
SIMULACIÓN DE UN MODELO M/M/1 E INVENTARIO

Abud Santiago Elias
Legajo 47015
sabudvicco@gmail.com

Castellano Marcelo
Legajo 39028
marce.geek22@gmail.com

Navarro Franco
Legajo 46387
franconavarro1889@gmail.com

22 de julio de 2022

ABSTRACT

En este trabajo estudiaremos dos modelos de simulación de eventos discretos, el primero es el comportamiento de líneas de espera. El cual es de gran ayuda para predecir el comportamiento de dichas líneas en situaciones del mundo real, desde la entrada y salida de autos de un estacionamiento hasta la utilización de una red distribuida de servidores a lo largo del mundo que alojan una página web para miles de usuarios. El siguiente será un modelo de inventario que nos permitirá saber los costos correspondientes al mantenimiento o de compras necesarias y será de gran ayuda para un inventario en la vida real

1. Introducción

En el siguiente trabajo simularemos mediante un programa desarrollado en lenguaje Python, una cola simple, en la que solamente habrá llegada de clientes, atendidos por un único servidor luego de haber realizado una espera determinada en la cola y posteriormente partirán. Adicionalmente se simulará una cola con una cantidad de servidores predeterminada. El modelo de inventario contará con el mismo formato y consistirá en un proceso por mes en el que se realiza una orden de compra a los proveedores, mientras se trata de alcanzar la demanda de los clientes a través de distintos indicadores de escasez y existencia. Finalmente se realizará una evaluación y se calcularán las estadísticas correspondientes. A su vez estudiaremos la eficacia de dichos programas al comparar ciertas medidas de rendimiento observadas con las teóricas computadas y a su vez con las obtenidas mediante una calculadora web, y de otra simulación desarrollada en el aplicativo Anylogic.

2. Marco teórico

2.1. Proceso Estocástico

En la teoría de la probabilidad, un proceso estocástico es un concepto matemático que sirve para representar magnitudes aleatorias que varían con el tiempo o para caracterizar una sucesión de variables aleatorias (estocásticas) que evolucionan en función de otra variable, generalmente el tiempo. Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y pueden o no estar correlacionadas entre sí.

Cada variable o conjunto de variables sometidas a influencias o efectos aleatorios constituye un proceso estocástico. Un proceso estocástico X_t puede entenderse como una familia uniparamétrica de variables aleatorias indexadas mediante el tiempo t . Los procesos estocásticos permiten tratar procesos dinámicos en los que hay cierta aleatoriedad.

3. Modelo M/M/1

Un sistema de espera M/M/1 es aquel que considera un servidor, con tiempos exponenciales de servicio y entre llegadas de clientes. La implicancia que los tiempos de servicio se distribuyan exponencial es que existe una preponderancia de tiempos de servicio menores al promedio combinados con algunos pocos tiempos extensos. Un ejemplo de ello es lo que sucede en las cajas de los bancos donde la mayoría de las transacciones requieren poco tiempo de proceso por

parte del cajero, no obstante algunas transacciones más complejas consumen bastante tiempo. Por otra parte afirmar que los tiempos entre llegadas se distribuyen exponencial implica una preponderancia de tiempos entre llegadas menores que el promedio en combinación con algunos tiempos más extensos. Lo anterior tiene relación con la aleatoriedad del proceso de llegada de clientes que permite establecer la Propiedad de Falta de Memoria o Amnesia de la Distribución Exponencial y con los conceptos presentados en el artículo Qué son las Líneas de Espera (Teoría de Colas), donde queda en evidencia que la formación de las colas o filas esta asociada a la variabilidad del sistema.

En este contexto consideremos la siguiente notación, donde valores usuales para A y B son M (distribución exponencial) y G (distribución general).

En el estudio con colas simples, cuya notación de Kendalls son M/M/1, para la primera, donde se posee un sólo servidor, las distribuciones de arribo y de servicio proceden de manera Markoviana y su cola es infinita, y en el caso de la siguientes M/M/C dónde solamente se añadiran mas servidores. Para la distribución de arribos, se tiene un parámetro, la tasa de arribos de clientes, de él obtenemos la media de clientes entrantes por unidad de tiempo $1/\lambda$. Por otro lado, la tasa de tiempo de servicios μ y la media de tiempos de servicios $1/\mu$. A partir de estos dos parámetros, obtenemos la tasa de utilización del servidor $= 1/\lambda\mu$, la relación entre los clientes arribando y los clientes partiendo, siendo s el número de servidores, en nuestro caso $s = 1$ (luego se permitira variarlo). Como la llegada de clientes y su tiempo en ser servidos son variables aleatorias, se pueden considerar nuevas variables aleatorias, que son combinaciones lineales de las anteriores, por ejemplo:

Considere un sistema de colas de un solo servidor para el cual el intervalo entre llegadas son A_1, A_2, \dots veces con variables aleatorias independientes e idénticamente distribuidas.

A partir de una única ejecución de la simulación con retrasos de los clientes D_1, D_2, \dots, D_n , es obvio que el estimador de $d(n)$ es:

$$\hat{d}(n) = \frac{\sum_{i=1}^n D_i}{n} \quad (1)$$

que es solo el promedio de los D_i 's que se observaron en la simulación.

Una de las medidas para nuestro modelo simple es el número promedio esperado de clientes en la cola (pero sin ser servida), denotada por $q(n)$, donde la n es necesaria en la notación para indicar que este promedio se toma durante el período de tiempo necesario para observar los n retrasos que definen nuestra regla de parada. Este es un tipo diferente de "promedio," comparación del retraso promedio en cola, porque se toma el tiempo (continuo), en lugar de los clientes (siendo este discreto). Por lo tanto, necesitamos definir qué significa este número promedio de tiempo de clientes en cola. Para hacer esto, sea $Q(t)$ el número de clientes en cola en el tiempo t, para cualquier número real $t \geq 0$, y sea $T(n)$ el tiempo requerido para observar nuestros n retrasos en la cola. Entonces para cualquier tiempo t entre 0 y $T(n)$, $Q(t)$ es un entero no negativo. Además, si dejamos que p_i sea la proporción esperada (que estará entre 0 y 1) del tiempo que $Q(t)$ es igual a i, entonces una definición razonable de $q(n)$ sería:

$$q(n) = \sum_{i=0}^{\infty} i p_i \quad (2)$$

Por lo tanto, $q(n)$ es un promedio ponderado de los posibles valores de i para la longitud de la cola $Q(t)$, siendo las áreas la proporción esperada de tiempo que la cola pasa en cada uno de sus posibles longitudes. Para estimar $q(n)$ a partir de una simulación, simplemente reemplazamos los p_i con estimaciones de ellos y obtenemos:

$$\hat{q}(n) = \sum_{i=0}^{\infty} i \hat{p}_i \quad (3)$$

donde i es la proporción observada (en lugar de la esperada) del tiempo durante la simulación donde que había i clientes en la cola. Computacionalmente, sin embargo, es más fácil reescribir (n) usando algunas consideraciones geométricas. Si dejamos que T_i sea el total de tiempo durante la simulación donde la cola tiene una longitud i, entonces $T(n) = T_0 + T_1 + T_2 + \dots$ y $i = T_i/T(n)$, por lo que podemos reescribir la ecuación 1.1 como:

$$\hat{q}(n) = \frac{\sum_{i=0}^{\infty} i T_i}{T(n)} \quad (4)$$

El numerador de la ecuación anterior, representa el área bajo la función $Q(t)$:

$$\sum_{i=0}^{k-1} i T_i = \int_0^{T(n)} Q(t) dt \quad (5)$$

Reemplazándolo en la ecuación anterior, nos queda:

$$\hat{q}(n) = \frac{\int_0^{T(n)} Q(t) dt}{T(n)} \quad (6)$$

Esta integral puede ser calculada como la suma de rectángulos formados por la base tiempo de cierta cantidad de clientes en cola por la altura dicha cantidad de clientes en cola.

La proporción esperada de tiempo del servidor en estado ocupado $u(n)$, deviene de la probabilidad de que el servidor no esté vacío, $p_N > 0 = 1 - p_0$. Para calcular su estimador $\hat{u}(n)$, primero se define la "función ocupada":

$$B(t) = \begin{cases} 1 & \text{si } Nt > 0 \\ 0 & \text{si } Nt = 0 \end{cases} \quad (7)$$

Entonces $u(n)$ es la porción del tiempo total en la que $B(t) = 1$. Al igual que a la medida anterior, podemos considerarlo como el área bajo $B(t)$, así,

$$\hat{u}(n) = \frac{\int_0^{T(n)} B(t) dt}{T(n)} \quad (8)$$

En definitiva, $u(n)$ es la sumatoria de áreas rectangulares, donde la base es el tiempo en el que el servidor está en un estado específico, y la altura es dicho estado, 0 o 1.

4. Modelo de Inventario

Los inventarios están presentes en todas las compañías que tratan con productos físicos, tales como fabricantes, distribuidores, comerciantes, etc. Las empresas necesitan inventarios de materias primas para la manufactura de productos y a su vez deben almacenar productos terminados en el almacén a la espera de ser vendidos. De manera similar, los distribuidores deben mantener inventarios de bienes que deberán estar disponibles cuando los consumidores los necesiten. Dado que los stocks representan una cantidad de dinero inmovilizada muy importante dentro de una empresa, la reducción de los costos de almacenamiento (evitando inventarios innecesariamente grandes) podría mejorar la competitividad de cualquier sistema productivo.

Cuando hay que analizar los inventarios con una demanda independiente los modelos de gestión de stocks que se utilizan son: el modelo de cantidad fija del pedido (EOQ) y el modelo de periodo de tiempo fijo (también llamado de revisión periódica, modelo P) [1-2]. En el modelo de cantidad fija de pedido se coloca un pedido cuando el inventario restante cae a un punto de pedido y se revisa el nivel de inventario continuamente. De esta manera, el modelo de cantidad fija de pedido es un sistema perpetuo que requiere que cada vez que se haga un retiro o una adición al inventario, los registros deban actualizarse para asegurar que el punto del nuevo pedido se ha alcanzado o no. En cambio en el modelo de periodo de tiempo fijo, el conteo tiene lugar solo durante el periodo de revisión.

Suponiendo una empresa que vende un solo producto le gustaría decidir cuántos artículos debería tener en inventario para cada uno de los siguientes n meses (n es un parámetro de entrada fijo). Los tiempos entre demandas son variables aleatorias exponenciales con una media de 0,1 mes. Los tamaños de las demandas, D , son variables aleatorias IID (independientes de cuando se presentan las demandas), con:

$$D = \begin{cases} 1 & \text{c.p. } 1/6 \\ 2 & \text{c.p. } 1/3 \\ 3 & \text{c.p. } 1/3 \\ 4 & \text{c.p. } 1/6 \end{cases} \quad (9)$$

donde c.p. significa "Con Probabilidad de".

Al comienzo de cada mes, la empresa revisa el nivel de inventario y decide cuántos artículos pedir a su proveedor. Si la empresa ordena Z artículos, incurre en un costo de $K + iZ$, donde $K = \$32$ es el costo de preparación e $i = \$3$ es el incremento de costo incremental por artículo pedido. (Si $Z = 0$, no incurre ningún costo). Cuando se realiza un pedido, el tiempo requerido para que llegue (llamado retraso en la entrega o tiempo de entrega) es una variable aleatoria que se distribuye uniformemente entre 0,5 y 1 mes.

La empresa utiliza una política estacionaria (s, S) para decidir cuánto pedir, es decir,

$$Z = \begin{cases} S - I & \text{if } I < s \\ 0 & \text{if } I \geq s \end{cases} \quad (10)$$

donde I es el nivel de inventario al comienzo del mes.

Para nuestro modelo, supondremos que la empresa incurre en un costo de mantenimiento de $h = \$1$ por artículo por mes mantenido en el inventario (positivo). El costo de mantenimiento incluye costos tales como alquiler de almacenes, seguros, impuestos y mantenimiento, así como el costo de oportunidad de tener capital inmovilizado en el inventario en lugar de invertirlo en otra parte. Hemos ignorado en nuestra formulación el hecho de que todavía se incurre en algunos costos de mantenimiento cuando $I = 0$. Sin embargo, dado que nuestro objetivo es comparar las políticas de pedido, ignorar este factor, que después de todo es independiente de la política utilizada, no afectará nuestra evaluación de qué política es la mejor. Ahora, dado que I es el número de elementos retenidos en el inventario en el momento t , la cantidad promedio de tiempo (por mes) de artículos mantenidos en el inventario durante el período de n meses es

$$\bar{I}^+ = \frac{\int_0^n I^+(t) dt}{n} \quad (11)$$

De manera similar, suponga que la compañía incurre en un costo de trabajo pendiente de $p = \$5$ por artículo por mes en reserva; esto representa el costo del mantenimiento adicional de activos cuando existe acumulación de pedidos, así como pérdida de la buena voluntad de los clientes. El número promedio de tiempo de los elementos en la reserva es:

$$\bar{I}^- = \frac{\int_0^n I^-(t) dt}{n} \quad (12)$$

por lo que el costo promedio de la cartera de pedidos por mes es πI .

5. Análisis de resultados

5.1. Modelo MM1 en AnyLogic

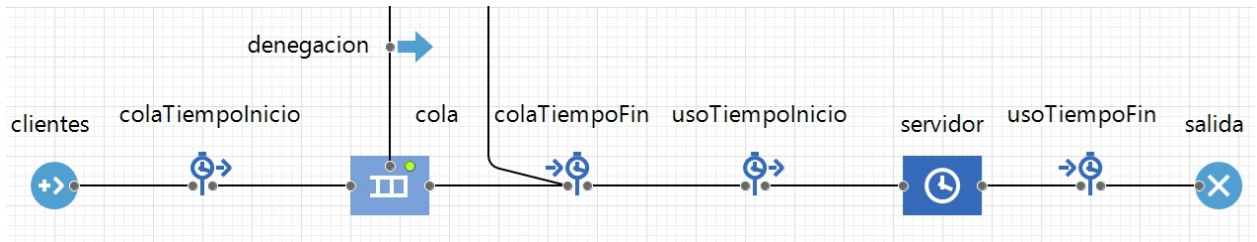


Figura 1: Bloques del modelo MM1 en Anylogic.

Para analizar el rendimiento del modelo, variamos la tasa de arribos (T_a) en base a la tasa de servicio (T_s), y la capacidad de la cola (cap).

Realizaremos 10 simulaciones de 1000 clientes cada una y promediamos los siguientes estadísticos:

- Promedio de clientes en el sistema ($q(n) + u(n)$)
- Cantidad de clientes en cola en promedio ($q(n)$)
- Demora promedio esperada en cola ($d(n)$)
- Tiempo promedio en el sistema ($d(n) + s(n)$)
- Ocupación del servidor ($u(n) * 100\%$)
- Probabilidad de denegación del servicio. ($p(den)$)
- Probabilidad de encontrar n clientes en cola. ($p(Q(t) = n) \times 100\%$)

Los primeros 5 fueron tabulados y el último fue graficado.

5.1.1. $cap = 0$

Los parámetros $d(n)$, $q(n)$ y $p(Q(t) = n) \times 100\%$ no aplican cuando la capacidad de la cola es 0. El promedio de clientes en sistema es en este caso, equivalente al factor de uso del servicio.

$\frac{T_a}{T_s} \times 100\% [\%]$	$d(n) + s(n) [\text{min}]$	$u(n) \times 100\% [\%]$	$p(den) [\%]$
25	2,005	19,823	19,46
50	2,059	33,656	34,09
75	1,93	42,212	41,65
100	1,985	49,165	50,12
125	2,004	55,434	55,15

5.1.2. $cap = 2$

$\frac{T_a}{T_s} \times 100\% [\%]$	$q(n) + u(n) [\text{min}]$	$q(n) [\text{clientes}]$	$d(n) [\text{min}]$	$d(n) + s(n) [\text{min}]$	$u(n) \times 100\% [\%]$	$p(den) [\%]$
25	0,313	0,068	0,548	2,54	24,546	1,26
50	0,739	0,266	1,148	3,186	47,27	6,39
75	1,168	0,524	1,65	3,681	64,373	15,71
100	1,499	0,748	2,014	4,039	75,164	25,3
125	1,756	0,934	2,26	4,246	82,168	32,93

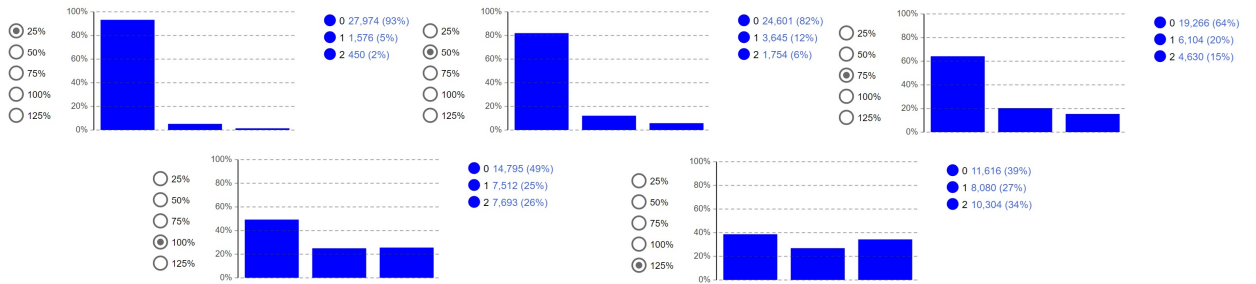


Figura 2: Probabilidad de encontrar n clientes en cola.

5.1.3. $cap = 5$

$\frac{T_a}{T_s} \times 100\% [\%]$	$q(n) + u(n) [\text{min}]$	$q(n) [\text{clientes}]$	$d(n) [\text{min}]$	$d(n) + s(n) [\text{min}]$	$u(n) \times 100\% [\%]$	$p(den) [\%]$
25	0,326	0,079	0,634	2,069	24,672	0,01
50	0,917	0,424	1,7	3,681	49,263	0,85
75	1,882	1,18	3,318	5,289	70,119	4,9
100	3,122	2,253	5,234	7,254	86,914	14,19
125	3,847	2,917	6,267	8,266	93,034	25,23

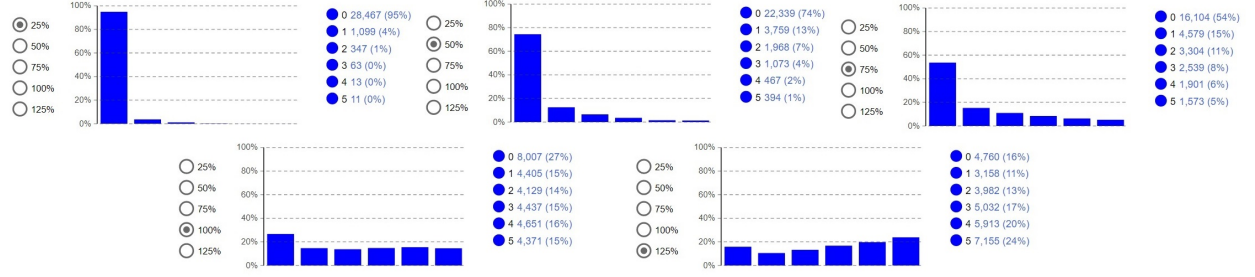


Figura 3: Probabilidad de encontrar n clientes en cola.

5.1.4. $cap = 10$

$\frac{T_a}{T_s} \times 100\% [\%]$	$q(n) + u(n) [\text{min}]$	$q(n) [\text{clientes}]$	$d(n) [\text{min}]$	$d(n) + s(n) [\text{min}]$	$u(n) \times 100\% [\%]$	$p(den) [\%]$
25	0,325	0,08	0,648	2,631	24,52	0
50	1,032	0,534	2,15	4,159	49,816	0,08
75	2,314	1,593	4,351	6,332	72,183	0,64
100	5,294	4,382	9,62	11,623	91,217	8,13
125	7,582	6,606	13,329	15,296	97,504	19,65

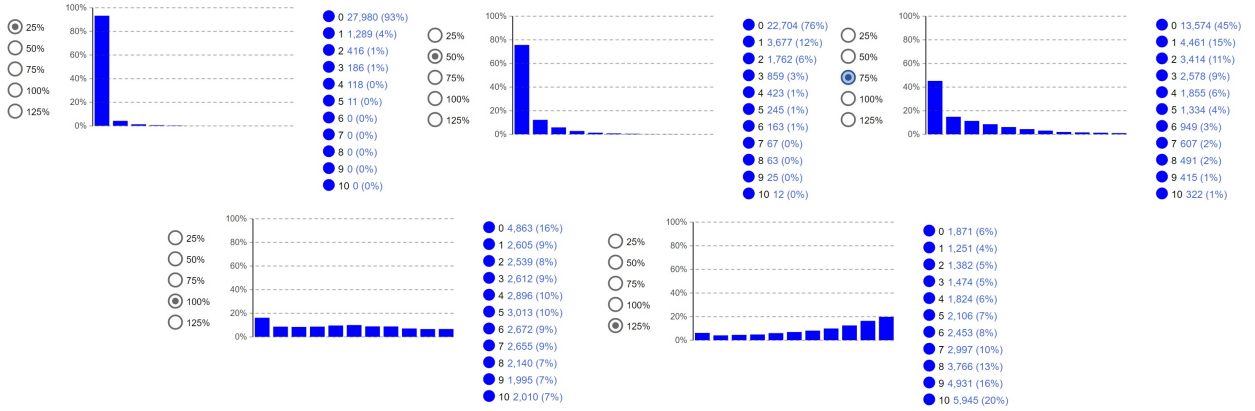


Figura 4: Probabilidad de encontrar n clientes en cola.

5.1.5. $cap = 50$

$\frac{T_a}{T_s} \times 100\% [\%]$	$q(n) + u(n) [\text{min}]$	$q(n) [\text{clientes}]$	$d(n) [\text{min}]$	$d(n) + s(n) [\text{min}]$	$u(n) \times 100\% [\%]$	$p(den) [\%]$
25	0,335	0,085	0,679	2,686	25,007	0
50	1,039	0,532	2,099	4,11	50,7	0
75	3,116	2,368	6,341	8,358	74,8	0
100	22,932	21,957	45,615	47,631	97,42	0,8
125	39,365	38,37	75,431	77,386	99,539	13,41

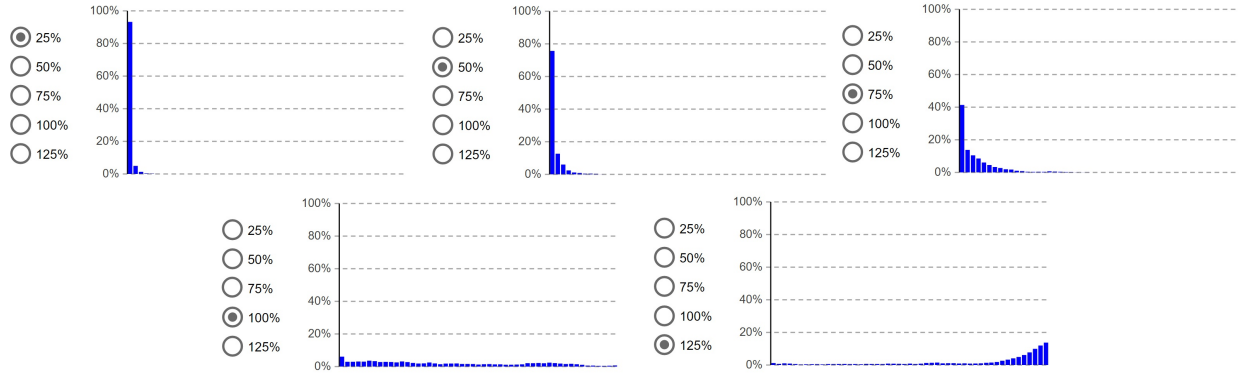


Figura 5: Probabilidad de encontrar n clientes en cola.

5.2. Modelo de Inventario en AnyLogic

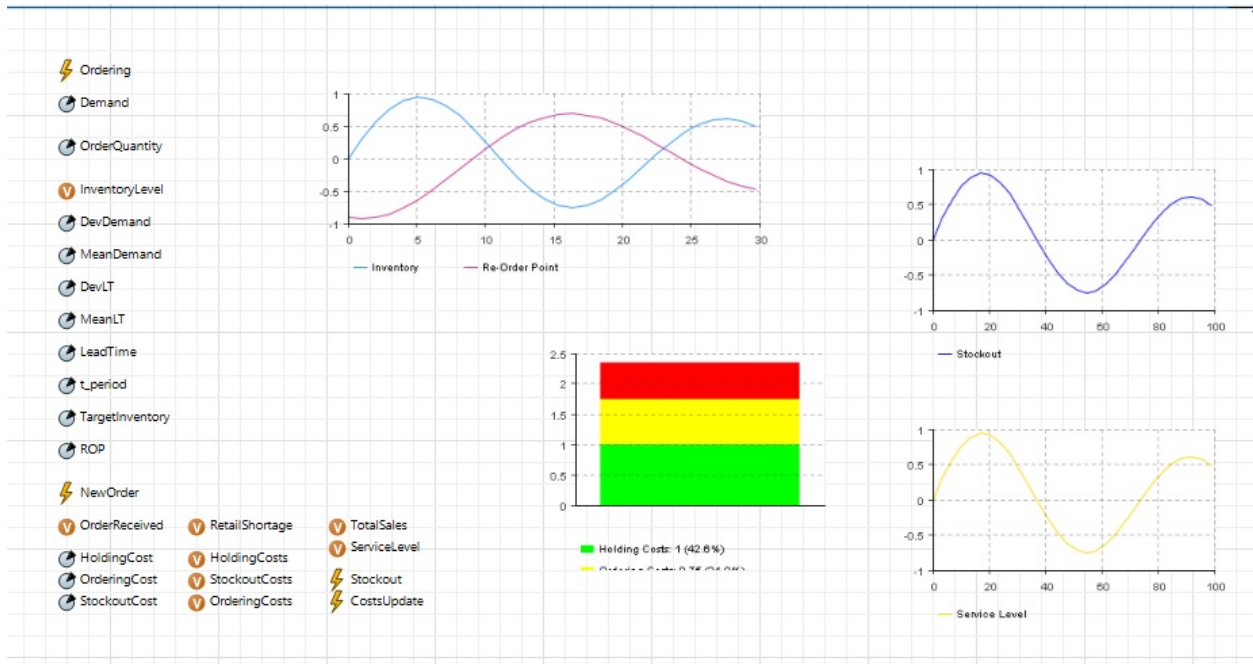


Figura 6: Modelo de Inventario con sus parámetros, variables y eventos utilizados en Anylogic.

Como podemos ver, hay ciertos parámetros que ya hemos utilizados en el modelo realizado en Python, pero fué necesario además agregar nuevos para poder utilizarlos con distintas funcionalidades.

Comenzando por el evento “Ordering” el cual nos permite programar los reabastecimientos del inventario a través de dos parámetros: (demand y order quantity). Además agregamos una variable llamada inventoryLevel la cual analizaremos con una línea de tiempo.

Este es un evento cíclico con un tiempo de recurrencia de un día.

Suponiendo que la demanda diaria y/o el tiempo de entrega se distribuyen normalmente. Necesitamos definir más parámetros para el tiempo de entrega, la demanda media, el tiempo de entrega promedio, la desviación estándar de la demanda y la desviación estándar del tiempo de entrega.

Para el sistema de revisión periódica, reabastecemos en fechas fijas, por lo que solo modificamos el tiempo de recurrencia del evento “Ordering” definiéndolo a través del parámetro “t_period”.

Para el sistema de revisión continua, la regla de reabastecimiento se puede definir con la ayuda del punto de pedido (ROP).

Además incluiremos stock de seguridad en nuestro modelo. Para modelar el tiempo de entrega, definimos el evento “NewOrder” y la variable “OrderReceived” de tipo “Booleano” (con valor inicial “false”) y vuelva a escribir la regla de reposición en el evento “Ordering”.

Si el nivel de inventario alcanza el punto de pedido, se genera un nuevo pedido de reabastecimiento el cual llegará en X días definidos en el parámetro LeadTime. Es por eso que dejamos que el evento “NewOrder” comience en X días. En el momento de la entrada del pedido y el aumento del nivel de inventario, la variable lógica “OrderReceived” toma el valor “false”. Esto significa que no hay otras órdenes recibidas por nuestro proveedor. El siguiente pedido se generará cuando el nivel de inventario vuelva a alcanzar el punto de pedido.

Para estimar la eficiencia de las diferentes políticas de pedidos, necesitamos estimar los costos de mantenimiento de inventario, pedidos y agotamiento de existencias. Primero, se necesitan nuevos parámetros y variables para estos tres tipos de costos. En segundo lugar, los diagramas deben construirse como gráficos de barras y diagramas de tiempo. Definimos tres nuevos parámetros “HoldingCost”, “OrderingCost” y “StockoutCosts” y tres nuevas variables de los mismos. Es necesario definir otras tres nuevas variables, “ServiceLevel”, “TotalSales” y “RetailShortage”.

Los costos, la escasez y el nivel de servicio se actualizan en el evento “CostsUpdate”.

A continuación se realizarán 10 simulaciones de 30 días cada una, donde se llevarán a cabo las mediciones de las siguientes variables:

- Costo de Orden
- Costo de Mantenimiento
- Costo de Faltante
- Costo Total

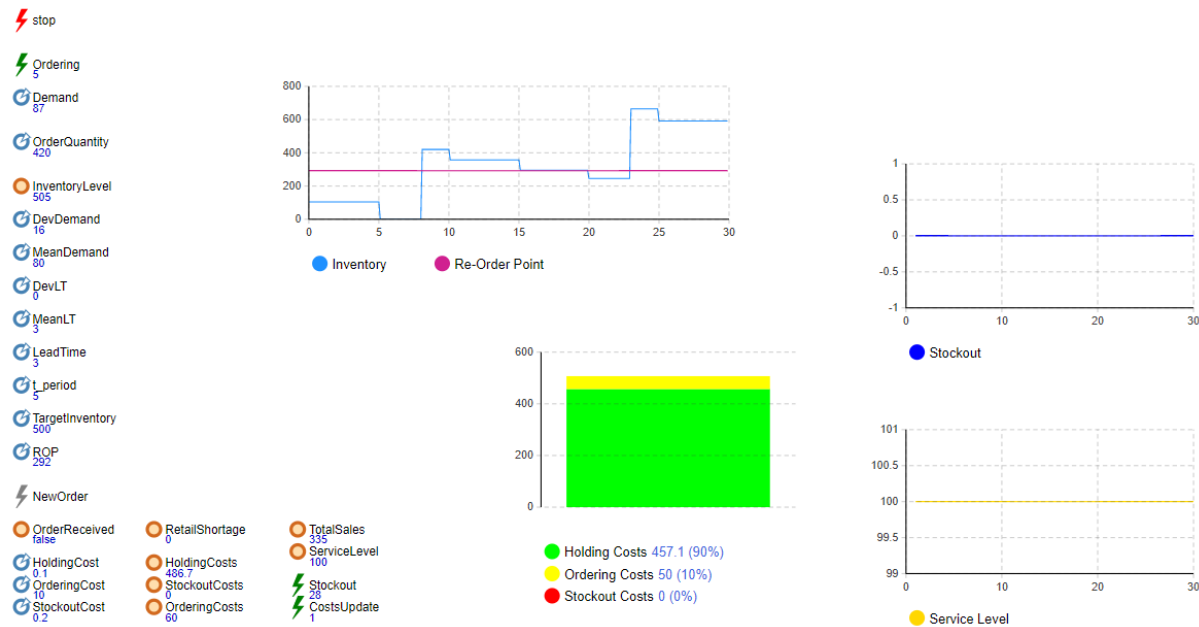


Figura 7: Simulación 1. Orden de compra alto.

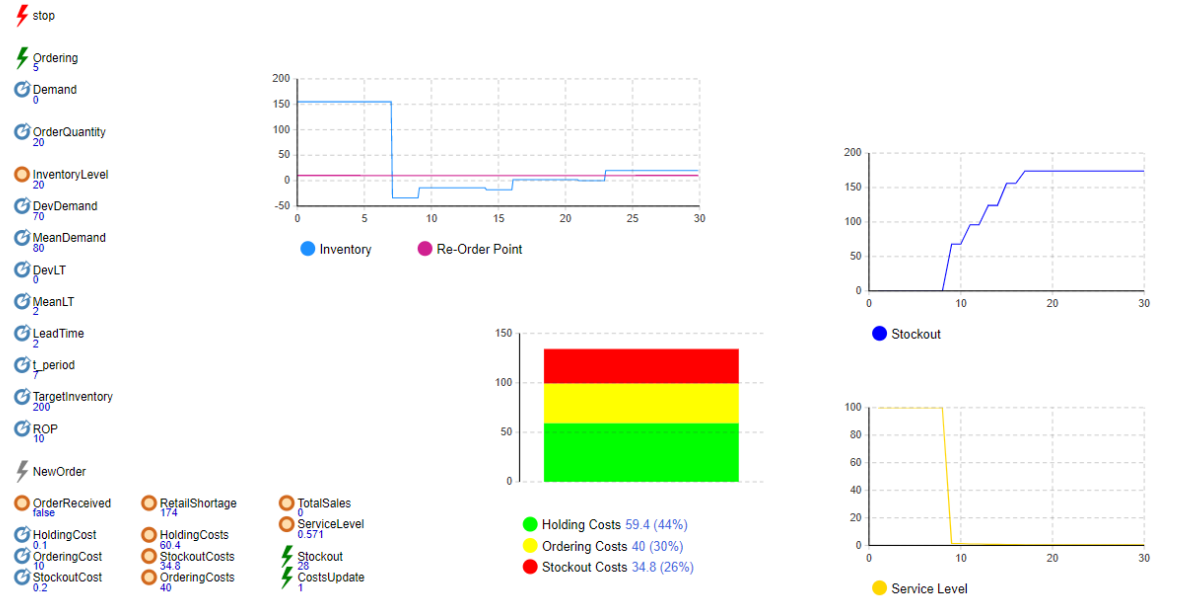


Figura 8: Simulación 2. Orden de compra y ROP bajo.

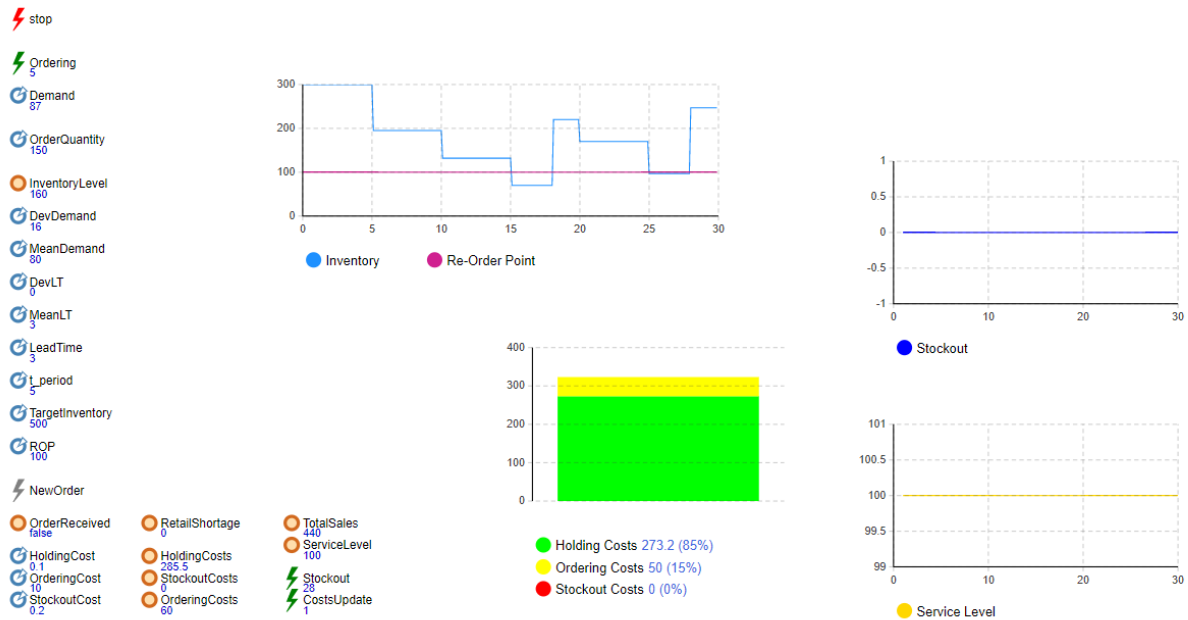


Figura 9: Simulación 3. Cantidad inicial alta y ROP a un tercio de la cantidad inicial.

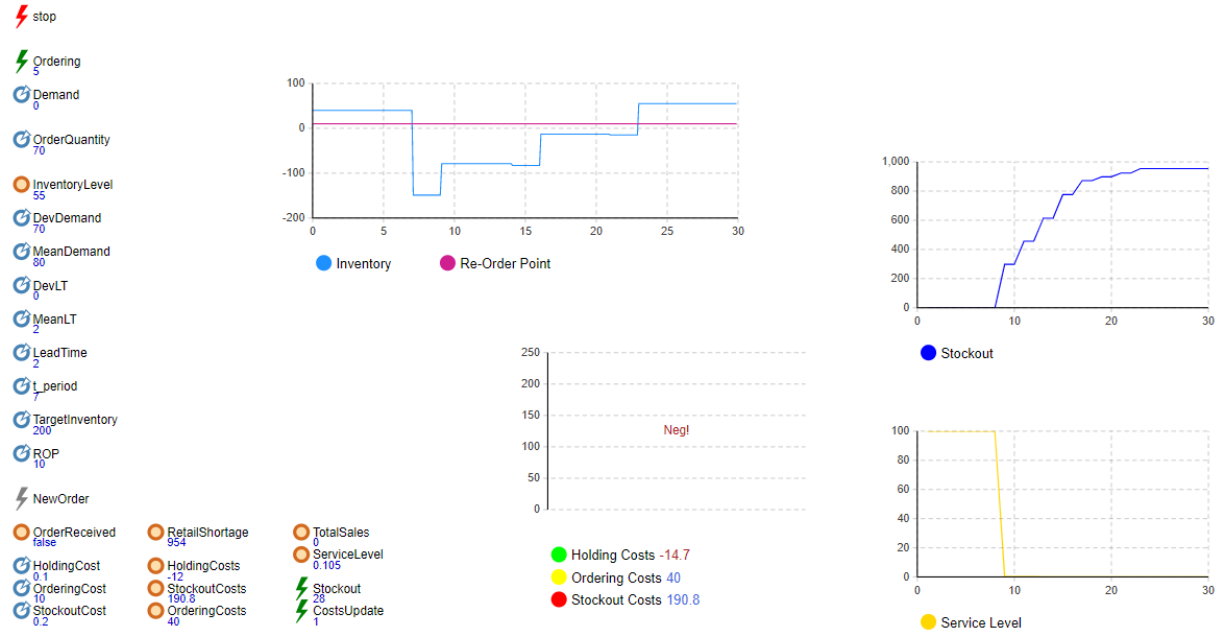


Figura 10: Simulación 4. Gran falta de stock.

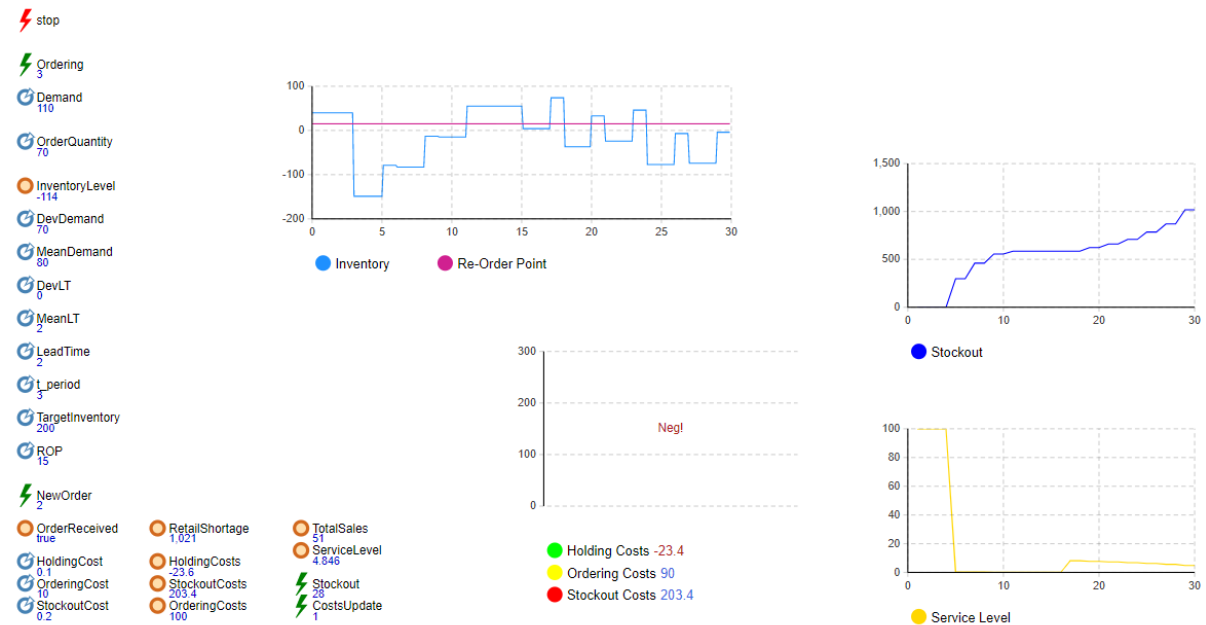


Figura 11: Simulación 5. Gran falta de stock con mayor velocidad de recepción.

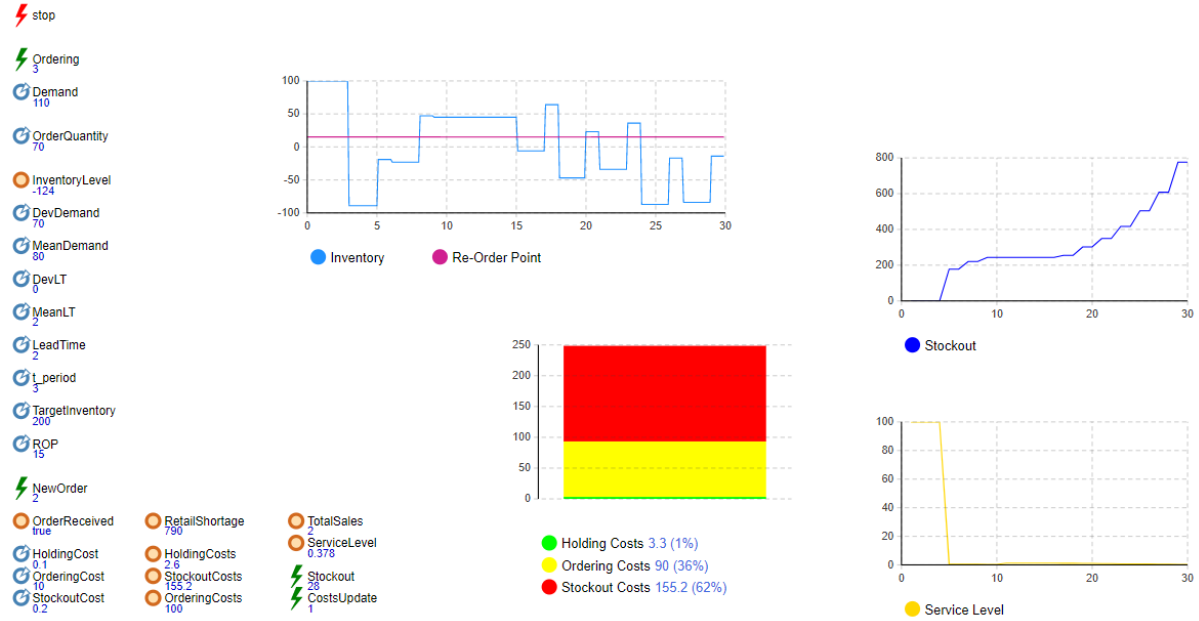


Figura 12: Simulación 6. Mismo Escenario con mayor cantidad de demanda.

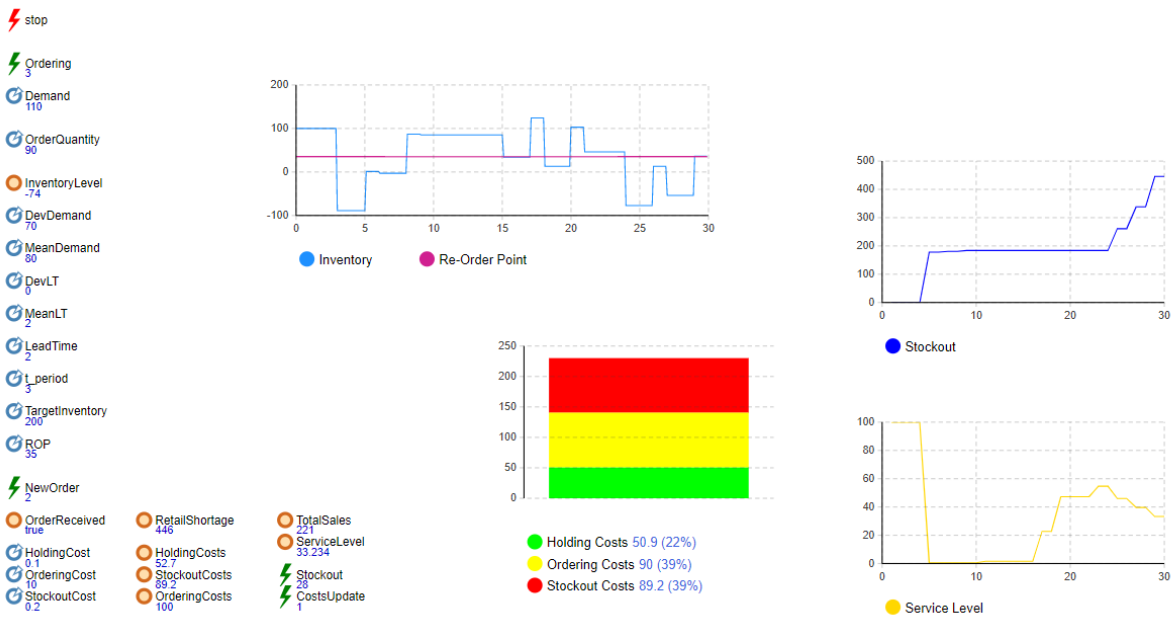


Figura 13: Simulación 7. Mismo Escenario con mayor cantidad de orden de compra.

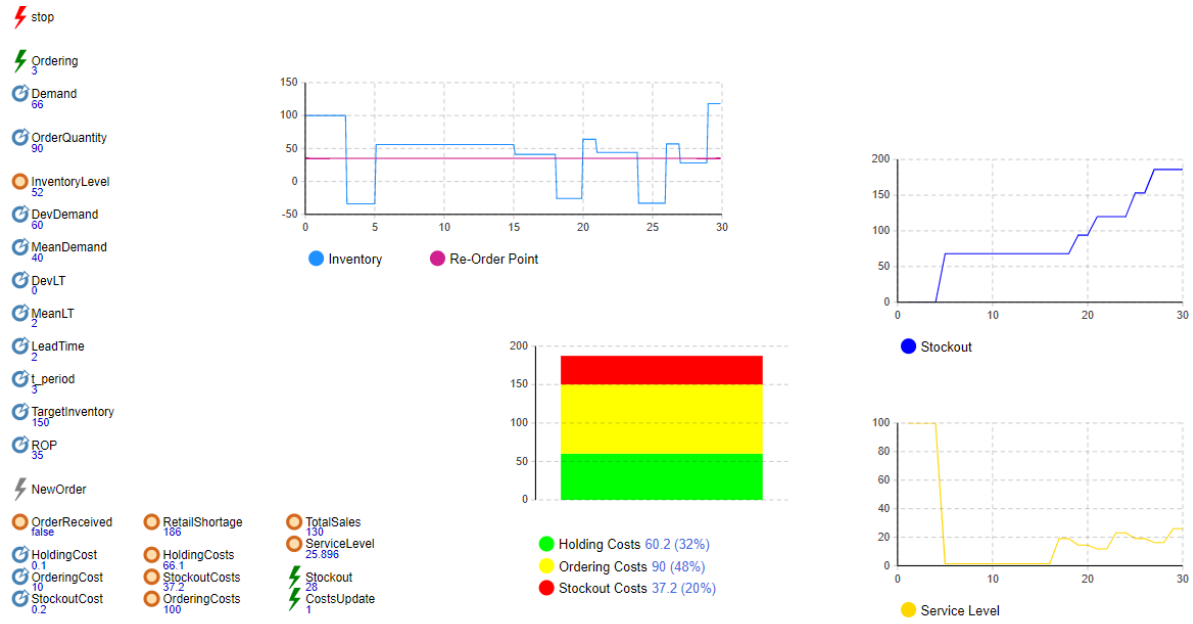


Figura 14: Simulación 8. Baja demanda.

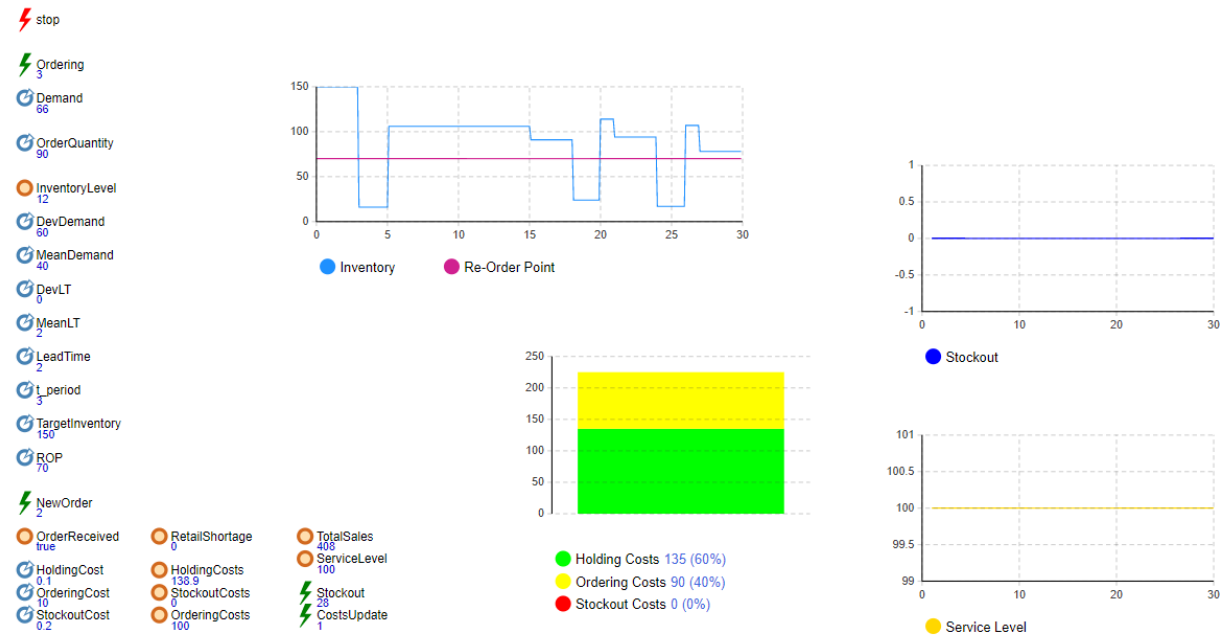


Figura 15: Simulación 9. Baja demanda y aumento de ROP.

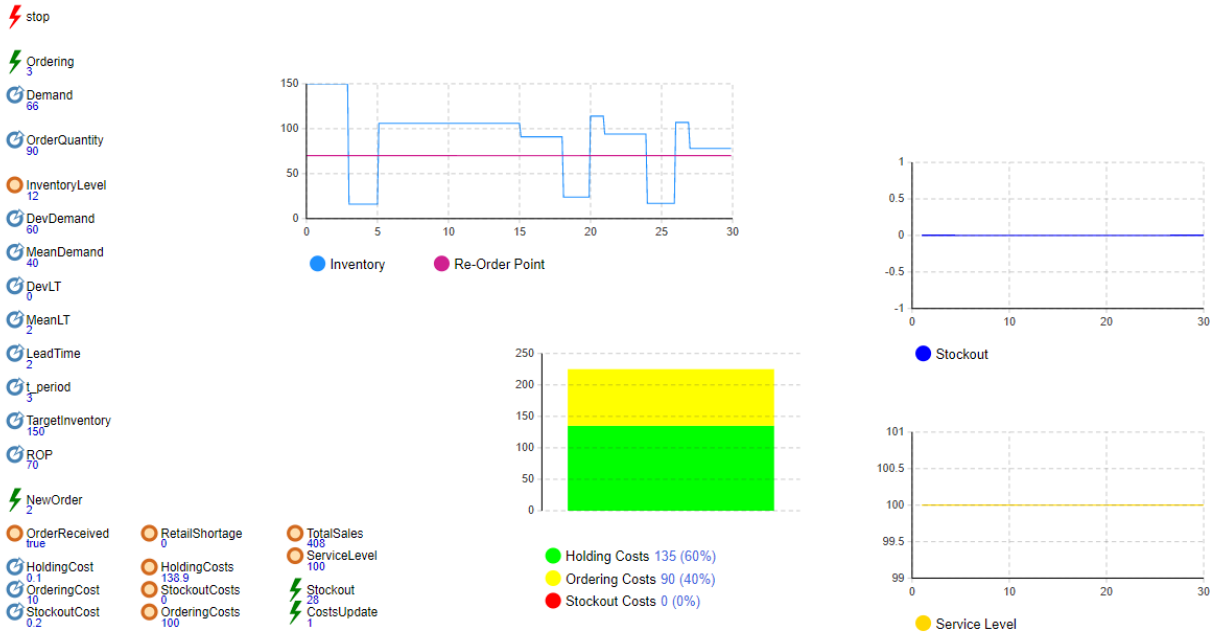


Figura 16: Simulación 10. Disminución de orden de compra.

Simulación nro	Orden	Mantenimiento	Faltante	TOTAL
1	60	486,7	0	546,7
2	40	34,8	60,4	135,2
3	60	285,5	0	345,5
4	40	0	190,8	230,8
5	100	0	204,4	304,4
6	100	2,6	155,2	257,8
7	100	52,7	389,2	541,9
8	100	66,1	37,2	203,3
9	100	138,9	0	238,9
10	100	124,4	0	224,4

5.2.1. Conclusión de las simulaciones

Cabe aclarar que los resultados negativos no significan que se “ahorró” el costo, esto cuenta como 0 debido a que no se pudo realizar el costo de mantenimiento del stock ya que hubo faltante del mismo. Además se puede observar como en la simulación 2 fué la que más pudo mantener bajos costos a comparación de las demás debido a que tuvo pocas ordenes de compra y una gran existencia al principio.

La Simulación 8 fué la siguiente con menos costo debido a que si bien tuvo muchos gastos en las ordenes de compra, pudo distribuir mejor la faltante y el mantenimiento de la existencia en el inventario.

A su vez la Simulación 1 fué la más costosa, debido a que mantuvo una gran cantidad de stock en el inventario mientras que la segunda mas costosa (la simulación 7) tuvo el mismo problema con la diferencia de que esta que tuvo mucha faltante.

Como conclusión se puede decir que el mayor problema radica en la falta de equilibrio en los costos de cada acción, ya que por ejemplo se pueden tener altos costos de orden de compra, pero esto no fué problema para la simulación 8 que distribuyó la faltante y el mantenimiento de existencia de manera que pudo reducir los costos. De la misma manera, hubo simulaciones con pocas ordenes de compra pero con altos costes de faltante o de mantenimiento, lo que llevo a grandes costos totales.

6. Conclusiones

7. Referencias

https://es.wikipedia.org/wiki/Proceso_estoc

Arash Mahdavi, Simulation Modeling Consultant, The AnyLogic Company M/M/1// (single-server queues). En The Art of Process-Centric Modeling with AnyLogic, pages 81–103. 2020. Disponible en <https://www.anylogic.com/resources/books/the-art-of-process-centric-modeling-with-anylogic/>

Averill M. Law, W. David. Kelton Simulation of a single-server queueing system. En Simulation Modeling Analysis, pages 13–74. 199