ESTUDIO DE MODELO DE COLA M/M/1 Y DE SISTEMA DE INVENTARIO

Ramiro Di Giacinti
Ingeniería en Sistemas de Información
Universidad Tecnológica Nacional - FRRO
Zeballos 1341, S2000, Argentina
dia.digiacinti.ramiro@gmail.com

Bruno Mollo
Ingeniería en Sistemas de Información
Universidad Tecnológica Nacional - FRRO
Zeballos 1341, S2000, Argentina
dia.mollo.bruno@gmail.com

Facundo Braida
Ingeniería en Sistemas de Información
Universidad Tecnológica Nacional - FRRO
Zeballos 1341, S2000, Argentina
facundobraida98@gmail.com

Lucía Cappellini
Ingeniería en Sistemas de Información
Universidad Tecnológica Nacional - FRRO
Zeballos 1341, S2000, Argentina
luciacappll@gmail.com

Adriel Gorosito
Ingeniería en Sistemas de Información
Universidad Tecnológica Nacional - FRRO
Zeballos 1341, S2000, Argentina
adrielgorosito14@gmail.com

March 28, 2024

ABSTRACT

A partir del conocimiento netamente teórico de Teoría de Colas y de Sistemas de Inventarios, se implementaron dos modelos de simulación que permitieron comprender, e incluso, predecir el comportamiento de los mismos. En el presente informe se detallarán la metodología empleada y los resultados obtenidos.

1 Introducción

1.1 Teoría de Colas

La teoría de colas es un campo de estudio que se centra en el análisis matemático y estadístico de los sistemas de espera. Estos sistemas pueden encontrarse en una amplia variedad de contextos, como las colas en los supermercados, los centros de llamadas, los sistemas de transporte y muchas otras aplicaciones.

El objetivo principal de la teoría de colas es comprender y predecir el comportamiento de los sistemas de espera. A su vez, se basa en ciertas suposiciones fundamentales, como el supuesto de que los clientes llegan al sistema de forma aleatoria y siguen un patrón estadístico predecible. Además, se asume que el servicio se realiza de acuerdo con ciertas reglas, como el principio de FIFO (primero en entrar, primero en salir).

1.2 Sistema de Inventario

En su esencia, el sistema de inventario busca equilibrar los costos asociados con el almacenamiento de inventario y las oportunidades perdidas debido a la falta de stock. El objetivo principal es encontrar un equilibrio óptimo que minimice los costos de inventario y, al mismo tiempo, garantice que haya suficiente inventario disponible para satisfacer la demanda de los clientes o las necesidades operativas.

Al utilizar la teoría de inventarios, se pueden tomar decisiones informadas sobre cuándo realizar pedidos de nuevos productos, cuánto stock mantener en el inventario, cómo establecer niveles de reordenamiento y cómo manejar la incertidumbre de la demanda, y como estos se verán reflejados en los costos de la empresa.

Dentro de la gestión de inventarios, existen diferentes tipos de políticas y enfoques que se pueden utilizar para alcanzar los objetivos. Estas políticas y enfoques se basan en diversas estrategias para gestionar el inventario de manera eficiente y satisfacer las demandas de los clientes. Cada política tiene sus propias características y consideraciones, y su elección dependerá de los requerimientos específicos de la empresa, los productos y la demanda.

Algunas políticas se centran en mantener niveles de inventario mínimos y eficientes, minimizando los costos de almacenamiento y evitando el exceso de stock innecesario. Otras políticas se enfocan en adaptarse a la demanda estacional o a la variabilidad de la demanda aleatoria, asegurando un inventario adecuado en momentos de alta demanda o incertidumbre.

En el informe, se analizará una política específica, la política "tS", que implica realizar pedidos de una cantidad fija S en intervalos de tiempo t. Esta política fue seleccionada debido a su reconocimiento como una política de inventario ampliamente utilizada y estudiada en la gestión empresarial. Además, ofrece beneficios como su simplicidad y facilidad de implementación, así como de su capacidad para equilibrar los costos de inventario y garantizar la disponibilidad de productos para satisfacer la demanda. Se evaluará su efectividad en términos de costos, disponibilidad de productos y satisfacción de la demanda.

2 Metodología

2.1 Modelo de Colas

Un sistema de colas se puede describir de la siguiente forma. Un conjunto de “clientes” llega a un sistema en busca de un servicio, esperan si este no es inmediato, y abandonan el sistema una vez hayan sido atendidos. El término “cliente” es usado con un sentido genérico. el cliente puede ser ya sea una persona esperando la cola de un supermercado, piezas esperando su turno para ser procesadas o una lista de trabajo esperando para ser impresos en una impresora en red.

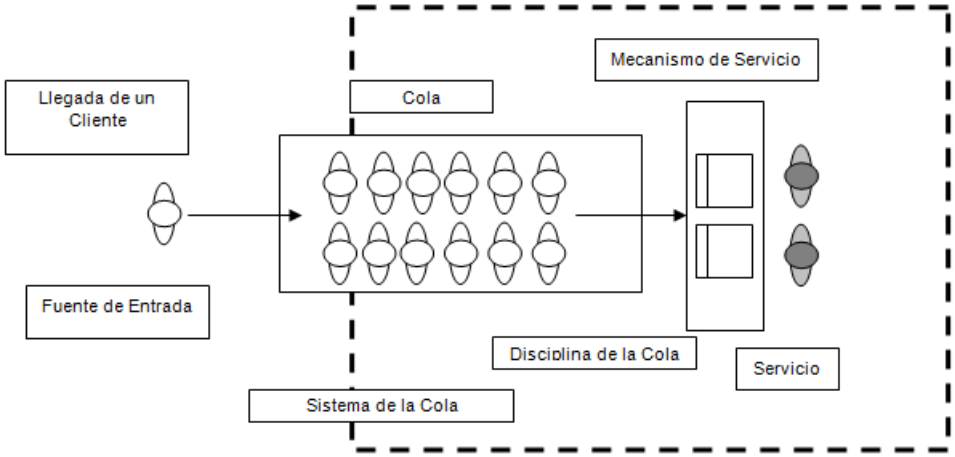

Figure 1: Modelo de colas genérico

2.1.1 Rendimiento de la cola

Notación:

- λ = Número de llegadas por unidad de tiempo
- μ = Número de servicios por unidad de tiempo si el servidor está ocupado
- ρ = Utilización del servidor
- $P_n(t)$ = Probabilidad que haya n clientes en el sistema en el instante t

- N Número de clientes en el sistema en el estado estable
- Pn = Probabilidad de que haya n clientes en estado estable
- Lq = Número medio de clientes en la cola
- Tq = Tiempo que un cliente invierte en la cola
- S = Tiempo de servicio
- T = Tiempo total que un cliente invierte en el sistema
- Wq = Tiempo medio de espera de los clientes en la cola
- Pb = probabilidad de que cualquier servidor esté ocupado
- t_i = tiempo de arribo (o entre llegadas)del i-ésimo cliente (t0=0)
- $A_i = t_i - t_{i-1}$: tiempo de arribos entre el $(i-1)$-ésimo y los arribos de los i-ésimos clientes
- S_i = tiempo que el servidor gasta atendiendo al i-ésimo cliente (Exclusivo del retraso del cliente en la cola)
- D_i: retraso de la cola del i-ésimo cliente
- $c_i = t_i + D_i + S_i$: tiempo en el que el i-ésimo cliente completa el servicio y se retira
- e_i: tiempo de ocurrencia del i-ésimo evento de cualquier tipo (i-ésimo valor que toma el reloj de la simulación, excluyendo el valor e0 = 0)

Considere un sistema de colas de servidor único para el cual los tiempos de arribo A1, A2, ... son variables aleatorias independientes, distribuidas idénticamente (ID) ("Distribuido idénticamente" significa que los tiempos entre llegadas tienen la misma distribución de probabilidad).

Un cliente quien llega y encuentra que el servidor inactivo ingresa al servicio inmediatamente, y los tiempos de servicio S1, S2, ... de los clientes sucesivos son variables aleatorias de ID que son independientes de los tiempos entre llegadas. Un cliente que llega y encuentra que el servidor está ocupado se une al final de una sola cola. Al completar el servicio para un cliente, el servidor elige a un cliente de la cola (si corresponde) en una primera entrada. primera salida (FIFO).

La simulación comenzará en el estado "vacío e inactivo"; es decir, no hay clientes presentes y el servidor está inactivo. En el momento 0, comenzaremos a esperar la llegada del primer cliente, que ocurrirá después del primer tiempo de arribo, A1 "en lugar de en el momento 0 (lo que sería posiblemente válido, pero diferente, supuesto de modelado).

Deseamos simular este sistema hasta que un número fijo (n) de clientes haya completado sus retrasos en la cola, es decir, la simulación se detendrá cuando el enésimo cliente entre en servicio. Tenga en cuenta que el tiempo de finalización de la simulación es, por lo tanto, una variable aleatoria, dependiendo de los valores observados para las variables aleatorias entre llegadas y tiempo de servicio.

Para evaluar el desempeño del sistema utilizaremos las siguientes medidas de desempeño:

- Utilización del servidor:
$$\rho = \frac{\lambda}{\mu}$$
donde λ es la tasa promedio de llegada de clientes y S es la tasa promedio de servicio del servidor.
- Promedio de clientes en el sistema:
$$L = \frac{\lambda}{\mu - \lambda}$$
donde L es el promedio de clientes en el sistema, L_q es el promedio de clientes en cola, λ es la tasa promedio de llegada de clientes y W es el tiempo promedio en el sistema.
  - Promedio de clientes en cola:
$$L_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$
  - Tiempo promedio en sistema:
$$W = \frac{L}{\lambda}$$
donde L es el promedio de clientes en el sistema (incluyendo los que están siendo atendidos y los que están en cola) y λ es la tasa promedio de llegada de clientes.

- Tiempo promedio en cola:

$$W_q = \frac{L_q}{\lambda}$$

donde L_q es el promedio de clientes en cola.

- Probabilidad de encontrar n clientes en cola:

$$P_n = (1-\rho)\rho^n$$

donde n es el número de clientes en cola.

- Probabilidad de denegación de servicio:

$$P_{den} = 1 - \sum_{i=0}^{n}(1-\rho)\rho^i$$

donde ρ es la utilización del servidor.

Por cuestiones de rendimiento en Python3, realizamos las simulaciones con 500 clientes; pero agregamos también una par de gráficas en las que llegamos a 4000 clientes en Anylogic.

2.2 Sistema de Inventario

Un sistema de inventario se puede describir de la siguiente forma. Se genera una demanda aleatoria a lo largo del tiempo, basada en una distribución de demanda esperada. A medida que se demanda, se realiza un seguimiento del inventario disponible y se aplican las políticas de reposición establecidas. Mientras tanto, se lleva el seguimiento de los costos relacionados con los costos de mantenimiento de productos que hayan quedado en stock, gastos por ventas perdidas, etc.

2.2.1 Variables del sistema

- A_i: tiempo entre arribos de clientes a comprar: $A_i \sim Exp(\lambda)$
  donde λ es la tasa de ocurrencia del evento.
- D: cantidad demandada por el cliente (Variable aleatoria con distribucion empirica): $D = \begin{cases} 1 & p = \frac{1}{6} \\ 2 & p = \frac{1}{3} \\ 3 & p = \frac{1}{3} \\ 4 & p = \frac{1}{6} \end{cases}$
- U: Demora del proveedor en traer pedidos $VA\ \mu \sim U(0.5; 1)$

2.2.2 Medidas de rendimiento finales

Para evaluar el desempeño, se evaluaran los siguientes costos:

- Costo de pedido:

$$Z = \begin{cases} S - I & siI < s \\ 0 & siI > s \end{cases}$$

$$C_{pedido} = k + iZ\ (si\ z \neq 0)$$

donde C_{pedido} es el costo de realizar un pedido de reabastecimiento, k es un costo fijo, i es el costo por unidad y Z es la cantidad de articulos pedidos

- Costo de mantenimiento:

$$C_{mantenimiento} = h \cdot \bar{I}^+$$

donde $C_{mantenimiento}$ es el costo de mantenimiento del inventario, h es el costo por unidad de mantener un producto en inventario y $\bar{I}^+$ es la cantidad de articulos en stock en promedio.

- Costo de faltante:

$$C_{faltante} = \pi \cdot \bar{I}^-$$

donde $C_{faltante}$ es el costo de faltante de productos, π es el costo por unidad faltante y U es la cantidad de productos que faltan en inventario en promedio.

- Costo total:

$$C_{total} = C_{pedido} + C_{mantenimiento} + C_{faltante}$$

donde C_{total} es el costo total del sistema de inventarios, que incluye el costo de pedido, el costo de mantenimiento y el costo de faltante.

3 Resultados

3.1 Teoría de colas M/M/1

3.1.1 Tasa de arribo igual a tasa de servicio

Realizamos experimentos con una tasa de arribo y tasa de servicio iguales a 1 y con una capacidad máxima de cola de 50.

Las gráficas a continuación muestran el tamaño de la cola respecto al tiempo, mostrando solo los primeros 100 arribos de clientes para que la gráfica se pueda apreciar bien:

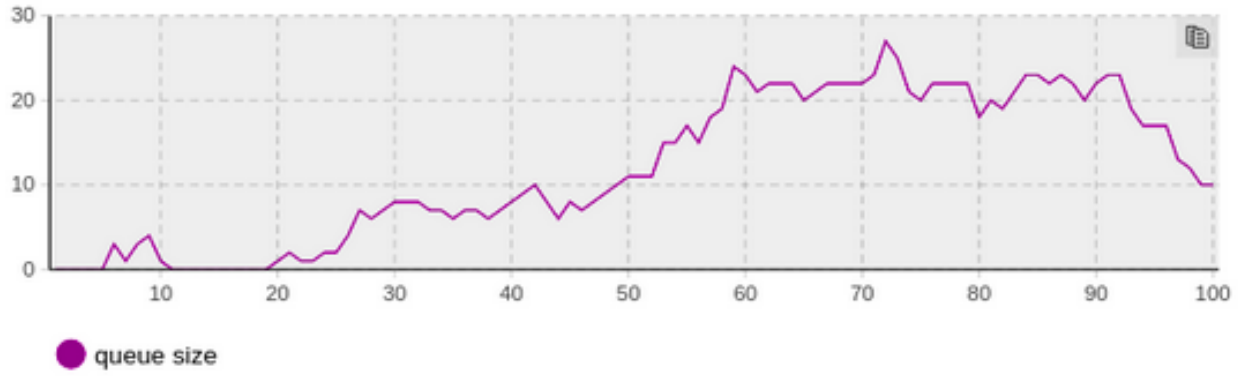

Figure 2: Tamaño de la cola en el tiempo, generado con Anylogic

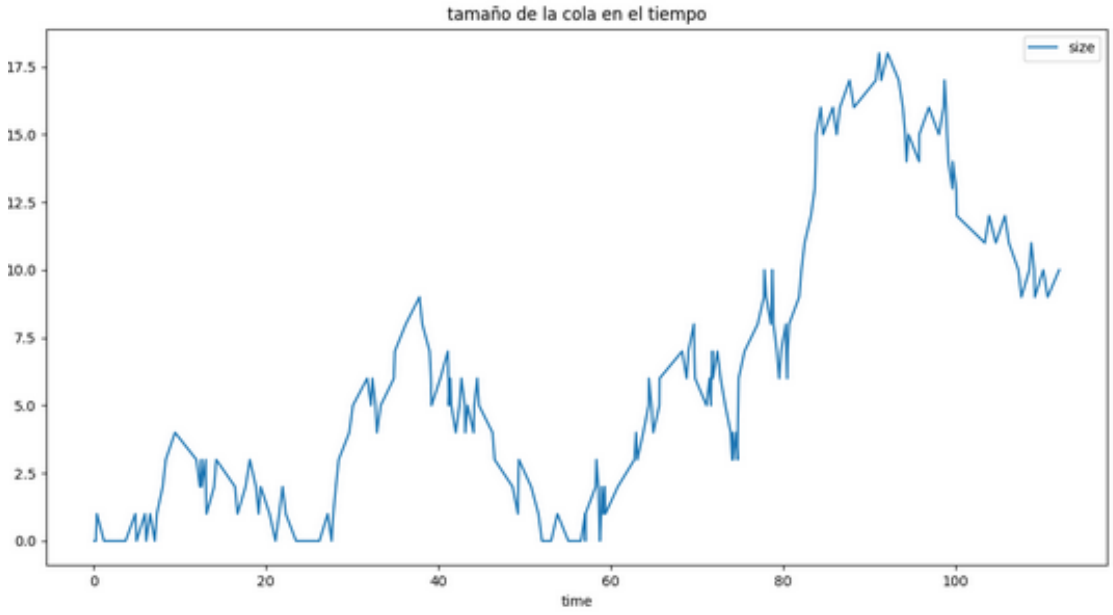

Figure 3: Tamaño de la cola en el tiempo, generado con Python3

Podemos apreciar que en ambas gráficas el tamaño de la cola muestra ciertas oscilaciones, pero una tendencia creciente mientras mas pasa el tiempo. Vemos que en una simulación tiene un valor máximo de cola mas grande, debido a la naturaleza estocástica del experimento y la muestra reducida de 100 clientes.

En las siguientes gráficas, se puede ver la distribución de frecuencias de los largos de la cola. Para generar estas gráficas usamos una muestra de 500 clientes.

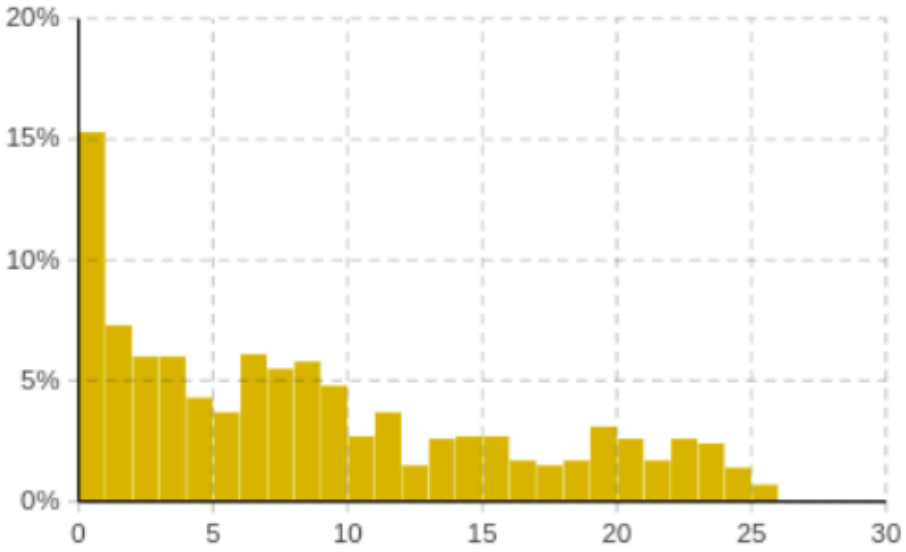

Figure 4: Frecuencia de los valores del tamaño de la cola, generado con Anylogic con 500 clientes

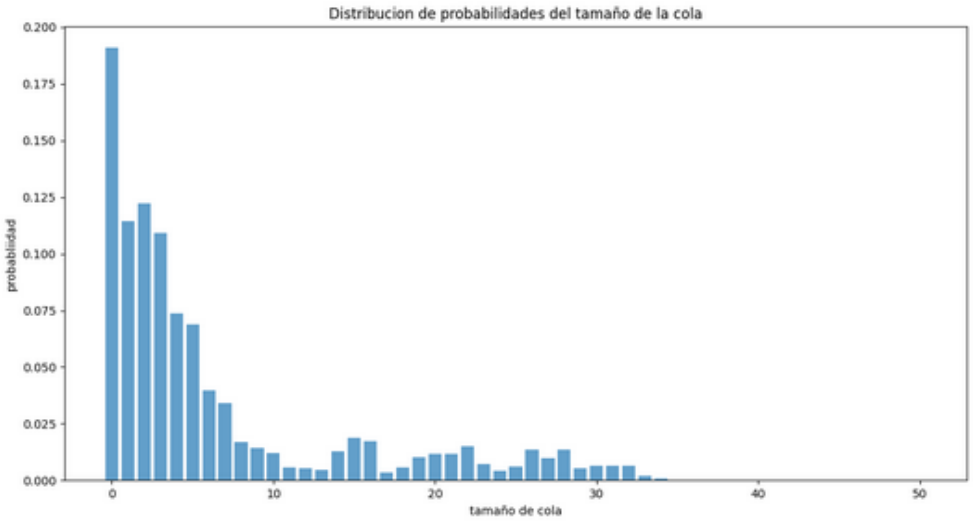

Figure 5: Frecuencia de los valores del tamaño de la cola, generado con Python3 con 500 clientes

Podemos apreciar, que con una muestra de de 500 clientes, el tamaño de la cola probablemente sea pequeño en un momento dado. Se puede ver una asimetría en al gráfica en que los valores mas cercanos a cero son mas frecuentes, siendo el cero el mas frecuente.

Aun así, decidimos llevar a cabo una simulación en Anylogic con 4000 clientes. Ya que por lo que deducimos al analizar las gráficas anteriores, el largo de la cola es proporcional al tiempo y a los clientes atendidos. El resultado de la grafica es fue el siguiente:

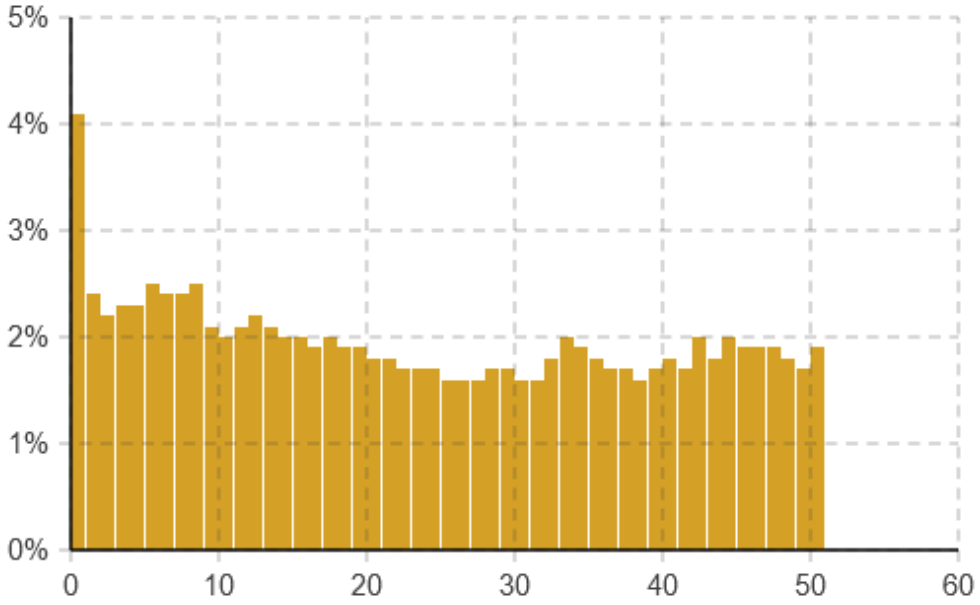

Figure 6: Frecuencia de los valores del tamaño de la cola, generado con Anylogic con 4000 clientes

Es evidente que, a pesar de que el cero sigue teniendo la probabilidad mas alta, la asimetría antes mencionada desparece con una muestra lo suficientemente grande, viéndose una distribución mas uniforme entre cero y la capacidad máxima de la cola.

	$Python$	$Anylogic$	$Teorico$
L	18.7329	19.927	Inf
W_q	16.0835	20.026	Inf

3.1.2 Tasa de arribo menor a tasa de servicio (50%)

Realizamos experimentos con una tasa de arribo de 1 y una tasa de servicio igual a 2, con una capacidad máxima de cola de 50. Tamaño de cola:

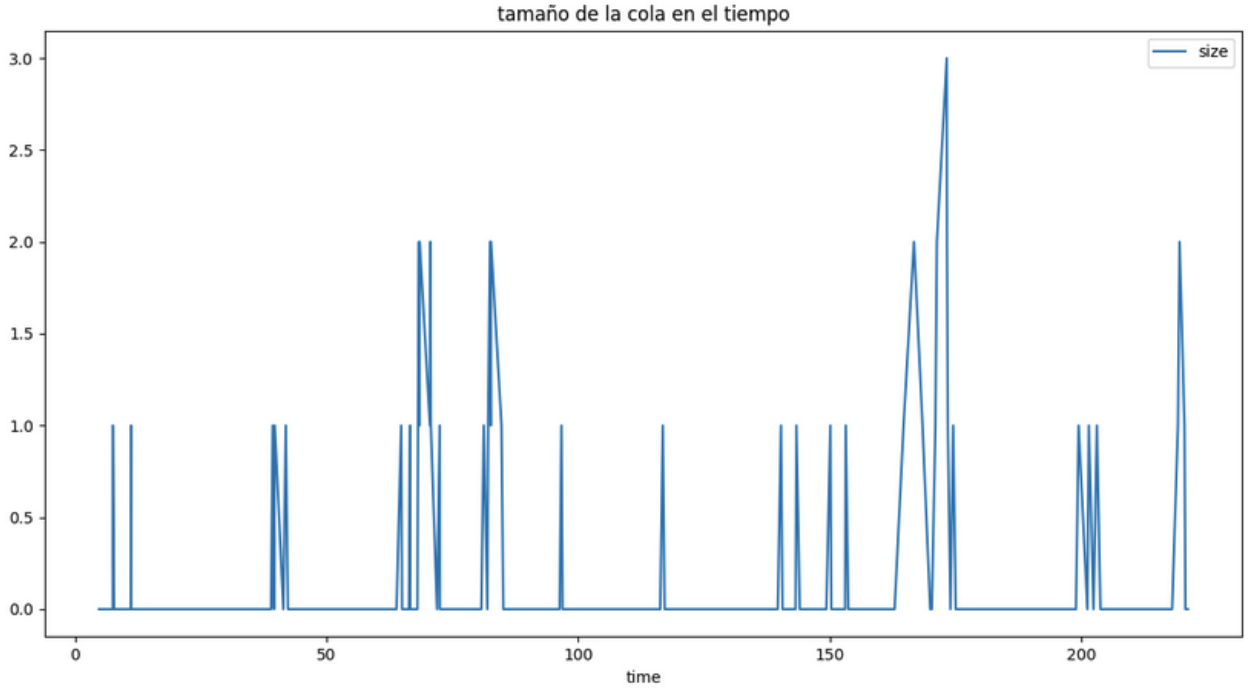

Figure 7: Tamaño de cola en Python

Figure 8: Tamaño de cola en Anylogic

Probabilidad de Tamaño de cola:

Como podemos observar, en este caso la cola nunca llega a ser muy extensa, ya que los clientes son atendidos mas rápido de lo que llegan nuevos clientes. Si se mide el largo de la cola en cualquier momento, es bastante probable que este vacía o que tenga un grupo reducido de clientes esperando, como se puede ver al analizar la frecuencias relativas.

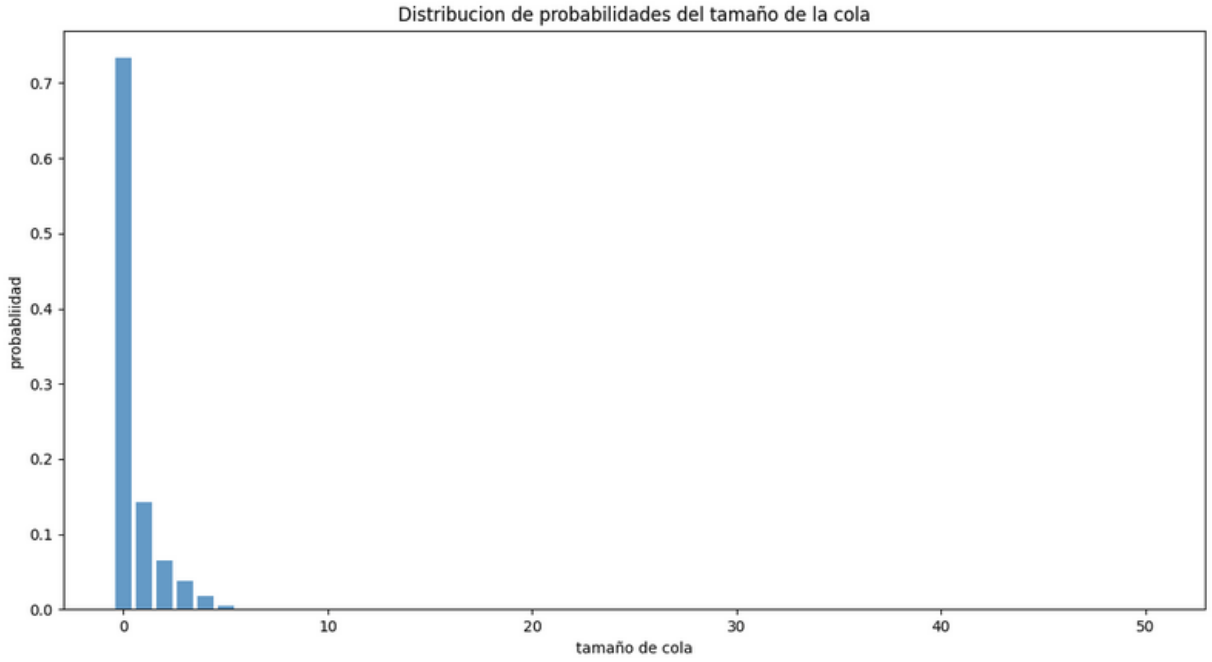

Figure 9: Probabilidad de Tamaño de cola en Python

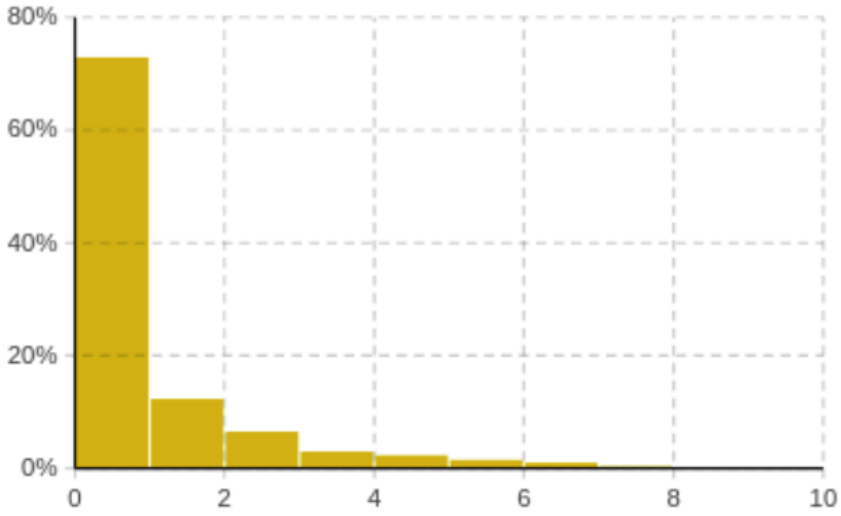

Figure 10: Probabilidad de Tamaño de cola en Anylogic

	$Python$	$Anylogic$	$Teorico$
L	47.2473	48.552	0.5
W_q	88.2017	95.533	1

3.1.3 Tasa de arribo mayor a tasa de servicio (125%)

Realizamos experimentos con una tasa de arribo de 5 y una tasa de servicio igual a 4, con una capacidad máxima de cola de 50.

Tamaño de la cola:

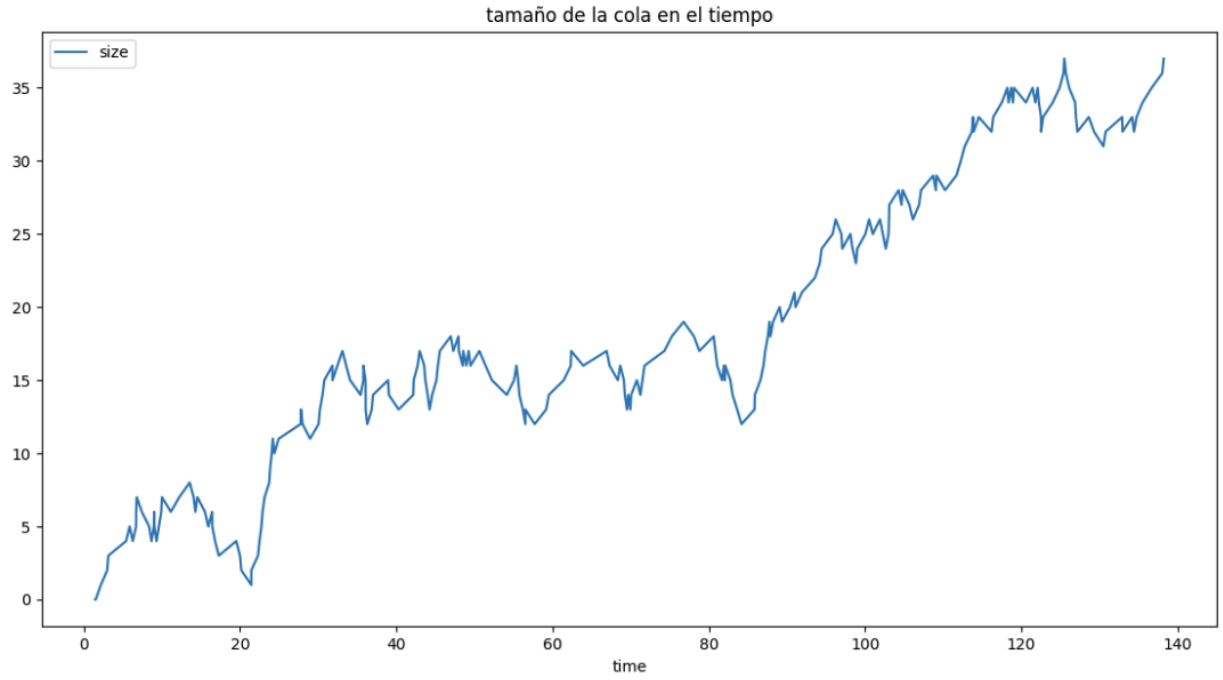

Figure 11: Tamaño de la cola en Python

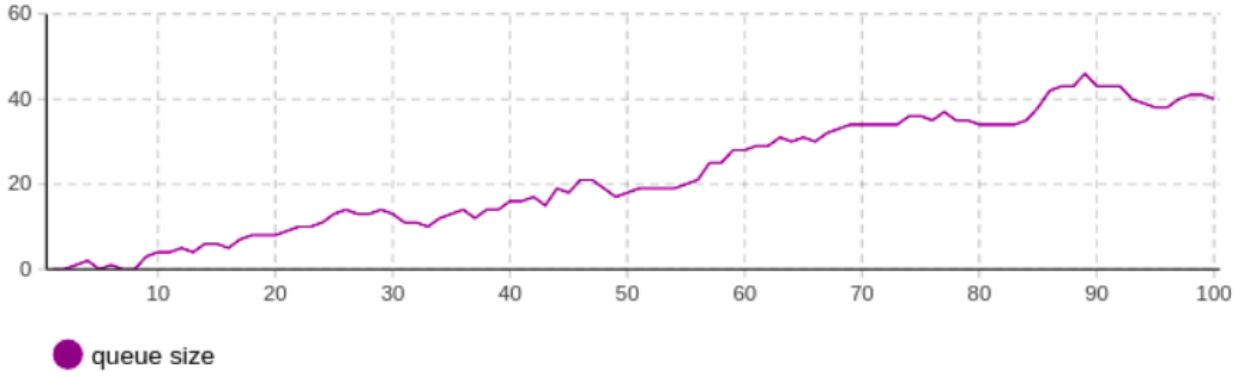

Figure 12: Tamaño de la cola en Anylogic

Probabilidades de longitud de cola:

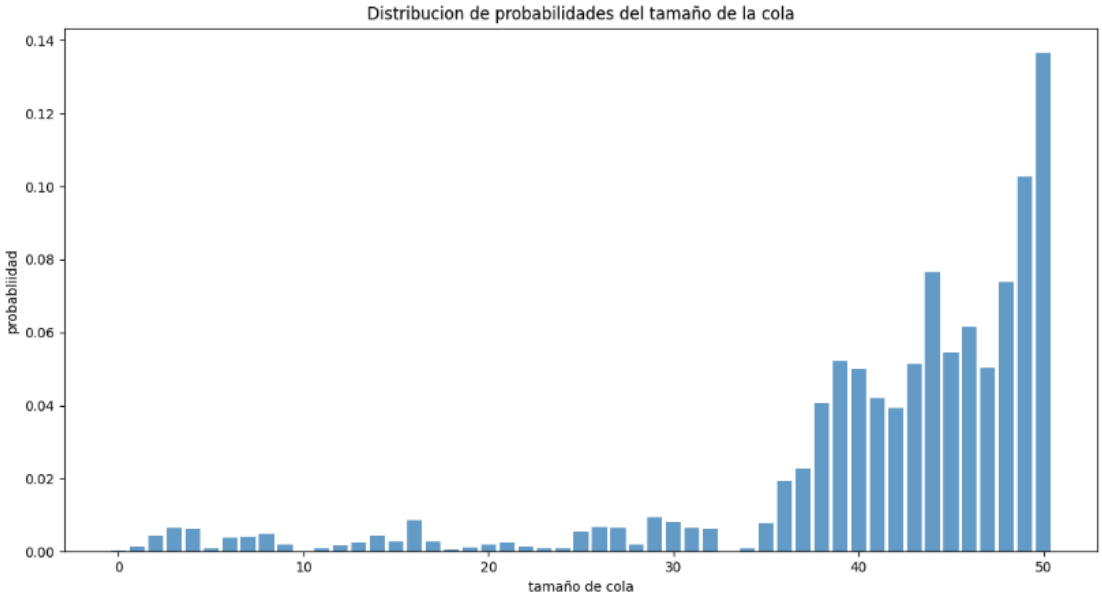

Figure 13: Probabilidades de longitud de cola en Python

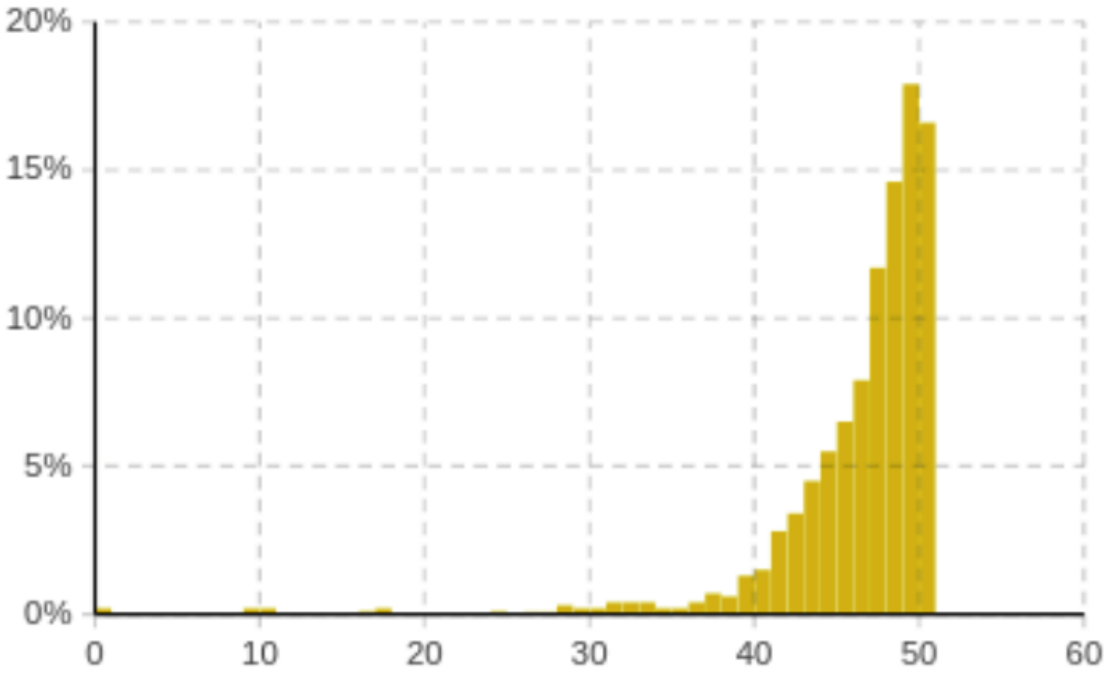

Figure 14: Probabilidades de longitud de cola en Anylogic

	$Python$	$Anylogic$	$Teorico$
L	38.547	44.452	5
W_q	42.3948	44.813	-5

En este caso, ocurre lo contrario al caso anterior. La cola tiende a estar colapsada y este crecimiento ocurre muy rápidamente. Si se mide la cola en cualquier instante, es muy probable que este llena; esto luego puede conllevar una gran cantidad de rechazos de clientes si tenemos una capacidad limitada en la cola.

3.1.4 Tamaños de cola y utilización del servidor

En la siguiente figura, se muestran gráficos de torta que indican el porcentaje de clientes atendidos (azul) y clientes rechazados por capacidad de cola alcanzada (rojo). Podemos apreciar que este estadístico depende de la capacidad de la cola y la relación del tiempo de arribo y el tiempo de servicio.

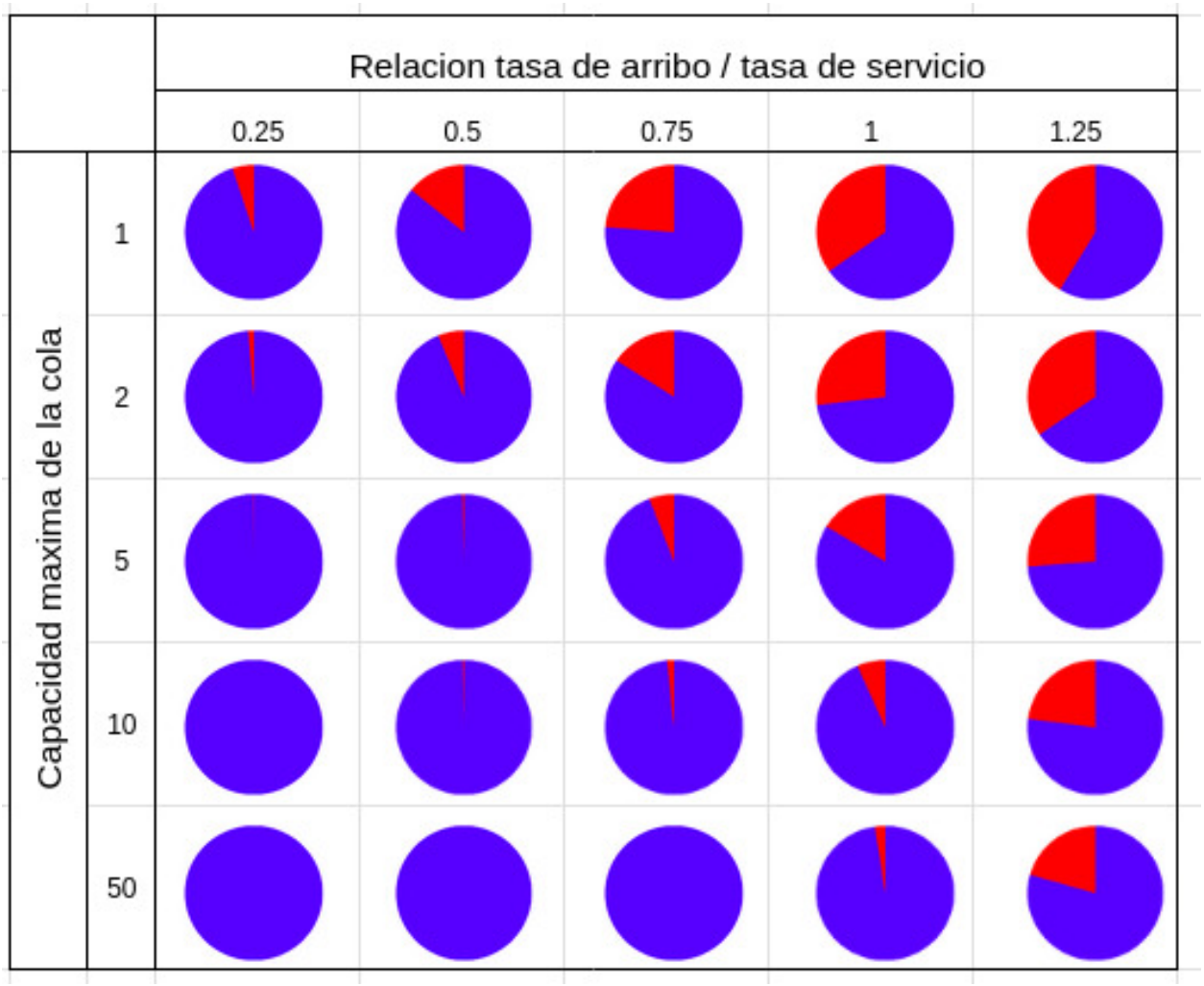

Figure 15: Tasa de Rechazos en base a la capacidad de la cola y la tasa de arribo/servicio

L_q	*Proporcion tasa de arribos a tasa de servicio*				
Tamano cola	25%	50%	75%	100%	125%
1	0.7757	0.5806	0.4161	0.3532	0.2052
2	1.6987	1.3239	1.0086	0.695	0.5591
5	4.6575	3.815	2.8442	1.811	1.3339
10	9.1975	8.2422	5.4558	3.4507	1.5489
50	48.7486	46.0661	40.9259	15.1541	2.4807

Figure 16: Longitud Promedio de Clientes en cola

3.2 Sistema de Inventario

En nuestro sistema de inventario hay establecidos dos niveles: el nivel mínimo (s) y el nivel máximo (S) de inventario. A continuación se presentan los resultados de diferentes corridas realizadas, en cada una de las cuales se optó por una combinación (s, S) diferente.

Para el modelo generado con Python3, se obtuvieron los siguientes resultados para cada política.

Policy	Average total cost	Average ordering cost	Average holding cost	Average shortage cost
(20, 40)	121.52	90.00	10.02	21.50
(20, 60)	140.60	106.44	16.21	17.94
(20, 80)	160.91	120.89	24.94	15.08
(20, 100)	143.99	101.67	35.85	6.47
(40, 60)	150.86	122.00	25.95	2.91
(40, 80)	140.21	101.89	29.40	8.92
(40, 100)	144.18	93.33	45.11	5.74
(60, 80)	156.29	106.56	48.11	1.63
(60, 100)	172.42	120.11	48.30	4.01

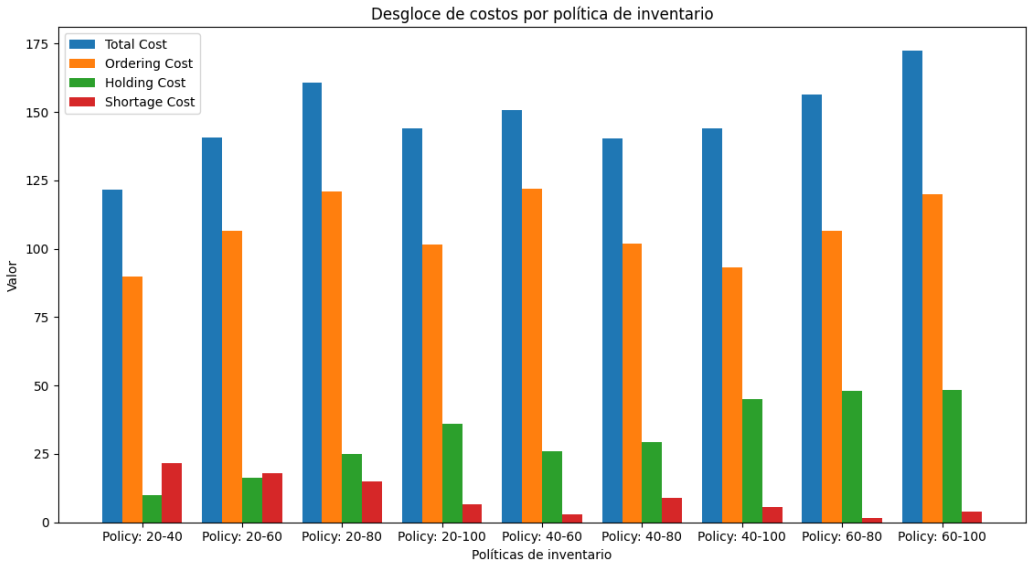

Figure 17: Costos para cada combinación de niveles de inventario en el modelo generado en Python3

A continuación se presentan los costos obtenidos en el modelo de inventario generado en Anylogic

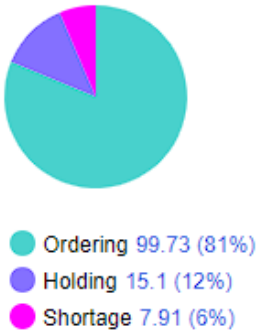

Figure 18: Política (20, 40)

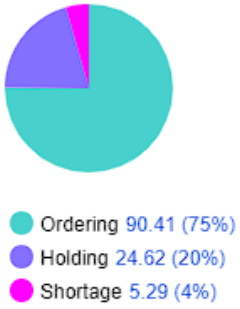

Figure 19: Política (20, 60)

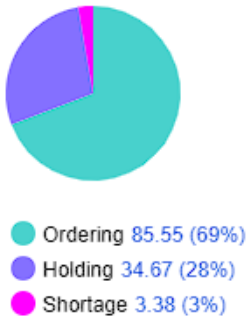

Figure 20: Política (20, 80)

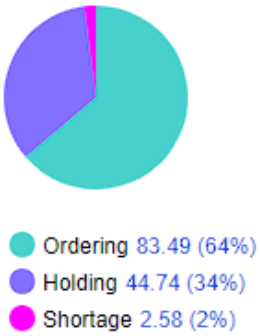

Figure 21: Política (20, 100)

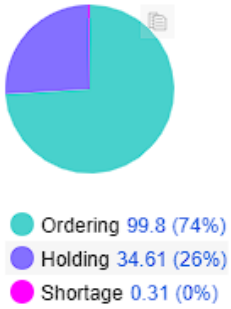

Figure 22: Política (40, 60)

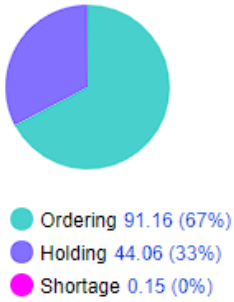

Figure 23: Política (40, 80)

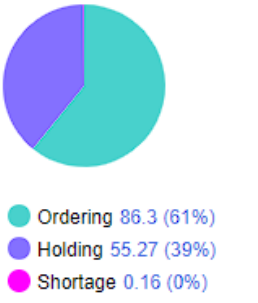

Figure 24: Política (20, 100)

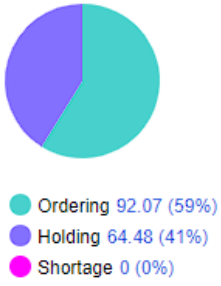

Figure 25: Política (60, 80)

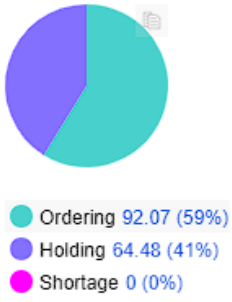

Figure 26: Política (60, 100)

Los resultados anteriores evidencian que el costo de pedido (ordering cost) es el más significativo en todos los casos. Sin embargo, hay una diferencia en la proporción que el mismo abarca dentro del costo total: cuanto mayor sea la diferencia entre el nivel máximo y el nivel mínimo de inventario, menor será el porcentaje que el costo de pedido representa dentro del costo total. Esto se debe a que una mayor discrepancia entre los niveles máximo y mínimo de inventario implica que la capacidad de almacenar productos será también mayor y que, por lo tanto, la necesidad de realizar pedidos a proveedores y afrontar sus gastos será menor.

El costo por mantenimiento de inventario (holding cost) es también sensible a la diferencia entre los niveles de inventario. Dicho costo tiende a aumentar cuando la diferencia el nivel máximo y el nivel mínimo del inventario es mayor, simplemente porque habrá una mayor cantidad de productos almacenados durante períodos prolongados que conllevarán diversos gastos asociados a su mantenimiento.

El costo por faltante (shortage cost), por su parte, varía únicamente en función del nivel inferior del inventario. Cuando el inventario cae por debajo del nivel inferior (s) revela que el inventario es insuficiente y que son probables las situaciones en las que no se pueda satisfacer con la demanda de manera efectiva a causa de falta de stock. Cuanto más alto el nivel inferior de inventario sea establecido, menos posibilidades habrá de no cumplir con las demandas y por lo tanto, el costo debido a faltante de stock será menor.

4 Conclusión

En este informe se detallan las bases y los resultados de dos modelos de simulación realizados: la teoría de colas y el sistema de inventarios.

Para el modelo de teoría de colas se optó por emplear el modelo M/M/1, en el que la llegada de clientes es aleatoria y predecible. De este modelo pueden desprenderse algunas métricas de desempeño como el uso de servidor, número promedio de clientes en cola y en el sistema, el tiempo promedio de espera en el sistema y en la cola, etc.; todas ellas han sido evaluadas en el informe como así también durante la simulación. Los resultados de ésta última indican que el tamaño de la cola tiende a aumentar con el tiempo, aunque con fluctuaciones. Además se observó que existe una asimetría en la distribución de frecuencias del tamaño de la cola, en la que los valores cercanos a cero son más frecuentes y que al aumentar el tamaño de la muestra, la asimetría disminuye y la distribución se asemeja más a una distribución uniforme.

En el caso del sistema de inventarios se tuvieron en cuenta el intervalo entre arribo de los clientes, la demanda que éstos realizan, el tiempo que tarda el proveedor en cumplir con la entrega de pedidos y, principalmente, el nivel máximo (S) y el nivel mínimo (s) de inventario. En el informe se evalúa como cambian los costos para diferentes políticas (s, S).

En resumen, la confección de los modelos significó un profundo análisis de las bases de sistemas de espera y de inventarios. Los resultados obtenidos proporcionan información que puede, y debe, ser considerada durante la toma de decisiones estratégicas en cada uno de sus respectivos campos.

5 Referencias

- https://github.com/Ramiro-DG/Simulacion2023/tree/main/TP3
- https://cloud.anylogic.com/model/caf3e9ce-bd85-4e70-9820-64700c9bcc28?mode=SETTINGS
- https://cloud.anylogic.com/model/8a8a549a-5d67-436a-9b8e-6547126091bd?mode=SETTINGS
- http://web.archive.org/web/20181117064024/https://www.supositorio.com/rcalc/rcalclite.htm
- https://www.anylogic.com/resources/books/the-art-of-process-centric-modeling-with-anylogic/
- https://en.wikipedia.org/wiki/M/M/1$_queue$