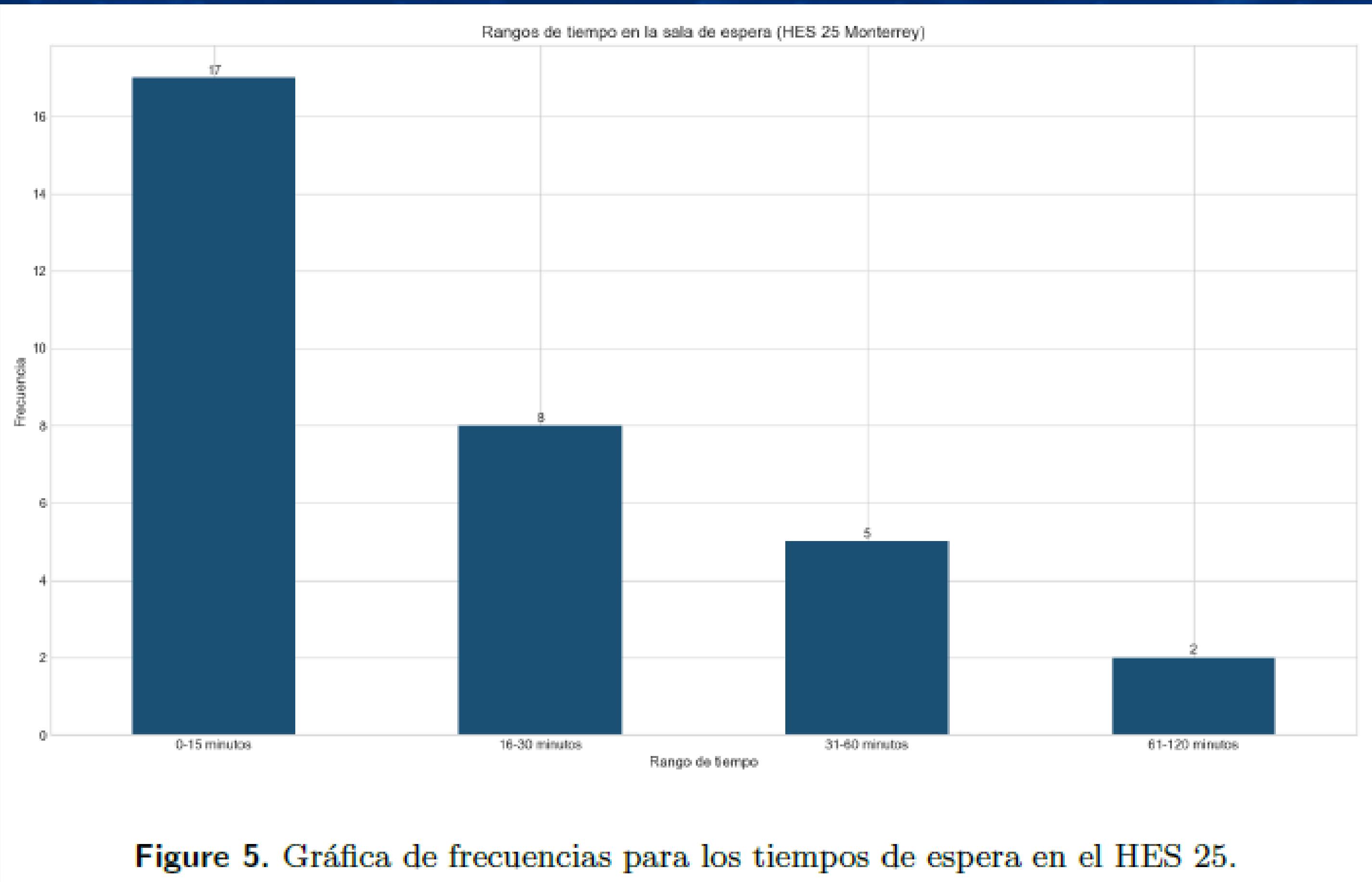
Finalmente, medimos los siguientes parámetros para esta cola:

$$\left\{ \begin{array}{l} \frac{1}{\lambda} = \frac{720 \text{ minutos}}{715 \text{ pacientes}} = 1.01 \text{ minutos por paciente,} \\ \lambda = 0.99 \text{ pacientes llegan en promedio por minuto,} \\ \frac{1}{\mu} = 30 \text{ minutos por paciente,} \\ \mu = 0.033 \text{ pacientes por minuto atendidos en promedio,} \\ W_q = 18.51 \text{ minutos en promedio en la cola,} \\ W = 48.51 \text{ minutos en promedio en el sistema,} \\ [L] = 49 \text{ pacientes en promedio en el sistema,} \\ [L_q] = 18 \text{ pacientes en promedio en la cola} \end{array} \right.$$

Para obtener esta información bastó con revisar el conjunto de datos para obtener dos cosas principalmente: el tiempo de espera promedio en la cola mediante una media de datos agrupados, según los datos que se muestran en la figura 5, y los pacientes diarios que acuden con un especialista en 2016. Por otro lado, para los valores de L y L_q se optó por usar la función piso o techo, según si los decimales eran despreciables o no, respectivamente, ya que tienen que ser números enteros.

Igual que en la primera cola, a partir de L y L_q podemos deducir que $s_2 \geq 31$ para que el sistema sea estable y que se cumpla la siguiente desigualdad:

$$\rho = \frac{\lambda}{s_2 \mu} = \frac{0.99}{(31)(0.033)} < 1$$



COLA DE ESPERA EN LA UMF 28

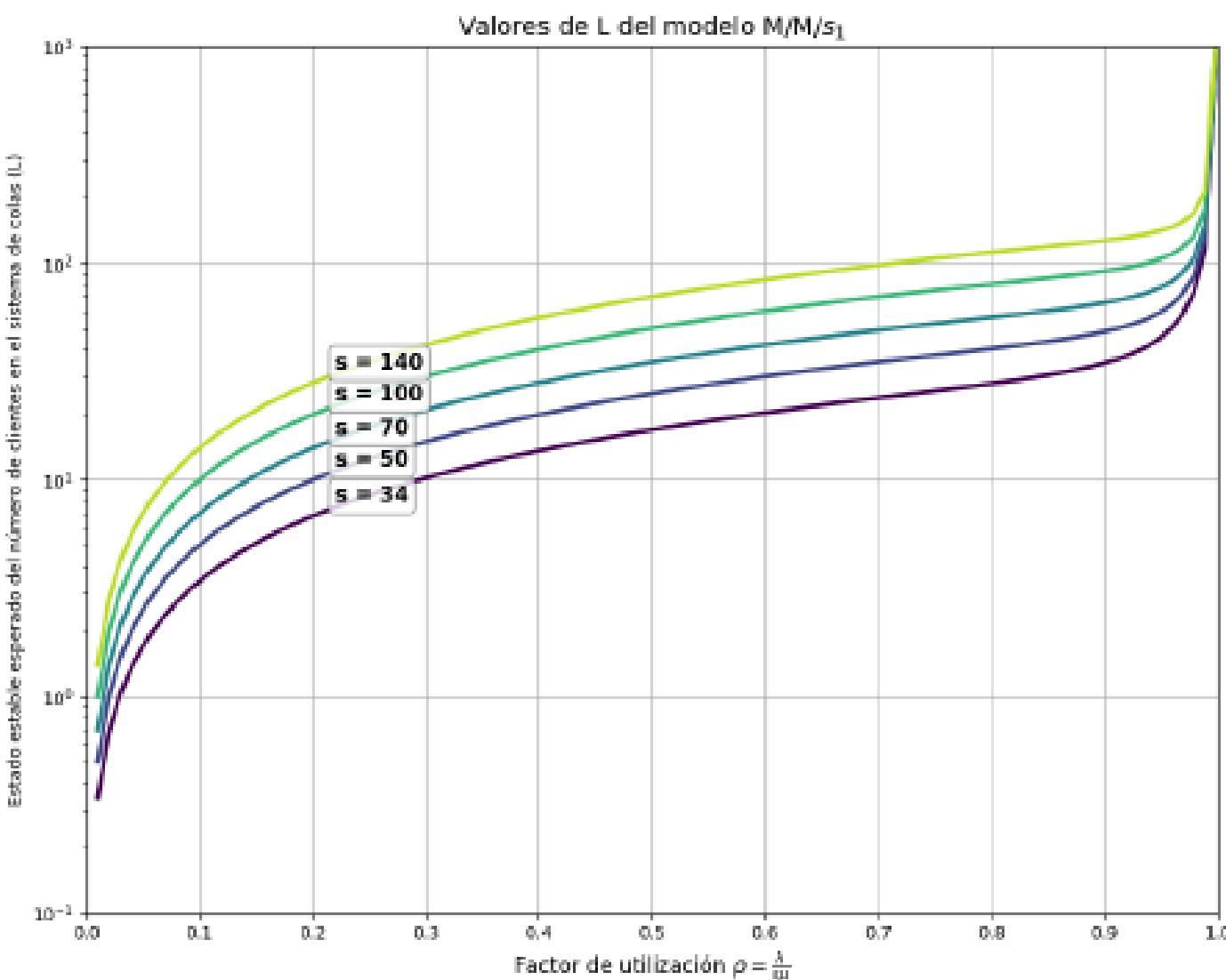
Si bien no es posible saber el número de servidores exactos debido a la falta de información, podemos dividir el análisis en dos partes: una cola estable y otra inestable.

Cola estable

Tomando en cuenta que $s_1 = 34$ (mínimo de consultorios para que sea estable), entonces ρ está muy cercano a uno, lo que significa que, esencialmente, la cola siempre estará saturada; es decir, las colas siempre serán largas y los servidores estarán en constante trabajo. Esto es negativo por diversas razones, las personas que requieren atención médica tendrán que esperar 18 minutos en promedio para ser atendidas, y los médicos generales estarán dando consultas continuas, lo cual implicaría mayores gastos para la unidad médica al tener que pagar más o mejores sueldos.

Tomando en cuenta los distintos valores de s_1 mostrados en la grafica, calculamos ρ :

$$\left\{ \begin{array}{l} \rho_1 = \frac{2.22}{(34)(0.066)} \approx 0.989; L_1 \approx 120 \\ \rho_2 = \frac{2.22}{(50)(0.066)} \approx 0.673; L_2 \approx 34 \\ \rho_3 = \frac{2.22}{(70)(0.066)} \approx 0.481; L_3 \approx 34 \\ \rho_4 = \frac{2.22}{(100)(0.066)} \approx 0.336; L_4 \approx 34 \\ \rho_5 = \frac{2.22}{(140)(0.066)} \approx 0.240; L_5 \approx 34 \end{array} \right.$$



(a) Valores de L para distintos servidores.

(b) Convergencia de L .

A pesar de que una cola con menos servidores es más eficiente bajo una carga relativa de trabajo igual, al aumentar el número de servidores en la cola podríamos reducir la cantidad de pacientes en espera.

Cola inestable

Si consideramos $s_1 \leq 33$, entonces esta cola estaría siempre saturada, operando bajo una sobrecarga de pacientes, lo que ocasionaría que no todos sean atendidos y que los servidores nunca dejen de trabajar. Esto sería un problema bastante grave, ya que la cola crecería indefinidamente con el tiempo hasta que se dejé de ofrecer servicio en la unidad médica (por los horarios de atención), y los tiempos de espera aumentarían durante el transcurso del día, generando disconformidad en los pacientes.





Cola de espera en sistema de citas

Para la segunda cola en el sistema, podemos asumir que se trata de una cola estable, pero extremadamente saturada. Esto explicaría los grandes tiempos de espera para poder acudir con cita al hospital deseado. Además, todos los pacientes en la cola, en algún momento, serán atendidos eventualmente debido a que asumimos que el "sistema" nunca descansa y siempre se están asignando citas.

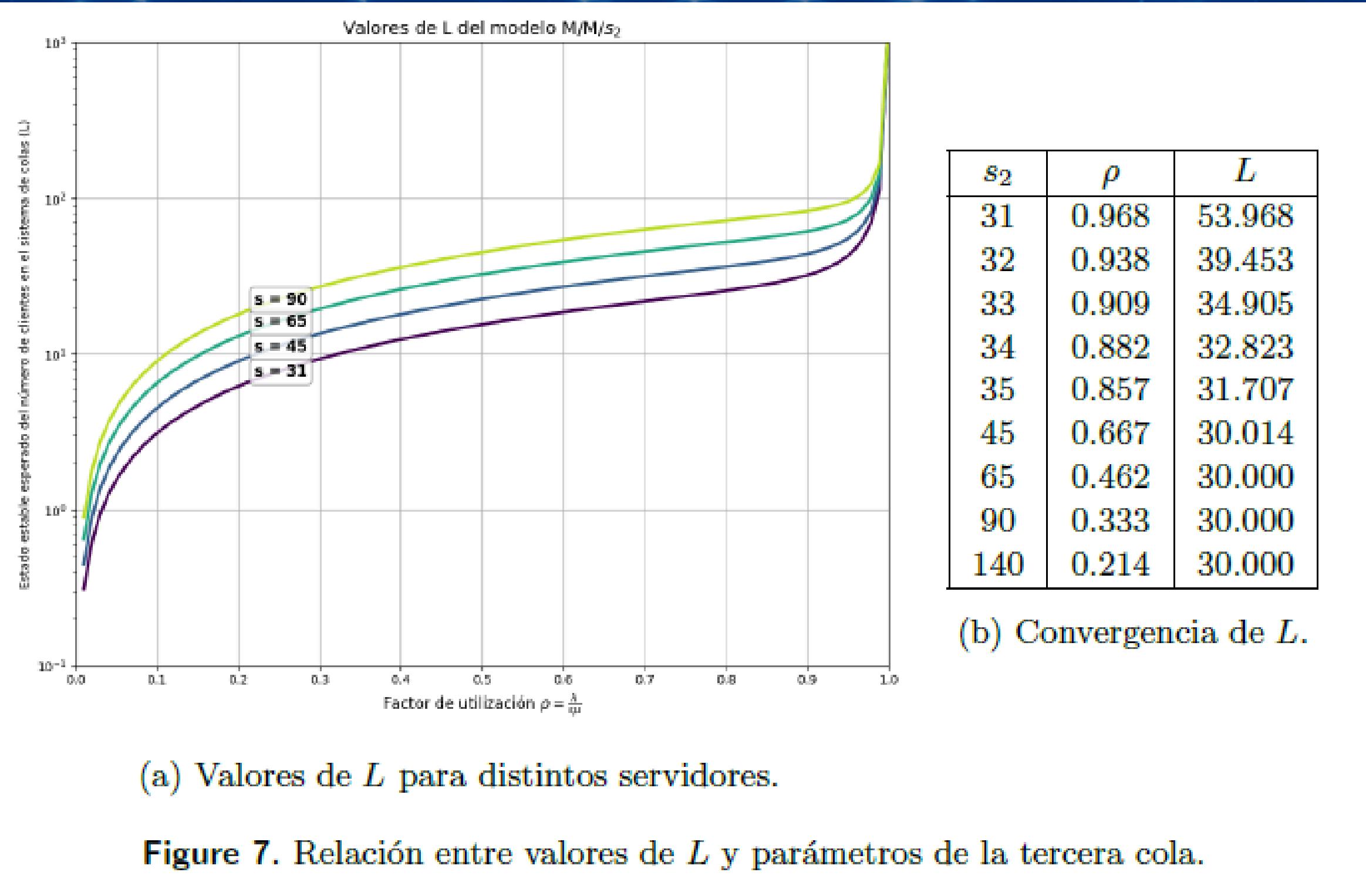
En este caso, es irrelevante hacer cálculos para ρ debido a lo ya mencionado y a que, al ser un sistema nacional, es normal que esté sobrecargado. La única manera de agilizar este procedimiento sería construyendo más centros de salud según el crecimiento de la fuente de entrada.

COLA DE ESPERA EN HES 25

Para que esta cola sea estable, se requiere que existan al menos 31 consultorios en el Hospital de Especialidades 25. Al no poder saber el número de servidores reales, el análisis se dividirá en dos partes, como en la primera cola: una estable y otra inestable.

Cola estable

Al igual que en la primera línea de espera, cuando $s_2=31$, entonces ρ está muy cerca de 1, indicando una gran carga de trabajo para el sistema. Aquí, los pacientes tienen que esperar en promedio 18 minutos antes de ser atendidos por un especialista; además, al tratarse de médicos especialistas, el sueldo que se tiene que pagar es mayor, generando más gastos para la unidad médica.



Observamos en la figura 7a que un sistema con menos servidores es más eficiente relativamente bajo un mismo valor de ρ ; pero si aumentamos el número de servidores, ρ disminuye considerablemente rápido. Además, con un aumento de 4 consultorios estaríamos prácticamente eliminando las colas de espera.

Cola inestable

Tomando en cuenta $s_1 \leq 30$, la cola estaría siempre bajo una carga excesiva de pacientes, ocasionando que no todos sean atendidos y que los médicos que ofrecen consultas nunca dejen de trabajar. Esto, además de generar los mismos problemas que el caso visto en el caso anterior, sería todavía más grave, ya que los pacientes que acuden a las unidades médicas de tercer nivel por lo general requieren seguimiento médico lo antes posible.



SISTEMA EN SERIE

En general, podemos decir que, en promedio, un paciente que se adapta a nuestro caso general explicado en la sección 2 dura aproximadamente 17 días con 13 horas en el sistema: 17 días con 12 horas en recibir una cita para el HES 25 y cerca de 1 hora esperando en filas.

Desde el punto de vista del paciente, esta cantidad excesiva de tiempo de espera no es ideal, aunque es hasta cierto punto normal por la cantidad de demanda que hay. Este tiempo podría deberse a varias razones, como el desequilibrio entre las tasas de llegada y la capacidad de servicio en cada una de las colas, las distancias entre unidades médicas que podrían ocasionar una sobrecarga de pacientes debido a que no hay ninguna otra opción cercana, o la cantidad de médicos disponibles.

SIMULACIÓN

A través de un programa en python se llevó a cabo una simulación de cada cola con número de servidores variable.

Es importante mencionar que los valores de esta tabla son distintos a los de la tabla 6b debido a que estos datos son simulados suponiendo un caso ideal, a diferencia de los datos en 6b que fueron realmente medidos y aproximados.

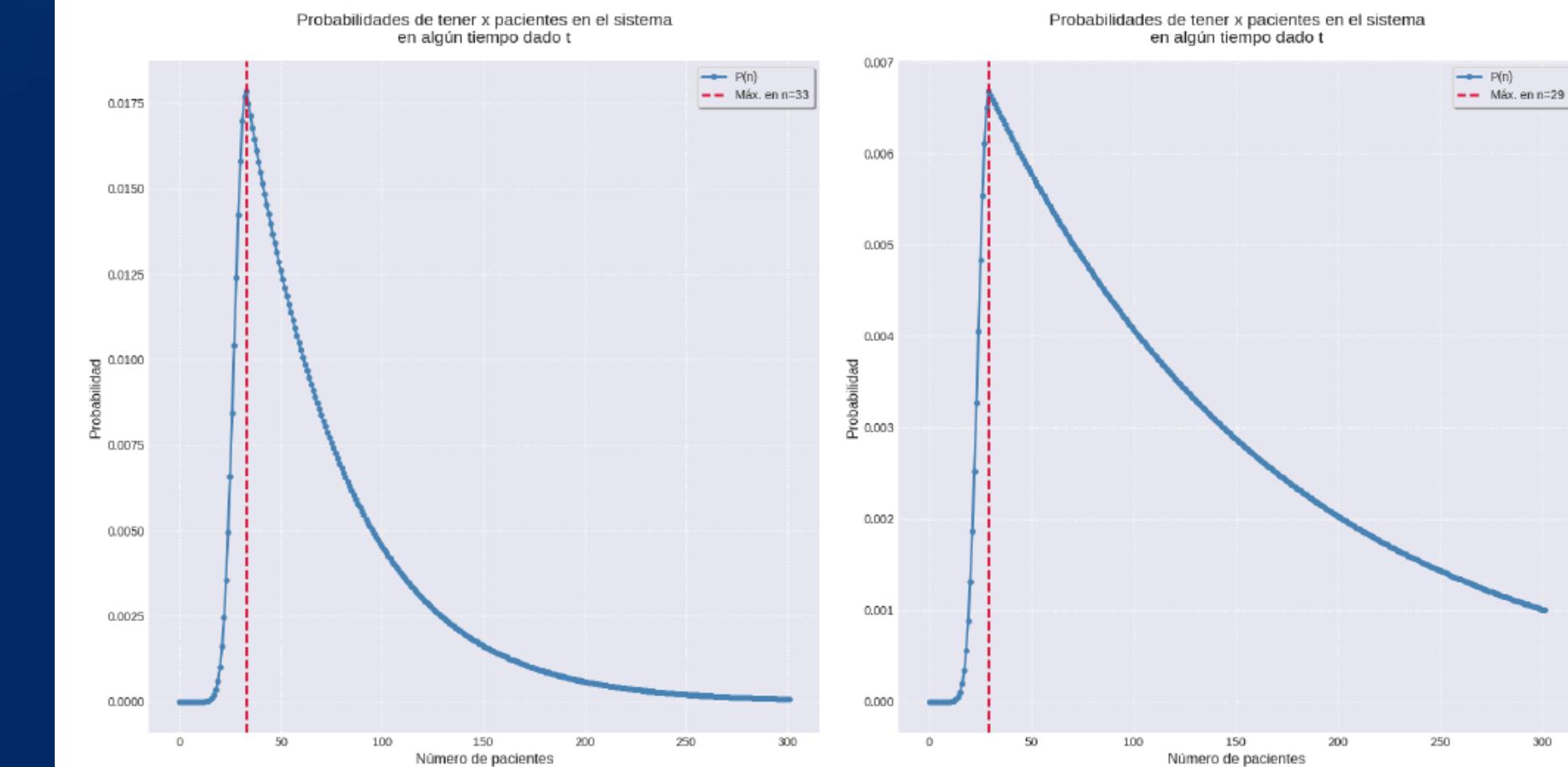
Pacientes en sistema	Servidores
75.2105	34
46.9958	35
40.1179	36
37.1919	37
35.6683	38
34.7909	39
34.2563	40
33.9198	41
33.704	42

(a) Distintos valores de L según s_1 .

Pacientes en sistema	Servidores
166.377	30
48.5426	31
37.8557	32
34.0949	33
32.2989	34
31.3168	35
30.74	36
30.3873	37
30.1668	38

(b) Distintos valores de L según s_2 .

Figure 8. Relación entre pacientes en promedio en el sistema y los servidores en la primera y tercera cola.



(a) Valores de P_n para distintos n en la primera cola.
(b) Valores de P_n para distintos n en la tercera cola.

Figure 9. Probabilidad de tener n pacientes en la primera y tercera cola.

Conclusiones

Según los datos publicados por el Instituto Mexicano del Seguro Social y las suposiciones hechas a lo largo del estudio, los resultados indican que el sistema es ineficiente a un grado tristemente normal. Si bien no se contó con la suficiente información para la investigación y el estudio se simplificó bastante, es difícil considerar la posibilidad de tener más de 34 médicos generales dando consultas en una sola Unidad de Medicina Familiar y más de 31 médicos especialistas en un Hospital de Especialidades.

Algunas posibles soluciones a este problema podrían ser el aumento de servidores en cada unidad médica, como se mencionó en la sección 4 (según el presupuesto), o una disciplina de cola distinta, como una cola de prioridad. Aunque estos cambios son difíciles de implementar, garantizan una mayor satisfacción en los pacientes atendidos.

Referencias

- [1] Capacidad hospitalaria: Factores, Gestión y Optimización. URL <https://www.medesk.net/es/blog/planeacion-de-la-capacidad-hospitalaria/>.
- [2] UNIFILA: Pacientes sin cita. URL <https://www.gob.mx/imss/acciones-y-programas/unifila-pacientes-sin-cita>.
- [3] IMSS: EnSAT 2017, 2017. URL <https://datos.gob.mx/busca/dataset/ensat-2017>.