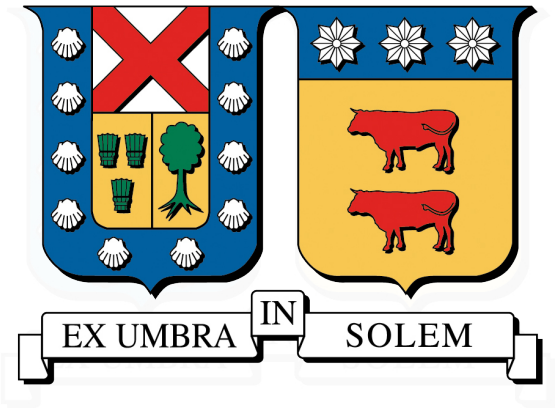
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
VALPARAÍSO - CHILE



A BRANCH & CUT ALGORITHM TO SOLVE THE JOINT LOCATION INVENTORY PROBLEM UNDER FILL-RATE SERVICE LEVEL CONSTRAINTS AND A FULL BACKORDERS APPROACH

CRISTÓBAL FERNANDO MORA QUIJADA

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA INDUSTRIAL
Y AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

PROFESOR GUÍA	:	DR. PABLO ESCALONA RODRÍGUEZ
PROFESOR CORREF. INTERNO	:	DR. RODRIGO MENA BUSTOS
PROFESOR CORREF. EXTERNO	:	DR. ALEJANDRO ANGULO CÁRDENAS
PROFESOR CORREF. EXTERNO	:	DR. FRANCISCO TAPIA UBEDA

MAYO, 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☒ Tesis de Postgrado

Título del trabajo: **A Branch & Cut algorithm to solve the joint location inventory problem under fill-rate service level constraints and a full backorders approach.**

Nombre del candidato(a): **Cristóbal Fernando Mora Quijada**

Carrera / Grado: **Ingeniería Civil Industrial y Magíster en Ciencias de la Ingeniería Industrial**

Campus: **Casa Central** **Departamento:** **Departamento de Industrial**

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, **Pablo Felipe Escalona Rodríguez** en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

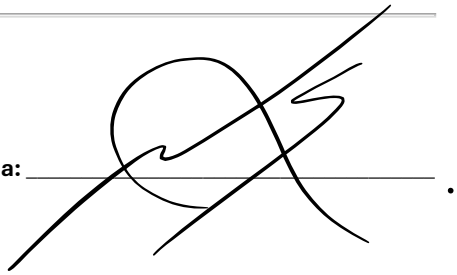
_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: **28 de Mayo, 2026** **Firma:** _____



Estudiante o Candidato(a):

Fecha: **28 de Mayo, 2026** **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

RESUMEN EJECUTIVO

Esta tesis aborda el problema conjunto de localización e inventario para el diseño de una red de distribución de dos niveles bajo demanda estocástica, en la que los centros de distribución candidatos operan con una política de revisión continua (r, Q), tiempos de entrega determinísticos y full backorders. El problema integra simultáneamente decisiones de apertura de centros de distribución, asignación de retail y definición de los parámetros óptimos de la política de inventario, incorporando además restricciones de nivel de servicio basadas en fill-rate en cada instalación. Con el fin de representar adecuadamente estas interacciones, se formula un modelo no convexo de programación no lineal entera mixta (MINLP) que describe explícitamente el inventario disponible, los pedidos pendientes esperados y las restricciones de nivel de servicio bajo demanda normalmente distribuida. Posteriormente, se desarrolla una reformulación convexa equivalente mediante la introducción de variables auxiliares para los términos de agregación de demanda, obteniéndose un problema convexo de programación no lineal entera mixta con restricciones de conos de segundo orden. Sobre esta base, se propone un enfoque de resolución que combina Outer Approximation y Branch & Cut para resolver el problema de manera eficiente. El enfoque propuesto permite estudiar el desempeño de ambos métodos comparando sus tiempos computacionales y gap de optimalidad para ver qué método resulta más conveniente para este tipo de problemas.

Índice de Contenidos

1 | Introducción

En el entorno empresarial actual, las organizaciones enfrentan presiones crecientes para optimizar sus cadenas de suministro, equilibrando la reducción de costos operacionales con el mantenimiento de altos niveles de servicio al cliente. Las decisiones de localización de instalaciones y la gestión de inventarios, tradicionalmente abordadas de manera secuencial, requieren una integración que permita capturar las interdependencias entre ambas dimensiones para lograr soluciones verdaderamente óptimas.

El problema conjunto de localización e inventario (JLIP) surge como respuesta a esta necesidad, permitiendo la optimización simultánea de decisiones estratégicas de ubicación y operacionales de inventario bajo condiciones de demanda incierta. La incorporación explícita de restricciones de nivel de servicio, particularmente aquellas basadas en fill-rate, añade realismo al modelo pero también complejidad computacional, constituyendo un desafío metodológico relevante tanto para la investigación de operaciones como para la práctica empresarial.

Un problema de diseño de una red de distribución, como se muestra en Figura 1.1, considera una planta con capacidad ilimitada que abastece a cada centro de distribución (CD o DC por sus siglas en inglés) instalado, y que cada CD instalado puede abastecer a un conjunto de tiendas de retail (clientes), teniendo en cuenta que cada cliente puede ser abastecido por un único CD. Luego, el problema a resolver es determinar la red de distribución de recursos, lo que implica problemas tales como cuántos centros de distribución instalar, donde instalarlos, qué tienda retail abastecer desde que centro de distribución y las variables de las políticas de inventario para cada centro de distribución (cantidades de pedido y puntos de reorden) para minimizar el costo total por unidad de tiempo en cada centro de distribución. Este problema se conoce como el problema conjunto

de ubicación e inventario (Joint location inventory problem, JLIP).

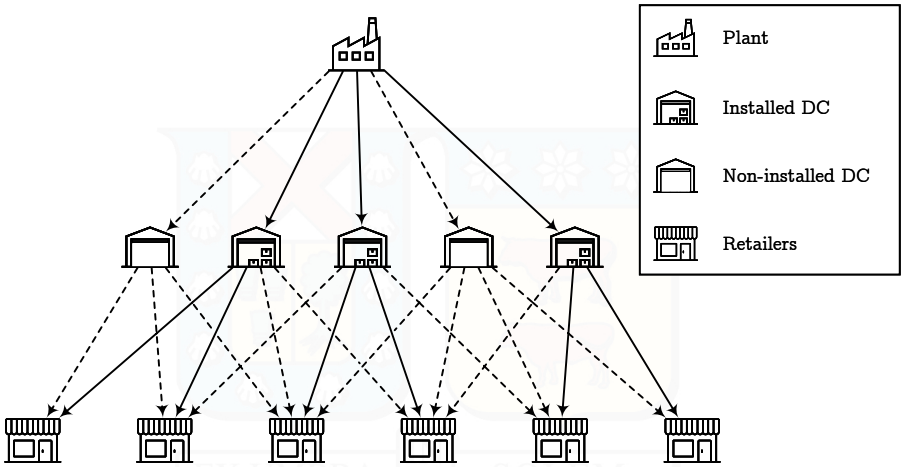


Figura 1.1: Ejemplo de una cadena de suministro (2 echelons).
Fuente: Elaboración propia.

Se conoce la ubicación de la planta, los sitios candidatos para los centros de distribución y las ubicaciones de los clientes. Sea I ($i = 1, \ldots, |I|$) el conjunto de retailers donde se asume que la demanda por unidad de tiempo en cada uno es independiente y se distribuye normalmente con μ_i y σ_i^2 como la demanda media y la varianza de la demanda, respectivamente. Sea J ($j = 1, \ldots, |J|$) el conjunto de emplazamientos candidatos para instalar centros de distribución. Cada centro de distribución j sigue una política de inventario de revisión continua (r, Q), siendo r la cantidad de inventario restante que activa el pedido de reposición de la cantidad Q, conocido como punto de reorden, con backorders y plazo de entrega determinista (L_j). Sea la posición de inventario el inventario disponible más los pedidos restantes menos los backorders, es decir, $IP_j(t) = IL_j^+(t + L_j) + D_j(t, t + L_j) - IL_j^-(t + L_j)$. Siempre que la posición de inventario de un centro de distribución cae por debajo de un punto de reorden fijo (r_j), se solicita a la planta una cantidad de pedido de tamaño constante (Q_j), que llega en un plazo de entrega fijo ($L_j > 0$) como se muestra en Figura 1.2.

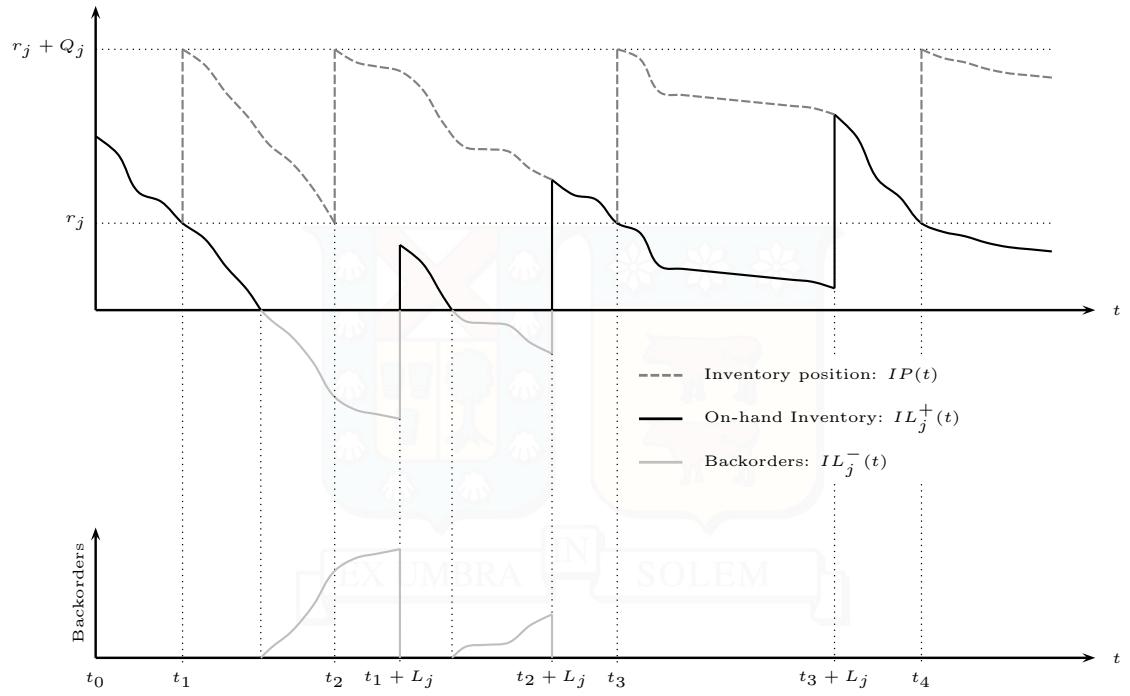


Figura 1.2: Comportamiento de la demanda con política de inventario (r, Q) con full-backorders y tiempos de entrega deterministas en cada CD_j.
Fuente: Escalona et al. (2021).

La escasez se controla mediante un fill-rate preestablecido, definido como la fracción de la demanda que puede satisfacerse utilizando el inventario disponible en cualquier momento (inventario on-hand), este parámetro se denomina $\bar{\beta} \in [0, 1]$.

Entonces finalmente el problema que se busca resolver es encontrar la red óptima de distribución sabiendo donde instalar cada centro de distribución, estos a que retailers van a abastecer y establecer las políticas de inventario acerca de qué tamaño debe ser el pedido del lote Q cuando el inventario llega al punto de reorden r.

1.1. Objetivos

Se busca desarrollar y resolver un modelo conjunto de localización e inventario que incorpore restricciones de nivel de servicio basadas en fill-rate y considerando que todo pedido pendiente se debe satisfacer, con el propósito de determinar la configuración óptima de la red de distribución y las políticas de inventario asociadas que minimicen el costo total del sistema.

1.1.1. Objetivos Específicos

- Formular el problema conjunto de localización e inventario considerando una red de distribución de dos niveles, demanda estocástica normalmente distribuida y políticas de inventario de revisión continua (r, Q) con full backorders.
- Incorporar explícitamente restricciones de nivel de servicio basadas en fill-rate, analizando su impacto sobre las decisiones de ubicación, asignación de clientes y parámetros de inventario.
- Formular el modelo original como un problema equivalente MINLP convexo, permitiendo su resolución mediante outer approximation y branch & cut.
- Implementar un algoritmo de solución basado en branch & cut que permita resolver instancias del problema.
- Evaluar el desempeño del modelo propuesto mediante experimentos numéricos, analizando sus resultados y rendimiento computacional.

2 | Marco Teórico

2.1. Control de Inventario

El control de inventario es fundamental en la gestión de cadenas de suministro y constituye uno de los problemas clásicos de la investigación de operaciones (Axsäter, 2015). Según Lewis (1998), la gestión efectiva del inventario requiere no solo del control de los niveles de stock, sino también de una adecuada predicción de la demanda que permita anticipar las necesidades futuras y ajustar las políticas de reposición para poder lograr mantener un buen nivel de servicio durante todo el horizonte de planificación. Esta integración entre control de inventario y predicción de demanda resulta crítica, ya que permite establecer políticas de reposición robustas que minimicen tanto los costos de mantenimiento de inventario como los costos asociados a situaciones de escasez, al mismo tiempo que se garantiza la disponibilidad de productos para satisfacer la demanda del cliente.

Los sistemas de control de inventario buscan responder dos preguntas fundamentales: ¿cuándo ordenar? y ¿cuánto ordenar? (Axsäter, 2015). La respuesta a estas preguntas depende de múltiples factores, incluyendo los patrones de demanda, los costos asociados al sistema (almacenamiento, pedido, faltantes), los tiempos de entrega, y los objetivos de servicio al cliente. Lewis (1998) enfatiza que la incertidumbre en la demanda es uno de los principales desafíos en el diseño de sistemas de inventario, ya que determina la cantidad de stock de seguridad necesaria para mantener niveles de servicio adecuados.

Es importante que los modelos de inventario se ajusten a las condiciones de la empresa para lograr su objetivo, debe representar de buena manera los escalones de la cadena de

suministro y sus características, para lo cual se deben considerar: los costos, políticas de pedido, como se afronta la escasez de inventario, como se comporta la demanda y de que forma establecer un buen nivel de servicio para el cliente.

2.1.1. Costos Considerados

La función objetivo de los modelos de control de inventario integra diversos componentes de costo que representan las diferentes operaciones y decisiones dentro de la cadena de suministro. Según Axsäter (2015), una correcta identificación y cuantificación de estos costos es fundamental para la toma de decisiones óptimas en la gestión de inventarios. El diseño de una cadena de suministro eficiente requiere equilibrar tanto costos estratégicos asociados a decisiones de largo plazo como costos operacionales que reflejan las operaciones diarias del sistema.

La inversión inicial y los gastos fijos asociados a la apertura y funcionamiento de instalaciones tales como centros de distribución o almacenes son parte de la estrategia. Esta característica binaria introduce complejidad computacional al problema e interactúa directamente con los costos de transporte, que operan tanto en el flujo ascendente desde proveedores o plantas hacia los centros de distribución, como en el flujo descendente hacia los clientes finales.

Dentro de cada instalación operan los costos de inventario, cuya estructura presenta una dinámica que caracteriza fundamentalmente las políticas de reposición óptimas. El mantenimiento de mercancía almacenada implica la inmovilización de recursos económicos que, además del costo de oportunidad, conlleva gastos asociados a deterioro, obsolescencia, seguros y espacio físico. Esta estructura de costos de almacenamiento crece proporcionalmente con el nivel de inventario mantenido, incentivando la reducción de stock. Sin embargo, la reducción del inventario promedio requiere incrementar la frecuencia de reposición. Este comportamiento configura el dilema central de las políticas de inventario: la frecuencia y magnitud de los pedidos debe equilibrar el costo de mantener inventario contra el costo de realizar reposiciones.

Los costos de escasez completan la estructura económica del problema, surgiendo cuando la demanda supera el inventario disponible y materializándose como ventas perdidas

o costos de penalización por pedidos atrasados. La dificultad práctica de cuantificar estas penalizaciones ha llevado a muchos modelos a regular su impacto indirectamente mediante restricciones de nivel de servicio, que garantizan que un cierto porcentaje de la demanda sea satisfecho directamente desde el inventario disponible (Zipkin, 1986). El modelado adecuado de la escasez y su interacción con las decisiones de inventario será abordado con mayor detalle en secciones posteriores.

2.1.2. Políticas de Pedidos

Para poder representar de forma realista un sistema de inventario es necesario implementar algún método que permita cuantificar las unidades de producto. Para esto la literatura implementa dos formas de revisión: continua y periódica. Además de implementar el concepto de posición de inventario que nos dará las herramientas que ayudarán a la cuantificación de inventario.

2.1.2.1. Posición del Inventario y Nivel de Inventario

Según Axsäter (2015), la decisión de cuándo y cuánto pedir no puede basarse únicamente en el inventario físico disponible. Para describir correctamente la situación del stock también es necesario considerar los pedidos ya realizados que aún no llegan y las unidades comprometidas con clientes que todavía no han sido entregadas. Por ello, en control de inventarios se utiliza el concepto de posición de inventario, definida como:

$$IP(t) = IL^{+}(t) + PP(t) - IL^{-}(t), \tag{2.1}$$

en donde $IP(t)$ es la posición del inventario en un tiempo t, $IL^{+}(t)$ nivel de inventario disponible a la mano (también llamado on-hand inventory $OH(t)$), $PP(t)$ las órdenes pendientes en curso y $IL^{-}(t)$ los pedidos atrasados (backorders). Se asume que en un largo plazo el efecto de los pedidos pendientes es despreciable, por lo tanto, se pueden descartar.

Por otro lado, si bien las decisiones de reabastecer inventario dependen de la posición del inventario, los costos de mantener el inventario y la escasez de este dependerán del

nivel de inventario definido como:

$$IL(t) = IL^{+}(t) - IL^{-}(t). \tag{2.2}$$

2.1.2.2. Revisión Continua

La política de revisión continua, constituye uno de los enfoques más utilizados para la gestión de inventarios en entornos donde es posible el monitoreo constante de los niveles de stock. Este sistema se fundamenta en el seguimiento permanente de la posición del inventario, permitiendo detectar inmediatamente cuando esta desciende por debajo de un umbral predefinido denominado punto de reorden. Una vez que se alcanza este nivel crítico, se emite una orden de reposición por una cantidad fija, la cual se recibirá después de transcurrir un periodo determinístico L conocido como lead-time o tiempo de entrega (Van Horenbeek et al., 2013).

A partir de esto las dos políticas de revisión continua más comunes son: (r, Q) y (s, S). En el primer caso, se realiza una orden de reabastecimiento de tamaño Q cuando la posición de inventario disminuye hasta el punto de reorden r, debido a los requerimientos de demanda $D(t)$ de los clientes. (Silver et al., 1998). Producto de esto, la posición de inventario alcanza $r + Q$ unidades, que se puede interpretar como el inventario virtual que se tiene actualmente, mientras que la posición del inventario o inventario a la mano (on-hand inventory) solo es actualizada una vez llega realmente el reabastecimiento luego de un tiempo L producto del lead-time. Esta política de revisión es muy útil para bienes de rápido movimiento (Fast moving consumer goods), ya que son productos que tienen una demanda y tasa de rotación muy alta, por lo cual requieren de un constante monitoreo.

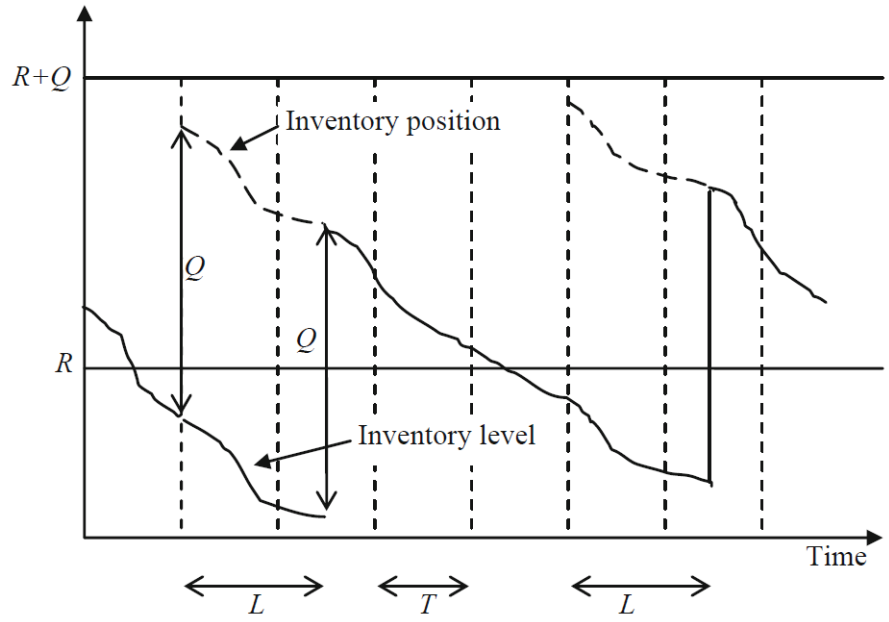


Figura 2.1: Política de revisión periódica (r, Q) con demanda normalmente distribuida.
Fuente: Axsäter (2015).

Por otro lado, la política (s, S) muy similar a la anterior, con la diferencia que el punto de reorden ahora llamado s y el inventario en vez de reabastecerse en un tamaño de lote fijo Q, se hace un pedido tal que el inventario llegue a un nivel objetivo S, además considerar que $s \leq S$. Es decir, que una vez la posición del inventario haya llegado al punto s, se hace un pedido que reabastezca hasta tener S unidades (Van Horenbeek et al., 2013). Esta política es particularmente útil para bienes de baja rotación, que requieran mantener un stock no muy alto en todo momento. Se puede establecer una relación entre las políticas (r, Q) y (s, S) cuando se cumple que $S = r + Q$.

2.1.2.3. Revisión Periódica

A diferencia del sistema de revisión continua, esta política implica que la información sobre los niveles de stock se obtiene únicamente en momentos específicos, con intervalos de tiempo fijos T previamente determinados (Waters, 2003).

Este enfoque ofrece beneficios específicos, particularmente cuando se busca sincronizar las órdenes de abastecimiento para múltiples productos. Si bien los avances tecnológicos han disminuido significativamente los gastos asociados con el monitoreo de inventarios, la implementación de sistemas de revisión periódica contribuye adicionalmente a la reducción

de costos operacionales del control de inventario. Esta ventaja resulta más pronunciada en productos que presentan niveles elevados de rotación. En contraste, para productos de baja rotación, el costo adicional de implementar revisión continua es marginal, y frecuentemente los beneficios de este último sistema superan a los de la revisión periódica.

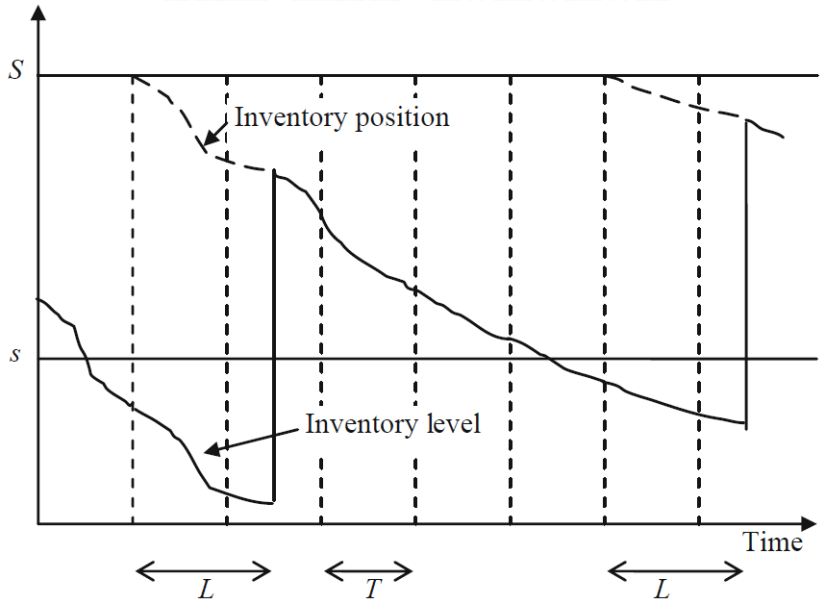


Figura 2.2: Política de revisión periódica (s, S) con demanda normalmente distribuida.
Fuente: Axsäter (2015).

Por consiguiente, en aplicaciones reales se tiende a emplear revisión continua para productos de baja rotación, mientras que la revisión periódica se reserva para aquellos con mayor volumen de demanda. Es importante destacar que cuando el intervalo de revisión T es suficientemente reducido, el comportamiento del sistema de revisión periódica se aproxima considerablemente al de revisión continua. (Axsäter, 2015)

Lo que está claro es que con los avances en la tecnología, ya no es costoso ni añade mayor dificultad implementar un sistema de revisión continua para saber cuanto inventario se posee en todo momento (Wang and Chen, 2022). Esto facilita la adopción de esta política para prevenir quiebres de stock. Este trabajo se enfoca principalmente en una revisión continua con política de abastecimiento (r, Q).

2.1.3. Escasez de Inventario

El indicador fundamental para evaluar la eficiencia en los sistemas de control de inventario es la disponibilidad de productos. Las situaciones de desabastecimiento se presentan cuando la demanda excede el inventario disponible, generando quiebres de stock. Ante estas circunstancias, la literatura identifica dos estrategias principales para afrontarlo: Ventas perdidas y Pedidos Pendientes (Estellés-Miguel et al., 2014).

2.1.3.1. Ventas perdidas

El modelo de ventas perdidas (lost sales) representa una situación donde la demanda no satisfecha se pierde definitivamente cuando el nivel de inventario ($IL(t)$) es insuficiente para cubrirla. Formalmente, cuando el inventario disponible (on-hand inventory) es menor que la demanda solicitada, la porción no satisfecha de la demanda no se convierte en una orden pendiente, sino que se considera como una venta perdida que no se recuperará en el futuro. Debido a esto, la posición de inventario no puede ser negativa ($IL^{-}(t) = 0$) (Axsäter, 2015).

En determinadas circunstancias, considerar que todos los pedidos insatisfechos se mantienen como órdenes pendientes puede no resultar realista. Por ejemplo, resulta más apropiado modelar los quiebres de stock como ventas perdidas cuando los retailers operan en mercados altamente competitivos donde los clientes pueden recurrir fácilmente a empresas alternativas para adquirir el producto deseado. (Andersson and Melchiors, 2001)

2.1.3.2. Backorders

En sistemas donde la escasez se maneja mediante pedidos pendientes, un indicador de desempeño fundamental es el nivel promedio de órdenes atrasadas, representado por $B(r, Q)$. Esta métrica constituye una medida relevante del nivel de servicio y forma parte de los componentes estándar utilizados en la expresión del inventario promedio (Zipkin, 1986).

Los pedidos atrasados (backorders) representan, por consiguiente, la demanda insatisfecha que permanece registrada como pedido pendiente y será satisfecha cuando se realice la

próxima reposición de inventario posterior al episodio de desabastecimiento.

Liu et al. (2010) desarrollan un modelo de revisión periódica con nivel base de inventario (Base Stock Level) para una demanda continua, estocástica y normalmente distribuida, incorporando costos por pedidos pendientes. Para ello, proponen una aproximación lineal de los backorders a partir del inventario disponible al término de cada ciclo de reposición. Por otro lado, Berman et al. (2012) formulan un modelo bajo una política de revisión periódica (R, S), donde la función objetivo incluye costos de faltantes estimados como una proporción de la demanda. En cambio, Miranda and Garrido (2009) presentan una expresión exacta para calcular los pedidos pendientes, modelándolos como la esperanza de que la demanda durante el tiempo de entrega supere o iguale el punto de reorden; no obstante, no consideran un esquema de backorders completos, sino que emplean esta formulación como herramienta de solución. En consecuencia, obtener una expresión exacta para la esperanza de los backorders ($\mathbb{E}[IL^{-}]$) resulta un problema complejo y no fácil de abordar.

Zhang (1998) y Zipkin (1986) demuestran que la función que describe el comportamiento de los backorders es convexa para una política de revisión continua (r, Q), siendo conjunta convexamente en r y Q, asumiendo una demanda estocástica con distribución normal.

2.1.4. Control de escasez de inventario

La ocurrencia de eventos de escasez puede disminuirse mediante la implementación de políticas de inventario que garanticen elevados niveles de disponibilidad de productos. Para la determinación de los parámetros óptimos de estas políticas, la literatura se ha concentrado en dos categorías principales de problemas de control de inventario: el modelo de costo total (Full-cost) y el modelo con restricciones de nivel de servicio.

En el modelo full-cost, el objetivo consiste en identificar los parámetros óptimos de una política de inventario que minimice la suma de los costos de mantenimiento, pedido y escasez. Para asegurar altos niveles de disponibilidad, estos modelos utilizan penalizaciones por situaciones de quiebre de stock. Por otro lado, el enfoque basado en nivel de servicio sustituye el costo de escasez por una restricción de nivel de servicio. El propósito es proporcionar un nivel específico de servicio minimizando únicamente los costos de pedido

y mantenimiento de inventario (Escalona et al., 2021).

2.1.4.1. Full-Cost

En este modelo, la política de inventario analiza tres elementos clave: el costo de mantener existencias por unidad y por unidad de tiempo (h), el costo fijo de realizar un pedido (S) y las penalizaciones derivadas de la falta de stock, buscando minimizar el costo total resultante (Schneider, 1981). Estelles-Miguel et al. (2014) señalan que, en situaciones reales, cuantificar económicamente las faltas es complejo, por lo que gran parte de la investigación se orienta hacia modelos basados en niveles de servicio. Aun así, algunos trabajos emplean directamente estos costos para reducir los backorders. En este contexto, Axsäter (2015) diferencia dos formas de penalización por escasez: una proporcional al tiempo de atraso por unidad y otra aplicada por cada unidad no satisfecha.

El costo de escasez por unidad y por unidad de tiempo, denotado como b_1, se expresa en las mismas unidades que el costo de almacenamiento. Este puede interpretarse como el valor económico asociado al tiempo de espera del cliente o como el costo esperado de mantener pedidos pendientes. Considerando una política de revisión continua (r, Q) con demanda aleatoria durante el lead time, caracterizada por media μ y desviación estándar σ, los costos de mantenimiento y de escasez pueden describirse mediante la expresión:

$$h \cdot (IL^+) + b_1 \cdot (IL^-) = -b_1 \cdot IL + (h + b_1) \cdot (IL^+) = h \cdot IL + (h + b_1)(IL^-), \tag{2.3}$$

donde $h \cdot (IL^+)$ representa el costo de inventario disponible (stock positivo) y $b_1(IL^-)$ el costo asociado a los faltantes o backorders (stock negativo). Esta formulación permite expresar ambos tipos de costo en función del nivel de inventario, facilitando su análisis dentro del modelo.

2.1.4.2. Restricción de Nivel de servicio

α_L-Service Level

El nivel de servicio α se entiende como la probabilidad de que no ocurra un quiebre de stock en un instante cualquiera (Schneider, 1981). Esta métrica, también denominada ready-

rate, puede interpretarse como la proporción del tiempo en que el inventario disponible es positivo. No obstante, dado que únicamente considera la ocurrencia de faltantes y no su magnitud, para políticas de revisión continua se introduce una medida adicional: el nivel de servicio α_L. Este indicador evalúa la probabilidad de evitar quiebres de stock durante un ciclo completo de reabastecimiento, siendo L el tiempo de entrega (lead-time).

Si $\alpha_L(r)$ representa el nivel de servicio asociado al punto de reorden r, entonces existirá desabastecimiento en un ciclo si la demanda durante el lead-time supera estrictamente dicho punto, es decir, si $D(L) > r$. En consecuencia, el nivel de servicio puede expresarse como:

$$\alpha_L(r) = \mathbb{P}(D(L) \geq r) = F_{D(L)}(r). \tag{2.4}$$

β-Service Level

El nivel de servicio β, conocido también como fill-rate, fue planteado por Brown (1967) y corresponde a la proporción de la demanda total que logra satisfacerse sin incurrir en pérdidas de ventas. En otras palabras, mide qué fracción de la demanda es atendida directamente a partir del inventario disponible (inventario on-hand) en un momento cualquiera, reflejando así la capacidad efectiva del sistema para cumplir con los requerimientos de los clientes. A diferencia de otras métricas, esta no se limita a identificar si ocurre o no un quiebre de stock, sino que cuantifica cuánto de la demanda queda realmente insatisfecha.

Posteriormente, Schneider (1981) enfatiza que esta medida incorpora no solo la ocurrencia de faltantes, sino también el volumen de los pedidos pendientes (backorders), lo que la convierte en un indicador más completo del desempeño del inventario. Desde esta perspectiva, β puede interpretarse como una representación proporcional del costo asociado a la acumulación de demanda no atendida, ya que penaliza en mayor medida los faltantes de gran magnitud.

Bajo este enfoque, Zipkin (1986) formaliza el nivel de servicio objetivo $\beta(r, Q)$ dentro de políticas de revisión continua, expresándolo mediante una función que depende del punto de reorden r y del tamaño del lote Q, lo que permite vincular directamente la elección de la política de inventario con el grado de servicio esperado. En consecuencia, $\beta(r, Q)$ se

define mediante la siguiente expresión:

$$\beta(r, Q) = 1 - \frac{A(r, Q)}{\mu}, \tag{2.5}$$

con $A(r, Q)$ es la función que describe el promedio de quiebre de stock por unidad de tiempo, μ la media de la demanda distribuida normalmente, por unidad de tiempo. Además, Escalona et al. (2021) establece que esta función corresponde a la demanda promedio por la probabilidad de que la demanda durante el leadtime en un periodo sea menor que cero, expresado matemáticamente: $A(r, Q) = \mu\mathbb{P}(IL(t + L) < 0)$.

γ-Service Level

Por otro lado, se puede controlar la duración de un quiebre de stock mediante el nivel de servicio γ. Schneider (1981) lo define como la proporción de la demanda que pasa a ser un backorder en un periodo y se define en la siguiente expresión:

$$\gamma(r, Q) = 1 - \frac{B(r, Q)}{\mu}, \tag{2.6}$$

donde $\gamma(r, Q)$ es el nivel de servicio proporcionado.

2.1.5. Política de inventario con Revisión Continua (r, Q)

Los bienes de consumo rápido (FMCG) son productos no duraderos que satisfacen las necesidades diarias de los consumidores (Bala and Kumar, 2011). Son productos que tienen una alta rotación y demanda, se venden al por menor y no son altamente costosos, ejemplos de estos son: alimentos y bebidas, productos de higiene personal, productos de limpieza, entre otros. Es importante resaltar que en el diseño de una cadena de suministro para FMCGs hay una diferencia entre los consumidores y el cliente final, que para este caso, son los retailers que venden estos productos (Aljunaidi and Ankrak, 2014), por lo cual podemos catalogar esta industria como una del tipo business-to-business.

Axsäter (2015) señala que, cuando se trata de productos con alta tasa de demanda, resulta más conveniente emplear aproximaciones de tipo continuo sobre esta. De hecho, distintos estudios que analizan la demanda de bienes de consumo de rápida rotación a lo

largo del tiempo indican que una forma eficiente de representarla es mediante un proceso estocástico estacionario con distribución normal (Nasiri et al., 2021; Chung et al., 2009). En cambio, para productos de alto valor, baja rotación, demanda reducida y elevados costos de inventario, es habitual modelar la demanda mediante un proceso estocástico con distribución de Poisson compuesta (Axsäter and Zhang, 1999).

2.1.6. Nivel de inventario

Como hemos visto a lo largo del documento, los modelos que sugieren diversos autores dependen del comportamiento de la demanda, en particular, del nivel de inventario y de qué forma distribuye este. Se toma como supuesto que esta se comporta de manera normal con media μ y varianza σ^2.

A partir de lo planteado por Axsäter (2015), se establece que, en estado estacionario, la posición de inventario $IP(t)$ se distribuye de manera uniforme dentro del intervalo $[r, r+Q]$. En consecuencia, su función de densidad puede expresarse como $f(x) = \frac{1}{Q}$ para todo $x \in [r, r+Q]$. Por otro lado, la demanda observada durante el intervalo $[t, t+L]$ sigue una distribución normal con media $\mu' = \mu L$ y desviación estándar $\sigma' = \sigma\sqrt{L}$, parámetros que representan la demanda acumulada durante el lead-time. Bajo estas consideraciones, la función de distribución acumulada que describe el inventario disponible puede expresarse de la siguiente forma:

$$F(x) = P(IL \leq x) = \frac{1}{Q}\int_{R}^{R+Q}\left[1 - \Phi\left(\frac{u - x - \mu'}{\sigma'}\right)\right]du, \tag{2.7}$$

donde $\Phi(x)$ es la función de distribución normal. Por otro lado, se define la función de pérdida $G(x)$ como:

$$G(x) = \int_{x}^{\infty}(v - x)\varphi(v)dv = \varphi(x) - x(1 - \Phi(x)), \tag{2.8}$$

notando que $G(x)$ es una función decreciente y convexa para x además notar que $G'(x) =$

$\Phi(x) - 1$ y que $\varphi(x)$ es la función de densidad normal. Esto permite reformular (2.7) como:

$$F(x) = \frac{1}{Q}\int_{R}^{R+Q}\left[-G'\left(\frac{u-x-\mu'}{\sigma'}\right)\right]du \tag{2.9}$$

$$= \frac{\sigma'}{Q}\left[G\left(\frac{R-x-\mu'}{\sigma'}\right) - G\left(\frac{R+Q-x-\mu'}{\sigma'}\right)\right]. \tag{2.10}$$

2.1.7. Control de inventario con restricción de nivel de servicio fill-rate

Escalona et al. (2021) define un modelo de control de inventario bajo una política de revisión continua (r, Q), con full backorders, demanda estocástica normal y una restricción de nivel de servicio fill-rate. En este enfoque, la decisión de reabastecimiento se toma de manera conjunta a través del punto de reorden r y del tamaño de lote Q, considerando el comportamiento aleatorio de la demanda y el efecto de los backorders acumulados. Este modelo, denominado $\beta -$ SLC, busca minimizar el costo total de mantener inventario junto a las reposiciones y a los pedidos pendientes:

$$\beta - \text{SLC:} \quad \min_{r,Q} \quad S\frac{\mu}{Q} + h\left(\frac{Q}{2} + r - \mu L + B(r,Q)\right) \tag{2.11}$$

$$\text{s.t.} \quad \beta(r,Q) \geq \bar{\beta} \tag{2.12}$$

$$r \geq \mathbb{E}(D(L)) \tag{2.13}$$

$$Q \geq Q_{EOQ}. \tag{2.14}$$

La función objetivo del modelo (2.11), representa los costos totales de inventario que considera en su primera expresión los costos de set-up por unidad de tiempo y la segunda expresión los costos de almacenamiento por unidad de tiempo. Mientras que en las restricciones tenemos la restricción asociada al nivel de servicio (2.12), en ella se establece que el nivel de servicio alcanzado debe ser al menos igual a $\bar{\beta} \in [0, 1]$, valor que representa el nivel de servicio preestablecido. Este parámetro indica la fracción mínima de la demanda que debe satisfacerse directamente con el inventario disponible en un momento cualquiera. La restricción (2.13) establece que el inventario de seguridad no puede tomar valores negativos, ya que se define como $r - \mathbb{E}(D(L))$. La restricción (2.14) establece que la

cantidad de pedido debe ser al menos igual al tamaño del lote económico $Q_{EOQ} = \sqrt{\frac{2\mu S}{h}}$. Esta condición se impone para evitar que el primer término de la función objetivo resulte indefinido aunque bajo condiciones KKT, la variable Q será estrictamente positiva.

2.2. Localización de instalaciones

La localización de instalaciones se puede categorizar según distintos criterios o clasificaciones, pero en general todos los modelos de localización comparten un objetivo, determinar las ubicaciones óptimas de las instalaciones que se necesitan localizar y como asignar la demanda de los clientes. Algunas de las clasificaciones pueden ser si se tratan de modelos continuos, de redes o discretos, es decir, en el caso continuo la ubicación de un punto en cualquier parte del plano. O en una red establecida descrita por un grafo del tipo $G(N, A)$, donde N son los nodos candidatos de instalación y A los arcos que los unen. Todo esto bajo el supuesto de la ubicación de cada nodo y sitio candidato es conocida así como también su demanda.

Por otro lado, tenemos la diferencia entre que tipo de objetivo se busca minimizar o maximizar, ya que depende de las decisiones de quien requiera el modelo. En el caso del sector privado, por ejemplo, los costos o beneficios son descritos monetariamente mientras que en el sector público estos son descritos principalmente por el cliente a quien se quiera atender.

Otros tipos de clasificaciones consideran las metodologías usadas: como se calcula la distancia entre nodos, o el tipo de instalaciones, si tienen límites de capacidad o no, o si se entrega un solo producto o múltiples, entre otras. Daskin (2013) agrupa estas clasificaciones en los siguientes grupos: Problemas de cobertura, problemas de centro, problemas de p-mediana y problemas de localización de instalaciones de carga fija.

2.2.1. Modelo UFLP

Para este caso se analizan modelos de localización discreta desde la perspectiva del sector privado de proveedores de bienes de consumo de alta rotación (FMCG). Se consideran instalaciones sin restricciones de capacidad, dedicadas a un único producto y evaluadas en

un solo período de tiempo. Un modelo que reúne estas características es el Uncapacitated Facility Location Problem (UFLP), que consiste en determinar la ubicación de instalaciones y la asignación de clientes de manera que se minimice el costo total.

Modelo propuesto por Stollsteimer (1961) y Balinski (1964), con objetivo de instalar una cantidad indefinida de instalaciones al menor costo. Considera los costos fijos de instalación en términos anuales y costos variables asociados a la satisfacción de la demanda, por ejemplo, el costo de transporte.

Sean los conjuntos que describen los puntos de demanda I indexado en $i = 1, ..., |I|$ y las instalaciones candidatos J indexado en $j = 1, ..., |J|$. Donde el objetivo del modelo es minimizar los costos de instalación y asignación de demanda, de los nodos i a los centros j, con f_j el costo fijo de instalar un servidor en el sitio candidato j y C_{ij} es el costo del servicio asociado a asignar la demanda del sitio i al centro j. El modelo es un problema entero (IP), el modelo es el siguiente:

$$\text{UFLP:} \quad \min_{X,Y} \quad \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} C_{ij} Y_{ij} \tag{2.15}$$

$$\text{s.t.} \quad \sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{2.16}$$

$$Y_{ij} \leq X_j \qquad \forall i \in I,\ j \in J \tag{2.17}$$

$$X_j,\ Y_{ij} \in \{0, 1\} \qquad \forall i \in I,\ j \in J, \tag{2.18}$$

donde las variables de decisión X_j, es 1 si se instala un centro candidato en la ubicación j, 0 si no. E Y_{ij} 1 si la demanda del cliente i está siendo satisfecha por el centro j, 0 si no. La primera expresión de la función objetivo (2.15), corresponde al costo fijo de instalación y la segunda expresión a los costos por satisfacción de la demanda, donde C_{ij} puede ser descrita como $C_{ij} = h_i d_{ij} c$, donde h_i es la demanda del nodo i, d_{ij} la distancia entre el cliente i y el centro j, y c el costo por unidad de distancia por unidad de demanda.

En cuanto a las restricciones: la restricción (2.16), describe que todos los clientes i deben ser atendidos por un, y solo un, centro j, también conocida como *single source constraint*. La restricción (2.17), también llamada restricción de Balinsky, describe que los clientes i solo pueden ser atendidos si el centro j se encuentra instalado. La restricción

(2.18) la integralidad de las variables de decisión.

2.3. Localización e Inventario conjunto

Hasta este punto solo hemos hablado de los problemas de inventario y de los problemas de localización por separado, pero el problema radica cuando se busca resolver un problema donde se requiere localizar instalaciones a la vez que resuelven las decisiones óptimas para la gestión del inventario, esto se denomina Joint Location-Inventory Problem (JLIP).

Manatkar et al. (2016) menciona que los factores que más afectan en el diseño de una cadena de suministro, son: localización, transporte, abastecimiento e inventario. Afectan no solo en su diseño sino que también en los costos asociados y además entre estos mismos, por lo que la decisión de abordarlos en conjunto es fundamental para lograr un diseño eficiente y óptimo de la cadena de suministro.

La interdependencia entre estos factores se manifiesta de múltiples formas. Las decisiones de localización de instalaciones determinan directamente los costos de transporte entre los diferentes eslabones de la cadena, ya que definen las distancias y rutas que deben recorrer los productos. A su vez, estas decisiones de ubicación influyen en las políticas de abastecimiento, pues afectan los tiempos de entrega (lead-times) y, por consecuencia, los niveles de inventario de seguridad requeridos para mantener un nivel de servicio adecuado.

Cuando estas decisiones se toman de manera secuencial o independiente, se puede incurrir en soluciones subóptimas que no reflejan las verdaderas compensaciones (trade-offs) existentes en el sistema (Escalona et al., 2015). Por ejemplo, una solución que minimiza únicamente los costos de localización podría generar instalaciones muy dispersas que, si bien reducen los costos fijos, incrementan significativamente los costos de transporte y los requerimientos de inventario debido a mayores tiempos de entrega.

Para esto Daskin et al. (2002) y Shen et al. (2003) proponen un modelo de localización con inventario conjunto en donde un conjunto de retail I ($i = 1, ..., |I|$) son atendidos por un conjunto de centros de distribución J ($j = 1, ..., |J|$). Buscando resolver cuatro preguntas principales: donde instalar los centros, a qué retail atenderán estos centros, cuanto será el punto de reorden en cada centro y el tamaño de lote en cada centro, es decir, cuando y

cuanto pedir al momento de reabastecer cada centro de distribución. Todo esto minimizando el costo total de localización, transporte e inventario, para asegurar el nivel de servicio establecido para los clientes Atamtürk et al. (2012).

Por otra parte, el JLIP bajo una política (r, Q) con nivel de servicio α_L es un problema ampliamente estudiado. Daskin et al. (2002) analizan este problema utilizando la cantidad económica de pedido (EOQ) y un punto de reorden estocástico como aproximaciones de los parámetros de una política de revisión continua (r, Q) con nivel de servicio α_L. La demanda de los clientes se modela mediante una distribución normal como aproximación para una demanda no negativa. El modelo se formula como un problema entero no lineal (INLP), debido a que el tamaño del lote EOQ y el punto de reorden se expresan en función de las variables de localización y asignación.

El problema se resuelve mediante relajación lagrangiana, aplicada en varias etapas para tratar la expresión no lineal asociada a la variable de asignación y posteriormente obtener una solución factible. Este trabajo motivó el desarrollo de diversos enfoques de solución para la misma formulación, entre ellos generación de columnas (Shu et al., 2005; Shen et al., 2003), relajación lagrangiana (You and Grossmann, 2008) y la reformulación CQMIP propuesta por Atamtürk et al. (2012), la cual puede resolverse directamente con softwares de optimización estándar.

El modelo también ha sido extendido en diversas direcciones. Entre ellas se encuentran versiones con restricciones deterministas de capacidad en los centros de distribución (Atamtürk et al., 2012; Ozsen et al., 2008; Miranda and Garrido, 2004), formulaciones con múltiples commodities (Atamtürk et al., 2012; Shen, 2005) y modelos que consideran correlación de demanda entre retailers (Shahabi et al., 2014; Atamtürk et al., 2012). Asimismo, se han desarrollado extensiones multi-echelon que incorporan centros de reprocesamiento para productos devueltos (Diabat et al., 2015), versiones estocásticas basadas en escenarios de demanda y costos (Snyder et al., 2007), modelos con niveles de servicio diferenciados para distintos tipos de retailers (Escalona et al., 2018, 2015) y formulaciones con capacidad estocástica que controlan la probabilidad de exceder la capacidad de inventario (Tapia-Ubeda et al., 2018; Miranda and Garrido, 2006, 2008)

En la mayoría de estos trabajos, al igual que en Daskin et al. (2002), los parámetros de la

política de control de inventario (r, Q) se aproximan mediante la EOQ y un punto de reorden estocástico. Además, se suele asumir que la componente de backorders es despreciable en la expresión del inventario disponible, ya que esta es controlada vía restricciones de servicio.

2.4. Programación Cónica

La programación cónica puede entenderse como una extensión de la programación lineal. Dentro de esta familia se distinguen principalmente dos subclases: la programación cónica de segundo orden (SOCP) y la programación semidefinida positiva (SDP). La primera se caracteriza por emplear conos de segundo orden, también conocidos como conos de Lorentz, mientras que la segunda se basa en la utilización de conos formados por matrices reales simétricas semidefinidas positivas.

2.4.1. Second order cone programming (SOCP)

En los problemas tipo SOCP se busca minimizar (o maximizar) una función lineal considerando como región factible la intersección entre un conjunto afín y uno o más conos de segundo orden. En consecuencia, los problemas de tipo SOCP corresponden a modelos de optimización convexa no lineal, ya que combinan restricciones lineales con restricciones cuadráticas convexas. A continuación, se presenta la formulación general de un modelo de programación cónica de segundo orden:

$$\min_{x} \quad f^T x \tag{2.19}$$

$$\text{s.t.} \quad \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad \forall i = 1, \ldots, N, \tag{2.20}$$

donde $x \in \mathbb{R}^n$ es la variable de optimización y los parámetros del problema son $f \in \mathbb{R}^n$, $A_i \in \mathbb{R}^{(n_i-1)\times n}$, $b_i \in \mathbb{R}^{n_i-1}$, $c_i \in \mathbb{R}^n$ y $d_i \in \mathbb{R}$. Además, $\|\cdot\|_2$ denota la norma euclidiana y la expresión (2.20) corresponde a una restricción cónica de segundo orden de dimensión n.

A modo general, en Boyd et al. (2004) se define un cono de segundo orden (también

conocido como cono de Lorentz) como el conjunto:

$$C = \{(x, t) \in \mathbb{R}^{n+1} \mid \|x\|_2 \leq t\}. \tag{2.21}$$

Geométricamente, el conjunto C representa todos los puntos cuyo vector x tiene una norma euclidiana menor o igual que el escalar t, formando una región convexa con forma de cono en el espacio $\mathbb{R}^{n+1}$. Esta propiedad de convexidad es fundamental, ya que garantiza que los problemas de optimización formulados con este tipo de restricciones puedan resolverse de manera eficiente mediante algoritmos de optimización convexa, como los métodos de punto interior.

2.5. Métodos de resolución para problemas MINLP

Un amplio rango de problemas de optimización no lineal involucran variables de decisión discretas junto con variables continuas, este tipo de problema se denominan mixed-integer nonlinear programming problems (MINLP)

Las variables enteras pueden emplearse para representar, por ejemplo, secuencias de eventos, alternativas entre distintos candidatos o la presencia o ausencia de ciertas unidades cuando se utilizan en su forma binaria. Por su parte, las variables discretas permiten modelar, por ejemplo, diferentes tamaños o capacidades de equipos. Finalmente, las variables continuas se utilizan para describir las relaciones de entrada y salida, así como las interacciones entre unidades u operaciones individuales y entre distintos sistemas interconectados.

Las principales aplicaciones de los enfoques MINLP en el área de diseño, programación y planificación de procesos batch dentro de la ingeniería química incluyen diversos problemas relacionados con la configuración y operación de plantas industriales. Entre estos destacan el diseño de plantas multiproducto y el diseño y la programación de plantas multipropósito, donde se busca determinar la mejor estructura y utilización de los equipos para producir distintos productos de manera eficiente. Estas aplicaciones se enfocan en optimizar tanto la estructura del proceso como la asignación y secuencia de las operaciones

a lo largo del tiempo (Floudas, 1995). Para este trabajo, el enfoque MINLP se utiliza en el contexto del diseño de redes, con el objetivo de determinar la configuración óptima del sistema considerado.

Los problemas de programación no lineal entera mixta (MINLP) pueden resolverse mediante diversos enfoques que combinan técnicas de optimización para manejar simultáneamente variables continuas, discretas y restricciones no lineales. Entre los métodos más conocidos se encuentra Generalized Benders Decomposition (GBD), que descompone el problema en subproblemas y genera cortes que se incorporan iterativamente a un problema maestro para aproximar la región factible. También destaca Extended Cutting Plane (ECP), el cual construye aproximaciones lineales externas de las restricciones no lineales mediante la incorporación progresiva de planos de corte. Además, en problemas de gran escala o alta complejidad computacional es común emplear heurísticas, que corresponden a estrategias de búsqueda diseñadas para encontrar soluciones factibles de buena calidad en tiempos de cómputo reducidos. Estas técnicas no garantizan necesariamente la optimalidad global, pero permiten explorar eficientemente el espacio de soluciones y obtener aproximaciones útiles cuando los métodos exactos resultan computacionalmente muy costosos. Entre los algoritmos exactos más utilizados se encuentran Branch and Bound, que explora el espacio de soluciones mediante un proceso de ramificación sobre las variables discretas utilizando cotas para descartar regiones no prometedoras; Branch and Cut, que extiende este esquema incorporando planos de corte para fortalecer las relajaciones del problema; y Outer Approximation, que utiliza aproximaciones lineales de las funciones no lineales para construir y resolver iterativamente problemas maestros que refinan la solución. En este trabajo se emplearán específicamente los métodos Branch & Cut y Outer Approximation para la resolución del modelo MINLP que se presentará.

2.5.1. Outer Approximation

Duran and Grossmann (1986) proponen su enfoque de aproximación externa a partir de

lo siguiente. Sea P un problema MINLP tal que:

$$P: \quad \min_{x,y} \; c^T\mathbf{y} + f(\mathbf{x}) \tag{2.22}$$

$$\text{s.t. } g(\mathbf{x}) + B\mathbf{y} \leq 0 \tag{2.23}$$

$$\mathbf{x} \in X \subset \mathbb{R}^n \tag{2.24}$$

$$\mathbf{y} \in Y, \tag{2.25}$$

donde $X = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n,\ A_1\mathbf{x} \leq a_1\}$ e $Y = \{\mathbf{y} : \mathbf{y} \in \{0,1\}^q,\ A_2\mathbf{y} \leq a_2\}$. Las condiciones que deben cumplirse para aplicar el método son las siguientes:

- X debe ser un poliedro no vacío, compacto y convexo.
- Y debe ser un conjunto discreto finito.
- Las funciones $f : \mathbb{R}^n \to \mathbb{R}$ y $g : \mathbb{R}^n \to \mathbb{R}^p$ deben ser convexas en $\mathbf{x}$.
- Las funciones f y g deben ser al menos una vez continuas y diferenciables
- Ya que se trata de un problema de descomposición, en el subproblema NLP, al fijar las variables enteras $\mathbf{y}$, debe cumplir al menos una restricción, como por ejemplo, la condición de Slater, la cual exige la existencia de un punto estrictamente factible (que satisfaga las desigualdades de forma estricta) y, en problemas convexos, garantiza dualidad fuerte, es decir, que el valor óptimo del problema primal coincide con el de su problema dual.

El método de Outer Approximation (OA) descompone el MINLP en dos subproblemas que comparten información de forma iterativa. Estos subproblemas, llamados primal y maestro, separan las variables continuas no lineales de las variables enteras. En el problema primal, las variables enteras se fijan como parámetros dados $\mathbf{y} = \mathbf{y}^k$ (provenientes de una solución factible), lo que permite resolver un problema no lineal continuo y obtener una cota superior del problema. Por otro lado, en el problema maestro la parte no lineal se aproxima mediante cortes lineales externos (de ahí el nombre del método); como estos cortes acotan la región factible desde el exterior, la solución obtenida corresponde a una

relajación del problema y proporciona una cota inferior. Ambos subproblemas, de forma iterativa, añaden nuevos cortes lineales que refinan la aproximación de la región factible y descartan soluciones enteras que no mejoran el resultado, hasta que la diferencia entre la cota superior y la inferior (GAP) queda dentro de una tolerancia aceptable.

2.5.1.1. Problema Primal

Floudas (1995) define el problema primal de la siguiente forma:

$$\text{PP}(\mathbf{y}^k): \quad \min_{x} \quad c^T\mathbf{y}^k + f(\mathbf{x}) \tag{2.26}$$

$$\text{s.t.} \quad g(\mathbf{x}) + B\mathbf{y}^k \leq 0 \tag{2.27}$$

$$\mathbf{x} \in X, \tag{2.28}$$

donde las variables $\mathbf{y} = \mathbf{y}^k$, con k contador de la iteración, corresponden a valores fijos de combinaciones binarias. El resultado de este problema puede ser o no factible, para lo cual se realiza lo siguiente:

Primal Factible

Si el problema primal es factible en la iteración k-ésima, su solución entra información sobre el óptimo $\mathbf{x}^k$, $f(\mathbf{x}^k)$ y obtenemos la siguiente cota superior $UB = c^T\mathbf{y}^k + f(\mathbf{x}^k)$. Con esta información, podemos linealizar sobre $\mathbf{x}^k$, las funciones convexas $f(\mathbf{x})$ y $g(\mathbf{x})$ y se satisfacen las siguientes relaciones:

$$f(\mathbf{x}) \geq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)\left(\mathbf{x} - \mathbf{x}^k\right), \quad \forall \mathbf{x}^k \in X, \tag{2.29}$$

$$g(\mathbf{x}) \geq g(\mathbf{x}^k) + \nabla g(\mathbf{x}^k)\left(\mathbf{x} - \mathbf{x}^k\right), \quad \forall \mathbf{x}^k \in X. \tag{2.30}$$

dada la convexidad de $f(\mathbf{x})$ y $g(\mathbf{x})$

Primal Infactible

Si el problema primal es infactible en la iteración k-ésima, entonces es necesario identificar un punto factible examinando el conjunto de restricciones:

$$g(\mathbf{x}) + B\mathbf{y}^k \leq 0, \tag{2.31}$$

y formulando un problema de factibilidad de manera similar a como se hace en el método GBD (Floudas, 1995)

Por ejemplo, si utilizamos la minimización l_1, se tiene

$$\min_{\mathbf{x}\in X} \quad \sum_{j=1}^{p} a_j \tag{2.32}$$

$$\text{s.t.} \quad g_j(\mathbf{x}) + B\mathbf{y}^k \leq a_j, \quad j = 1, 2, \ldots, p \tag{2.33}$$

$$a_j \geq 0 \tag{2.34}$$

Su solución proporcionará el punto correspondiente $\mathbf{x}^l$, a partir del cual podemos linealizar las restricciones:

$$g(\mathbf{x}) \geq g(\mathbf{x}^l) + \nabla g(\mathbf{x}^l)(\mathbf{x} - \mathbf{x}^l), \quad \forall \mathbf{x}^l, \tag{2.35}$$

donde el lado derecho constituye un soporte lineal válido.

2.5.1.2. Problema Maestro

Las dos ideas principales del problema maestro en OA son, la proyección de (2.22) en el espacio de $\mathbf{y}$ y la aproximación externa de la función objetivo y la región factible.

La proyección de (2.22) se puede escribir como:

$$\min_{y} \quad \inf_{x} c^T\mathbf{y} + f(\mathbf{x}) \tag{2.36}$$

$$\text{s.t.} \quad g(\mathbf{x}) + B\mathbf{y} \leq 0 \tag{2.37}$$

$$\mathbf{x} \in \mathbf{X} \tag{2.38}$$

$$\mathbf{y} \in \mathbf{Y}. \tag{2.39}$$

Nótese que el problema interno se escribe como el ínfimo con respecto a $\mathbf{x}$ para cubrir el caso en que exista una solución no acotada para un $\mathbf{y}$ fijo. Nótese también que $c^T\mathbf{y}$ puede sacarse fuera del ínfimo, ya que es independiente de $\mathbf{x}$.

Definamos $v(\mathbf{y})$:

$$v(\mathbf{y}) = \quad c^T\mathbf{y} + \inf_x f(\mathbf{x}) \tag{2.40}$$

$$\text{s.t.} \quad g(\mathbf{x}) + B\mathbf{y} \leq 0 \tag{2.41}$$

$$b\mathbf{x} \in \mathbf{X}, \tag{2.42}$$

donde $v(\mathbf{y})$ es la parametrización en la variable $\mathbf{y}$ y corresponde al óptimo del problema (2.22) para la variable $\mathbf{y}$ fija. Se define además el conjunto $\mathbf{V}$ de los $\mathbf{y}$ para los cuales existe una solución factible en las variables $\mathbf{x}$:

$$\mathbf{V} = \{\mathbf{y} : g(\mathbf{x}) + B\mathbf{y} \leq 0, \ \forall \mathbf{x} \in \mathbf{X}\}, \tag{2.43}$$

entonces el problema (2.22) puede ser escrito ahora como:

$$\min_y \quad v(\mathbf{y}) \tag{2.44}$$

$$\text{s.t.} \mathbf{y} \in \mathbf{Y} \cap \mathbf{V}. \tag{2.45}$$

Este problema es la proyección de P en el espacio de $\mathbf{y}$. El cual tiene que satisfacer las condiciones de factibilidad, lo cual es impuesto en su restricción.

Ahora bien, la aproximación externa de $v(\mathbf{y})$ se expresa en términos de la intersección de un conjunto infinito de funciones de soporte. Estas funciones de soporte corresponden a las linealizaciones de $f(\mathbf{x})$ y $g(\mathbf{x})$ en todos los puntos $\mathbf{x}^k \in X$. Entonces, se satisfacen las siguientes condiciones:

$$f(\mathbf{x}) \geq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^l)\left(\mathbf{x} - \mathbf{x}^k\right) \quad \forall \mathbf{x}^k \in X, \tag{2.46}$$

$$g(\mathbf{x}) \geq g(\mathbf{x}^k) + \nabla g(\mathbf{x}^l)\left(\mathbf{x} - \mathbf{x}^k\right) \quad \forall \mathbf{x}^k \in X, \tag{2.47}$$

debido al supuesto de convexidad y de diferenciabilidad continua. $\nabla f(\mathbf{x}^k)$ representa el vector gradiente de dimensión n de $f(\mathbf{x})$, y $\nabla g(\mathbf{x}^k)$ es la matriz jacobiana de dimensión $(n \times p)$ evaluada en $\mathbf{x}^k \in X$.

De esta forma, el problema maestro queda definido como:

$$\text{MP:} \quad \min_{\mathbf{x},\mathbf{y},\mu_{OA}} \quad c^T\mathbf{y} + \mu_{OA} \tag{2.48}$$

$$\text{s.t.} \quad \mu_{OA} \geq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k), \quad \forall k \in F \tag{2.49}$$

$$0 \geq g(\mathbf{x}^k) + \nabla g(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) + B\mathbf{y}, \quad \forall k \in F \tag{2.50}$$

$$\mathbf{x} \in X \tag{2.51}$$

$$\mathbf{y} \in Y \cap V, \tag{2.52}$$

donde

$$\mathbf{F} = \{k : \mathbf{x}^k \text{ solución factible del problema primal } PP(\mathbf{y}^k)\}. \tag{2.53}$$

Duran and Grossmann (1986) suponen que se puede reemplazar $\mathbf{y} \in Y \cap V$ por $\mathbf{y} \in Y$, argumentando que una representación de las restricciones $\mathbf{y} \in Y \cap V$ queda implícita en las linealizaciones del problema (2.48), siempre que se introduzcan los cortes enteros apropiados para excluir la posibilidad de generar las mismas combinaciones enteras.

2.5.2. Branch & Cut

Quesada and Grossmann (1992) proponen una extensión del método de *Outer Approximation* en la cual la generación de cortes se integra directamente dentro de un esquema de *Branch and Bound*. Este enfoque es conocido como algoritmo *LP/NLP Branch-and-Cut*. La motivación principal surge de la observación de que, en el método OA clásico, el problema maestro MILP debe resolverse repetidamente desde cero cada vez que se agregan nuevos cortes de aproximación externa (Melo et al., 2020).

En contraste, el enfoque de *Branch & Cut* incorpora estos cortes de manera dinámica durante la exploración del árbol de *branch-and-bound*. De esta forma, el problema maestro no se resuelve como una secuencia de MILP independientes, sino como una única relajación MILP que se va refinando progresivamente mediante la incorporación de linealizaciones obtenidas al resolver subproblemas NLP.

Sea L el conjunto de puntos de linealización generados hasta una iteración dada. En cada nodo del árbol correspondiente a una partición Y^k del espacio de las variables enteras,

se resuelve la siguiente relajación continua del problema maestro:

$$\tilde{M}_L(Y^k): \quad \min_{\mathbf{x},\mathbf{y},\mu} \quad c^T\mathbf{y} + \mu \tag{2.54}$$

$$\text{s.t.} \quad \mu \geq f(\mathbf{x}^l) + \nabla f(\mathbf{x}^l)(\mathbf{x} - \mathbf{x}^l), \quad \forall l \in L \tag{2.55}$$

$$0 \geq g(\mathbf{x}^l) + \nabla g(\mathbf{x}^l)(\mathbf{x} - \mathbf{x}^l) + B\mathbf{y}, \quad \forall l \in L \tag{2.56}$$

$$\mathbf{x} \in X \tag{2.57}$$

$$\mathbf{y} \in Y^k. \tag{2.58}$$

Si la solución obtenida para $\mathbf{y}$ es fraccional, el algoritmo procede a ramificar siguiendo la estrategia estándar de *branch-and-bound*. Sin embargo, si se obtiene una solución entera $\mathbf{y}^k$, se resuelve el subproblema NLP correspondiente:

$$\hat{P}_{\mathbf{y}^k}: \quad \min_{\mathbf{x}\in X} \quad c^T\mathbf{y}^k + f(\mathbf{x}) \tag{2.59}$$

$$\text{s.t.} \quad g(\mathbf{x}) + B\mathbf{y}^k \leq 0. \tag{2.60}$$

Dependiendo de la factibilidad de este subproblema, se generan nuevos cortes de aproximación externa que se agregan al conjunto L. Estos cortes se incorporan al problema maestro como restricciones adicionales y se imponen no solo en el nodo actual, sino también en todos los nodos futuros del árbol. De esta manera, la aproximación lineal del MINLP se vuelve progresivamente más precisa a medida que avanza la exploración del árbol.

En el esquema de *branch-and-bound*, las podas se realizan por infactibilidad o por cota. Cuando se encuentra una solución entera, el algoritmo entra en un ciclo en el cual el problema maestro se actualiza con nuevas linealizaciones y se resuelve nuevamente hasta que se obtiene una solución fraccional o se demuestra que la solución actual no puede mejorar la cota superior existente.

Desde el punto de vista computacional, esta estrategia permite evitar el reinicio repetido del problema MILP que ocurre en el método OA clásico. Además, los solucionadores modernos de MILP permiten implementar este algoritmo de manera eficiente mediante el uso de *user constraints* y *lazy constraints* y funciones *callback*, que permiten agregar

dinámicamente los cortes de aproximación externa durante la ejecución del procedimiento de *branch-and-bound*.

2.6. Resolución computacional

Para poder afrontar el problema es necesario ejecutarlo en un programa (llamados solvers) que sea capaz de resolver este tipo de problemas y modelos, para lo cual se genera una tabla comparativa para ver las distintas capacidades de estos.

Tabla 2.1: Tabla Solvers MINLP.

Solver MINLP	Global Solution (Para Non-Convex)	Soporta SOCP	Soporta GSL
Bonmin (Bonami et al., 2008)	Óptimo Local	Si	Si
SCIP (Hojny et al., 2025)	Óptimo Local	Si	No
Couenne (Belotti et al., 2009)	Si	Si	No
Knitro (Byrd et al., 2006)	Óptimo Local	Si	Si
Baron (Sahinidis, 1996)	Si	Si	No
Lindo Global (Lin and Schrage, 2009)	Si	Si	No
Xpress (Belotti et al., 2025)	Óptimo Local	Si	No
RAPOSa (González-Rodríguez et al., 2023)	Si	No	No
Octeract (Octeract Team, 2024)	Si	No	No
MOSEK (Andersen and Andersen, 2000)	Si	Si	No

La tabla anterior compara diferentes solvers MINLP en cuanto a su capacidad para encontrar soluciones globales en problemas no convexos, y su soporte para restricciones de tipo SOCP y para trabajar con GNU Scientific Library (GSL) que se necesita en particular para trabajar con las variables estocásticas que resultan del desarrollo de las funciones no triviales en los modelos. Algunos solvers como Couenne, Baron y Lindo Global garantizan soluciones globales, mientras que otros como Bonmin y SCIP solo alcanzan óptimos locales. Además, la mayoría soporta restricciones SOCP, pero solo unos pocos pueden trabajar con GSL. Dado esto, se decidió utilizar únicamente los solvers Bonmin y Knitro para abordar los problemas MINLP, soportando un benchmark específico con estos dos

solvers. Bonmin fue seleccionado porque implementa nativamente los algoritmos B-OA (Algoritmo de descomposición basado en Outer Approximation) y B-QG (Algoritmo de descomposición basado en Branch & Cut de Quesada and Grossmann (1992)), lo que permite comparar el desempeño de estos enfoques con el problema a trabajar. Knitro, por su parte, es ampliamente reconocido por su eficiencia y robustez en la resolución de problemas de optimización no lineal. Utiliza dos algoritmos principales para abordar problemas MINLP: el algoritmo de ramificación y acotación no lineal (NLPBB, por sus siglas en inglés) y el algoritmo de programación cuadrática secuencial mixto-entero (MISQP, Mixed-Integer Sequential Quadratic Programming). Estos algoritmos permiten a Knitro manejar eficientemente la naturaleza combinatoria y no lineal de los problemas MINLP. Esto mencionado es para resolver el modelo convexo sin descomponer, para este caso, la descomposición divide el problema en uno del tipo NLP y otro del tipo MILP, los cuales son resueltos por los solvers Ipopt y Gurobi respectivamente, Ipopt soporta las funciones de GSL nativamente y Gurobi permite sin problema la integración de callbacks que permiten modificar el comportamiento del solver durante la resolución, es decir, añadir *user constraints* y *lazy constraints* que irán acotando la región factible durante el proceso de resolución.

3 | Formulación de los modelos

Basándose en el análisis anterior, el problema conjunto de ubicación e inventario bajo una política de revisión continua (r, Q) con restricciones de nivel de servicio fillrate y demanda distribuida normalmente puede formularse como un modelo de programación no lineal entera mixta (MINLP), caracterizado por restricciones no lineales y una función objetivo no lineal. El modelo resultante se denomina $\beta-$JLIP.

3.1. Joint location inventory problem

El problema conjunto de localización e inventario (JLIP) integra decisiones estratégicas de ubicación de instalaciones con decisiones operacionales de políticas de inventario. El modelo busca determinar simultáneamente dónde instalar los centros de distribución, qué retail asignar a cada centro, y los parámetros óptimos de las políticas de inventario para cada instalación.

Para la formulación del modelo se definen las siguientes variables de decisión: X_j como variable binaria que indica si se instala un centro de distribución en el sitio candidato j; Y_{ij} como variable binaria que representa la asignación del retail i al centro de distribución j; r_j como el punto de reorden para cada centro j; y Q_j como el tamaño del lote para cada centro j.

El modelo resultante corresponde a un problema de programación no lineal entera mixta (MINLP) que minimiza los costos totales del sistema sujeto a restricciones de asignación, capacidad y nivel de servicio, bajo una política de inventario de revisión continua (r, Q).

3.1.1. Función de costo

El objetivo es minimizar el costo total por unidad de tiempo en cada CD j, el cual es la suma de: el costo fijo de instalación de cada CD, el costo de transporte por unidad desde la planta al CD, el costo de transporte por unidad desde cada CD j a los retailers, el costo de pedido y el costo de mantenimiento. Entonces la función de costo es:

$$AC_j(Q_j, r_j, X_j, \mathbf{Y}) = f_j X_j + a_j \sum_{i \in I} \mu_i Y_{ij} + \sum_{i \in I} d_{ij} \mu_i Y_{ij} + S_j \frac{\sum_{i \in I} \mu_i Y_{ij}}{Q_j} + h_j OH_j^{\infty}. \quad (3.1)$$

El primer término en (3.1), $f_j X_j$, representa el costo fijo por unidad de tiempo de apertura y operación de un CD en el sitio j, donde X_j es una variable binaria que indica si el CD está instalado. El segundo término, $a_j \sum_{i \in I} \mu_i Y_{ij}$, es el costo de transporte por unidad de tiempo desde la planta hasta el CD_j, considerando la tasa de demanda μ_i de cada tienda retail i servido por el CD_j. El tercer término, $\sum_{i \in I} d_{ij} \mu_i Y_{ij}$, representa el costo de transporte por unidad de tiempo desde el CD_j hasta los clientes que sirve. El cuarto término, $S_j \frac{\sum_{i \in I} \mu_i Y_{ij}}{Q_j}$, corresponde al costo de pedido por unidad de tiempo, donde la demanda total servida por el CD_j se repone en lotes de tamaño Q_j. El quinto término, $h_j OH_j^{\infty}$, es una aproximación del costo de mantenimiento por unidad de tiempo, donde OH_j^{∞} denota el inventario promedio disponible en el CD_j en estado estacionario. Siguiendo a Daskin et al. (2002), este nivel de inventario se aproxima al inventario promedio mantenido en el centro de distribución.

3.1.2. Inventario on-hand en estado estable

En una política de inventario de revisión continua (r, Q) con full backorders y tiempo de entrega determinístico L_j en cada centro de distribución candidato CD_j, la dinámica del inventario puede describirse mediante la identidad:

$$IL_j^+(t + L_j) = IP_j(t) - D_j(t, t + L_j) + IL_j^-(t + L_j), \quad (3.2)$$

porque $IL_j(t + L_j) = IP_j(t) - D_j(t, t + L_j)$ donde:

- $IL_j^+(t + L_j)$ es el inventario disponible en el tiempo $t + L_j$,

- $IP_j(t)$ es la posición del inventario en el tiempo t,
- $D_j(t, t + L_j)$ es la demanda total durante el intervalo de tiempo de entrega $(t, t + L_j]$ para cualquier CD_j candidato,
- $IL_j^-(t + L_j)$ es el nivel de backorders en el tiempo $t + L_j$ y,
- $IL_j(t + L_j)$ es el nivel de inventario en el tiempo $t + L_j$.

Asumiendo que la demanda total sigue un proceso de incremento estacionario, escribimos $D_j(t, t + L_j) := D_j(L_j)$, y bajo el supuesto de demanda independiente y normalmente distribuida, entonces: $D_j(L_j) \sim N\left(L_j \sum_{i\in I} \mu_i Y_{ij},\ L_j \sum_{i\in I} \sigma_i^2 Y_{ij}\right)$. En estado estacionario, la posición del inventario $IP_j(t)$ se distribuye uniformemente sobre el intervalo $[r_j, r_j + Q_j]$ (Zipkin, 1986), es decir: $IP_j(t) \sim U[r_j, r_j + Q_j]$.

Por lo tanto, el inventario disponible esperado en estado estacionario en el CD candidato j está dado por:

$$OH_j^\infty = \frac{Q_j}{2} + r_j - L_j \sum_{i\in I} \mu_i Y_{ij} + B_j^\infty, \tag{3.3}$$

donde $OH_j^\infty = \mathbb{E}(IL_j^+(t + L_j))$ y $B_j^\infty = \mathbb{E}(IL_j^-(t + L_j))$ el número esperado de pedidos pendientes en estado estacionario.

3.1.3. Backorders en estado estable

Bajo el supuesto de que la demanda está normalmente distribuida, el número esperado de pedidos pendientes en estado estacionario en el CD candidato j está dado por:

$$B_j^\infty = B\left(Q_j, r_j, \mathbf{Y}\right) = L_j \frac{\sum_{i\in I} \sigma_i^2 Y_{ij}}{Q_j} \left[H\left(\frac{r_j - L_j \sum_{i\in I} \mu_i Y_{ij}}{\sqrt{L_j \sum_{i\in I} \sigma_i^2 Y_{ij}}} \right) - H\left(\frac{r_j + Q_j - L_j \sum_{i\in I} \mu_i Y_{ij}}{\sqrt{L_j \sum_{i\in I} \sigma_i^2 Y_{ij}}} \right) \right], \tag{3.4}$$

donde $H(x) = \frac{1}{2}\left((x^2 + 1)(1 - \Phi(x)) - x\varphi(x)\right)$ con $\Phi(x)$ y $\varphi(x)$ las funciones de distribución y densidad de la distribución normal estándar, respectivamente. Debe notarse que $B(Q_j, r_j, \mathbf{Y}_j) \geq 0$ porque $H(x)$ es una función positiva y decreciente (Axsäter, 2006).

3.1.4. Restricción de nivel de servicio

La escasez de inventario es inherente en sistemas con demanda incierta. Para gestionar esto, adoptamos el nivel de servicio fill-rate introducido por Schneider (1981), que, bajo demanda normalmente distribuida, corresponde a la probabilidad de mantener inventario positivo en cualquier momento dado (Silver et al., 1998; Axsäter, 2015; Escalona et al., 2021). Esta medida controla efectivamente tanto el tamaño de los pedidos pendientes como la frecuencia de quiebres de stock. Sea $\beta(Q_j, r_j, \mathbf{Y})$ la tasa de llenado en el centro de distribución CD_j bajo una política de revisión continua (r, Q) con pedidos pendientes completos. Siguiendo a Axsäter (2006), está dada por:

$$\beta\left(Q_j, r_j, \mathbf{Y}\right) = 1 - \frac{\sqrt{L_j \sum_{i\in I} \sigma_i^2 Y_{ij}}}{Q_j}\left[G\left(\frac{r_j - L_j \sum_{i\in I} \mu_i Y_{ij}}{\sqrt{L_j \sum_{i\in I} \sigma_i^2 Y_{ij}}}\right) - G\left(\frac{r_j + Q_j - L_j \sum_{i\in I} \mu_i Y_{ij}}{\sqrt{L_j \sum_{i\in I} \sigma_i^2 Y_{ij}}}\right)\right], \tag{3.5}$$

donde $G(x) = \int_x^\infty (v - x)\varphi(v)dv = \varphi(x) - x(1 - \Phi(x))$ con $\Phi(x)$ y $\varphi(x)$ las funciones de distribución y densidad de la distribución normal estándar, respectivamente. También nótese que $H'(x) = -G(x)$.

Un fill-rate preestablecido mínimo $\bar{\beta} \in [0, 1]$ se impone para cada CD_j para asegurar el nivel de servicio deseado, es decir, $\beta(Q_j, r_j, \mathbf{Y}) \geq \bar{\beta}$ para todo $j \in J$.

3.1.5. Formulación matemática del modelo β-JLIP

Integrando todos los componentes desarrollados anteriormente —función de costo, inventario en estado estable, backorders y restricciones de nivel de servicio— se obtiene la formulación completa del problema conjunto de localización e inventario con restricciones

de fill-rate. El modelo matemático resultante se expresa como:

$$\beta\text{-JLIP:} \quad \min_{\mathbf{X},\mathbf{Y},\mathbf{r},\mathbf{Q}} \quad \sum_{j\in J}\left\{ f_j X_j + \sum_{i\in I} \hat{d}_{ij} Y_{ij} + S_j \frac{\sum_{i\in I}\mu_i Y_{ij}}{Q_j} + h_j\left(\frac{Q_j}{2} + r_j - L_j \sum_{i\in I}\mu_i Y_{ij} + B(Q_j, r_j, \mathbf{Y_j})\right)\right\} \tag{3.6}$$

$$\text{s.t.} \quad \sum_{j\in J} Y_{ij} = 1 \qquad \forall i \in I \tag{3.7}$$

$$Y_{ij} \le X_j \qquad \forall i \in I,\ \forall j \in J \tag{3.8}$$

$$\beta(Q_j, r_j, Y) \ge \bar{\beta} \qquad \forall j \in J \tag{3.9}$$

$$r_j \ge L_j \sum_{i\in I}\mu_i Y_{ij} \qquad \forall j \in J \tag{3.10}$$

$$Q_j \ge 0 \qquad \forall j \in J \tag{3.11}$$

$$X_j,\ Y_{ij} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J. \tag{3.12}$$

La restricción (3.7) asegura que cada tienda retail sea asignado a exactamente un centro de distribución, lo que garantiza exclusividad en la asignación. La restricción (3.8) asegura que un cliente solo puede ser asignado a un centro de distribución si ese centro está operativo, es decir, ha sido instalado. La restricción de nivel de servicio (3.9) impone que el nivel de servicio fill-rate en cada centro de distribución debe ser mayor o igual que el nivel preestablecido $\bar{\beta} \in [0, 1]$ para toda la red, asegurando así un nivel de servicio satisfactorio. La restricción (3.10) introduce un requerimiento de stock de seguridad, asegurando que el punto de reorden en cada centro de distribución sea lo suficientemente alto para cubrir la demanda esperada durante el tiempo de entrega. La restricción (3.11) refleja la condición natural del modelo al requerir que las cantidades de pedido sean no negativas. Finalmente, la restricción (3.12) impone la integralidad de las variables de decisión, requiriendo que las decisiones de localización y asignación consistan en determinar si se instala un centro o si se asigna un retail, según corresponda.

3.2. Reformulación convexa

Siguiendo un enfoque similar al planteado por Atamtürk et al. (2012), se realiza una reformulación convexa que consiste en introducir variables auxiliares y restricciones cónicas con el objetivo de representar ciertas expresiones cuadráticas del modelo en una forma que preserve la convexidad del problema.

En primer lugar, se redefinen algunos términos para simplificar la estructura del modelo. El *safety stock* se define como $v_j := r_j - L_j \sum_{i \in I} \mu_i Y_{ij}$, el cual representa el nivel de inventario de seguridad efectivo en la instalación j una vez considerada la demanda esperada de los clientes asignados a dicha instalación. Asimismo, el costo total de transporte esperado se redefine como $\hat{d}_{ij} := (a_j + d_{ij})\mu_i$, que abarca los costos de transporte desde la planta hacia los centros, y desde los centros a los retailers.

Posteriormente, se introducen las variables auxiliares $C1_j$ y $C2_j$ para representar términos cuadráticos que dependen de las variables de asignación. En particular, se definen mediante las siguientes restricciones cónicas:

$$C1_j = \sqrt{\sum_{i \in I} \mu_i Y_{ij}^2} \qquad \forall j \in J \tag{3.13}$$

$$C2_j = \sqrt{\sum_{i \in I} (\sigma_i Y_{ij})^2} \qquad \forall j \in J \tag{3.14}$$

$$C1_j \geq 0 \qquad \forall j \in J \tag{3.15}$$

$$C2_j \geq 0 \qquad \forall j \in J. \tag{3.16}$$

Estas variables permiten representar la norma cuadrática asociada a las asignaciones de demanda y su variabilidad. Al expresar dichos términos mediante variables auxiliares, las expresiones cuadráticas que aparecen en la función objetivo y en las restricciones pueden incorporarse al modelo mediante restricciones cónicas convexas.

En particular, las restricciones

$$C1_j \geq \sqrt{\sum_{i \in I} \mu_i Y_{ij}^2} \tag{3.17}$$

$$C2_j \geq \sqrt{\sum_{i \in I} (\sigma_i Y_{ij})^2}, \tag{3.18}$$

constituyen una representación equivalente en términos de conos de segunda orden.

Finalmente, reformulando el problema (3.6) añadiendo estas restricciones y cambios de variable.

$$\beta\text{-CQLIP:} \min_{\substack{\mathbf{X},\mathbf{Y},\mathbf{v},\mathbf{Q},\\ \mathbf{C1},\mathbf{C2}}} \sum_{j \in J} \left\{ f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + S_j \frac{C1_j^2}{Q_j} + h_j \left(\frac{Q_j}{2} + v_j + \hat{B}(v_j, Q_j, C2_j) \right) \right\} \tag{3.19}$$

$$\text{s.t.} \quad \sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{3.20}$$

$$Y_{ij} \leq X_j \qquad \forall i \in I, \forall j \in J \tag{3.21}$$

$$b(v_j, Q_j, C2_j) \leq 0 \qquad \forall j \in J \tag{3.22}$$

$$C1_j \geq \sqrt{\sum_{i \in I} \mu_i Y_{ij}} \qquad \forall j \in J \tag{3.23}$$

$$C2_j \geq \sqrt{\sum_{i \in I} (\sigma_i Y_{ij})} \qquad \forall j \in J \tag{3.24}$$

$$v_j, Q_j \geq 0 \qquad \forall j \in J \tag{3.25}$$

$$C1_j, C2_j \geq 0 \qquad \forall j \in J \tag{3.26}$$

$$X_j, Y_{ij} \in \{0, 1\} \qquad \forall i \in I, \forall j \in J, \tag{3.27}$$

donde $b(v_j, Q_j, C2_j)$ es la restricción que representa el nivel de servicio dada por:

$$b(v_j, Q_j, C2_j) = -\sqrt{L_j} C2_j (1 - \bar{\beta}) + \frac{L_j C2_j^2}{Q_j} \left[G\left(\frac{v_j}{\sqrt{L_j} C2_j} \right) - G\left(\frac{v_j + Q_j}{\sqrt{L_j} C2_j} \right) \right]. \tag{3.28}$$

Finalmente, se puede observar que el problema reformulado es convexo, ya que sus términos son lineales o cónicos convexos: las restricciones de asignación y no negatividad

son lineales, las restricciones cuadráticas pueden expresarse como conos de segundo orden, y la restricción asociada al nivel de servicio se apoya en una función convexa demostrada por Axsäter (2006).

Se presenta en el anexo un glosario de los conjuntos, variables y parámetros de los problemas presentados, junto con también la reformulación de las funciones $\hat{\beta}(v_j, Q_j, C2_j)$ y $\hat{B}(v_j, Q_j, C2_j)$.

3.3. Método de descomposición

Tal como se describió en el marco teórico para los métodos de *Outer Approximation* y *Branch & Cut*, la estrategia de solución adoptada para el modelo β-CQLIP se basa en un mismo esquema de descomposición primal-maestro. La idea central consiste en separar las decisiones discretas de localización y asignación, representadas por $(\mathbf{X}, \mathbf{Y})$, de las decisiones continuas asociadas al control de inventario, representadas por $(\mathbf{v}, \mathbf{Q}, \mathbf{C1}, \mathbf{C2})$. Esta separación permite explotar la estructura del problema: al fijar las variables binarias, el modelo resultante en las variables continuas conserva la parte no lineal del inventario y del nivel de servicio, mientras que el problema maestro mantiene las decisiones combinatorias y aproxima la componente no lineal mediante linealizaciones.

En consecuencia, el procedimiento iterativo opera en dos etapas. Primero, para una solución entera dada $(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$, se resuelve un subproblema primal de tipo NLP que entrega una solución continua factible y una cota superior del problema original. Luego, a partir de esa solución se construyen cortes de aproximación externa sobre la función de costo no lineal $g(\cdot)$ y sobre la restricción de nivel de servicio $b(\cdot)$, los cuales se incorporan al problema maestro. El problema maestro, formulado sobre las variables de localización, asignación y variables auxiliares, entrega una nueva combinación candidata y una cota inferior.

Es importante destacar que este es el mismo esquema utilizado en ambos algoritmos considerados en este trabajo, OA y Branch & Cut. La diferencia entre ellos no radica en la definición del subproblema primal ni del problema maestro, sino en la forma en que el maestro es resuelto y actualizado: en OA, el problema maestro se resuelve de manera

secuencial agregando cortes entre iteraciones; en Branch & Cut, esos mismos cortes se integran dinámicamente dentro del árbol de *branch-and-bound*. Por ello, la formulación que se presenta a continuación constituye la base común sobre la cual operan ambos enfoques.

3.3.1. Problema Primal (NLP)

El problema primal corresponde a la etapa continua de la descomposición, donde las variables de localización y asignación permanecen fijas. En este subproblema se optimizan únicamente las variables asociadas al inventario y a la restricción de nivel de servicio. De esta forma, se obtiene una solución continua factible asociada a la combinación entera fijada, que luego se evaluará iterativamente en el método.

El problema primal es el siguiente:

$$\text{PP:} \quad \min_{\mathbf{v},\mathbf{Q},\mathbf{C1},\mathbf{C2}} \sum_{j\in J}\left\{f_j X_j^{(k)} + \sum_{i\in I}\hat{d}_{ij}Y_{ij}^{(k)} + g(v_j,Q_j,C1_j,C2_j)\right\} \tag{3.29}$$

$$\text{s.t.} \quad b(v_j,Q_j,C2_j) \le 0 \qquad \forall j\in J \tag{3.30}$$

$$C1_j \ge \sqrt{\sum_{i\in I}\mu_i Y_{ij}^{(k)}} \qquad \forall j\in J \tag{3.31}$$

$$C2_j \ge \sqrt{\sum_{i\in I}\sigma_i^2 Y_{ij}^{(k)}} \qquad \forall j\in J \tag{3.32}$$

$$v_j, Q_j \ge 0 \qquad \forall j\in J \tag{3.33}$$

$$C1_j, C2_j \ge 0 \qquad \forall j\in J. \tag{3.34}$$

En esta formulación, la función $g(v_j,Q_j,C1_j,C2_j)$ representa de manera compacta toda la parte no lineal de la función objetivo asociada al control de inventario en cada centro j, es decir:

$$g(v_j,Q_j,C1_j,C2_j) = S_j\frac{C1_j^2}{Q_j} + h_j\left(\frac{Q_j}{2} + v_j + \hat{B}(v_j,Q_j,C2_j)\right), \tag{3.35}$$

los términos que dependen simultáneamente de las variables continuas del problema una vez fijadas las variables enteras. Escribiéndola en esta forma compacta se facilita la construcción de los cortes de aproximación externa, ya que *Outer Approximation* linealiza precisamente esta función en los puntos generados por el problema primal para incorporarlos al problema

maestro.

3.3.2. Problema maestro (MILP)

El problema maestro concentra las decisiones de localización y asignación, manteniendo explícitas las variables enteras del modelo. Su función es incorporar progresivamente cortes de aproximación externa que representen la parte no lineal del problema, representadas por W_j. Así, entrega nuevas combinaciones candidatas para refinar la búsqueda de la solución óptima.

$$\text{MP:} \quad \min_{\mathbf{X},\mathbf{Y},\mathbf{v},\mathbf{Q},\mathbf{C1},\mathbf{C2},\mathbf{W}} \sum_{j\in J}\left\{f_j X_j + \sum_{i\in I}\hat{d}_{ij}Y_{ij} + W_j\right\} \tag{3.36}$$

$$\begin{aligned}\text{s.t.} \quad W_j \geq\ & g(v_j^{(k)}, Q_j^{(k)}, C1_j^{(k)}, C2_j^{(k)}) + \frac{\partial g^{(k)}}{\partial v_j}(v_j - v_j^{(k)}) \\ & + \frac{\partial g^{(k)}}{\partial Q_j}(Q_j - Q_j^{(k)}) + \frac{\partial g^{(k)}}{\partial C1_j}(C1_j - C1_j^{(k)}) \\ & + \frac{\partial g^{(k)}}{\partial C2_j}(C2_j - C2_j^{(k)}) \qquad \forall j \in J,\ k \in \mathcal{K}\end{aligned} \tag{3.37}$$

$$\begin{aligned}0 \geq\ & b(v_j^{(k)}, Q_j^{(k)}, C2_j^{(k)}) + \frac{\partial b^{(k)}}{\partial v_j}(v_j - v_j^{(k)}) \\ & + \frac{\partial b^{(k)}}{\partial Q_j}(Q_j - Q_j^{(k)}) + \frac{\partial b^{(k)}}{\partial C2_j}(C2_j - C2_j^{(k)}) \\ & \qquad \forall j \in J,\ k \in \mathcal{K}\end{aligned} \tag{3.38}$$

$$\sum_{j\in J} Y_{ij} = 1 \qquad \forall i \in I \tag{3.39}$$

$$Y_{ij} \leq X_j \qquad \forall i \in I,\ \forall j \in J \tag{3.40}$$

$$C1_j^2 \geq \sum_{i\in I}\mu_i Y_{ij}^2 \qquad \forall j \in J \tag{3.41}$$

$$C2_j^2 \geq \sum_{i\in I}(\sigma_i Y_{ij})^2 \qquad \forall j \in J \tag{3.42}$$

$$v_j, Q_j \geq 0 \qquad \forall j \in J \tag{3.43}$$

$$C1_j, C2_j \geq 0 \qquad \forall j \in J \tag{3.44}$$

$$X_j, Y_{ij} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J. \tag{3.45}$$

En el problema maestro, la variable auxiliar W_j es una variable de epígrafo que actúa como sustituto lineal de la función no lineal $g(v_j, Q_j, C1_j, C2_j)$ en la función objetivo: en lugar de optimizar directamente sobre $g(\cdot)$, se minimiza W_j sujeto a que W_j sea siempre mayor o igual que la aproximación lineal de $g(\cdot)$, garantizando así que el valor óptimo de W_j converja al valor real de $g(\cdot)$ a medida que se acumulan cortes de *Outer Approximation*.

Se define $\mathcal{K}$ como el conjunto de cortes lineales (hiperplanos) que aproximan la función no lineal $g(\cdot)$ en el problema maestro. Cada elemento $k \in \mathcal{K}$ corresponde a un corte generado a partir de un punto de evaluación $(v_j^{(k)}, Q_j^{(k)}, C1_j^{(k)}, C2_j^{(k)})$ del problema primal y tiene la forma

$$W_j \geq g(v_j^{(k)}, Q_j^{(k)}, C1_j^{(k)}, C2_j^{(k)}) + \nabla g^{(k)\top} \left(\begin{pmatrix} v_j \\ Q_j \\ C1_j \\ C2_j \end{pmatrix} - \begin{pmatrix} v_j^{(k)} \\ Q_j^{(k)} \\ C1_j^{(k)} \\ C2_j^{(k)} \end{pmatrix} \right). \tag{3.46}$$

El conjunto exacto de todos los cortes que representan $g(\cdot)$ puede ser infinito (o tan grande que su representación explícita resulta intratable), por lo que en la práctica el maestro se resuelve con un subconjunto finito e incompleto $\mathcal{K}' \subset \mathcal{K}$, el cual se denomina como un conjunto relajado. Este conjunto no reproduce exactamente $g(\cdot)$, pero resulta imprescindible en la práctica debido a limitaciones computacionales. Esta relajación resulta entonces en un conjunto reducido y de cortes que es suficiente para alcanzar una solución de calidad aceptable según criterio de parada. La generación de cortes se puede detener cuando se cumple un criterio de convergencia ϵ o cuando la adición de nuevos cortes no aporta mejoras significativas a las cotas; en OA, en cada iteración se añade para todo $j \in J$ con $X_j = 1$ el corte lineal de $g(\cdot)$ evaluado en la solución del primal sobre W_j.

Las dos primeras familias de restricciones corresponden precisamente a dichos cortes lineales generados por el método de *Outer Approximation*. La primera familia entrega una aproximación lineal inferior de la parte no lineal de la función objetivo, a partir de la linealización de $g(\cdot)$ en los puntos obtenidos desde el problema primal. La segunda familia hace lo mismo para la restricción no lineal de nivel de servicio, reemplazando $b(\cdot)$ por sus hiperplanos de soporte.

El primer paso del procedimiento iterativo (contador de iteraciones $k = 0$) consiste en resolver un UFLP tal como está en (2.15) para obtener una configuración factible inicial de la red de distribución en términos de variables de localización-asignación. Sea $(X^{(0)}, Y^{(0)})$ y Z^*_{UFLP} las variables óptimas y la función objetivo del UFLP. Entonces, establecemos la cota inferior $LB^{(0)} = Z^*_{UFLP}$ y resolvemos el problema primal $P(X^{(0)}, Y^{(0)})$, a partir del cual obtenemos una solución óptima para $(v^{(0)}, Q^{(0)}, C1^{(0)}, C2^{(0)})$.

Se fija la mejor cota superior como ínf $UB = UB^{(0)} = Z^{(0)}_P$, donde $Z^{(0)}_P$ es el valor óptimo de la función objetivo de $P(X^{(0)}, Y^{(0)})$, y el contador de iteraciones se actualiza a $k = k + 1$.

El segundo paso del procedimiento iterativo consiste en resolver el problema maestro con $\mathcal{K} = \{0, \ldots, k-1\}$, denominado problema maestro relajado (RMP). Sea $(X^{(k)}, Y^{(k)})$ la solución óptima del problema maestro relajado y establecemos $LB^{(k)} = Z^*_{RMP}$ como la nueva cota inferior actual, donde Z^*_{RMP} es el valor óptimo de la función objetivo del modelo RMP.

Además, $(X^{(k)}, Y^{(k)})$ es el siguiente punto a considerar en el problema primal $P(X^{(k)}, Y^{(k)})$, del cual obtenemos una solución óptima para $(v^{(k)}, Q^{(k)}, C1^{(k)}, C2^{(k)})$ y la cota superior actual, es decir, $UB^{(k)} = Z^{(k)}_P$, donde $Z^{(k)}_P$ es el valor óptimo de la función objetivo de $P(X^{(k)}, Y^{(k)})$.

Si la cota superior actual es estrictamente menor que la mejor cota superior encontrada hasta el momento, entonces ínf $UB = UB^{(k)}$. Por lo tanto, si $\frac{\text{ínf } UB - LB^{(k)}}{\text{ínf } UB} \leq \epsilon$, entonces el algoritmo termina y ínf UB es una función objetivo ϵ-óptima para β-CQLIP, es decir, ínf $UB \leq Z^*_{\beta\text{-}CQLIP} = Z^*_{\beta\text{-}JLIP}$. De lo contrario, $k = k + 1$ y se regresa al paso 2.

Bajo la terminología de OA, en este punto se cerraría la iteración k y se pasaría a resolver un nuevo RMP con el conjunto de cortes $\mathcal{K}$ actualizado. En *Branch & Cut*, en cambio, esa actualización no ocurre en un ciclo externo maestro-primal, sino dentro del árbol de *branch-and-bound*: cuando aparece una solución entera candidata para $(\mathbf{X}, \mathbf{Y})$, se resuelve $P(X, Y)$, se generan cortes OA y estos se incorporan globalmente a la relajación lineal del nodo. Así, la lógica de OA se mantiene, pero embebida en el árbol: la cota inferior proviene de las relajaciones lineales de los nodos activos y la cota superior de la mejor solución incumbente factible.

4 | Experimentos Computacionales y comentarios

4.1. Metodología

Los experimentos computacionales se realizan con el fin de evaluar el desempeño de las formulaciones propuestas y de las estrategias de solución consideradas. En particular, se compara el comportamiento del modelo convexo β-CQLIP con el enfoque de descomposición basado en la separación entre un problema primal y un problema maestro.

En primer lugar, se resuelve β-CQLIP, el cual se utiliza como modelo de control o referencia. Este modelo proporciona una base de comparación que permite evaluar tanto la calidad de las soluciones obtenidas como el desempeño computacional de los métodos de descomposición que se aplican posteriormente.

A continuación, se considera la formulación descompuesta del problema, estructurada a partir del problema primal PP (3.29) y el problema maestro MP (3.36). Esta formulación se resuelve inicialmente mediante un esquema de Outer Approximation (OA), el cual permite aproximar progresivamente la región factible del problema original mediante cortes lineales agregados al problema maestro. Posteriormente, se incorpora un procedimiento de Branch-and-Cut (B&C) que opera sobre una única instancia del árbol de *branch-and-bound*, de tal modo que se espera obtener un desempeño computacional superior al de OA.

Se espera que B&C sea más eficiente que OA al mantener un único árbol de *branch-and-bound* que se va actualizando dinámicamente con cortes, en lugar de resolver múltiples instancias del MILP (RMP) en iteraciones externas. Los cortes se generan de dos formas:

lazy cuts en nodos enteros los cuales verifican que las soluciones candidatas cumplan los cortes de aproximación externa y *user cuts* en nodos fraccionarios que fortalecen la relajación lineal mejorando cotas duales, reforzando optimalidad. Esta integración de cortes en el árbol permite una exploración más eficiente del espacio de soluciones.

En cuanto a la implementación computacional, los modelos de programación no lineal (NLP) se resuelven utilizando IPOPT 3.12.13, mientras que los modelos de programación lineal entera mixta (MILP) se resuelven con Gurobi 13.0, y como se mencionó anteriormente, el modelo convexo sin descomponer se resuelven utilizando Bonmin 1.8.9 y Knitro 14.2.0. Todos los experimentos se ejecutan en un servidor equipado con un procesador Intel Xeon Gold a 2.10 GHz con 40 núcleos y 32 GB de memoria RAM.

Las instancias de prueba utilizadas en los experimentos corresponden a redes reportadas en Daskin (2011), las cuales consideran configuraciones de 49, 88 y 150 nodos. En estas instancias, cada nodo representa simultáneamente a un minorista y a un sitio candidato para la instalación de un centro de distribución, lo que refleja una estructura típica en problemas de localización y diseño de redes logísticas, que consideran grupos de ciudades. Los datos proporcionados por Daskin (2011) corresponden a coordenadas de ciudades ubicadas en distintos estados de Estados Unidos, distribuidas a lo largo del país. Debido a esto, las distancias consideradas se encuentran aproximadamente en un rango entre 0 y 4.500 km. Sin embargo, se tratan simplemente como unidades de distancia en una escala estándar.

Finalmente, para todos los experimentos se establece un criterio de parada basado en una tolerancia de optimalidad de $\epsilon = 10^{-4}$ y un tiempo máximo de resolución de 7200 segundos para cada instancia. En el caso de Branch & Cut, existe un criterio de parada adicional, controlado por el mismo solver, que es cuando se explora el árbol completo de Branch & Bound dentro del tiempo límite, es decir, explora todas las soluciones factibles sin encontrar una mejora que minimice más la ínfima solución óptima obtenida.

4.2. Data de prueba

Con el fin de tener una amplia muestra de resultados, tales que, describan de manera correcta sus resultados, se generaron 200 instancias aleatorias para cada conjunto de nodos.

La data de prueba se genera como lo realizan en Escalona et al. (2024), para definir criterios parametros y parámetros comunes. La base temporal definida para el conjunto de pruebas es el mes, la demanda por unidad de tiempo en cada tienda retail se distribuye normalmente con media $\mu_i = U[3000, 12000]$ para cualquier $i \in I$, y con coeficiente de variación $CV_i = U[0.1, 0.5]$ para cualquier $i \in I$. Cabe señalar que la distribución normal es una buena aproximación de la demanda no negativa cuando $CV \leq 0.5$ (Peterson and Silver, 1979), caso en el cual la probabilidad de que la demanda sea menor que cero es inferior a 0.0228. El nivel de servicio preestablecido para la red de distribución es $\bar{\beta} = U[0.75, 0.99]$. El costo fijo de instalar un centro de distribución (DC) en el sitio candidato j (por unidad de tiempo) es $f_j = U[4000, 8000]$ para cualquier $j \in J$. El costo de transporte por unidad desde la planta hasta el DC candidato j es $a_j = a = U[0.3, 0.7]$ para cualquier $j \in J$. El costo de transporte por unidad desde el DC j hasta el retail i, d_{ij}, es igual a la distancia entre el DC j y el retail i multiplicada por una tarifa de transporte $c_{ij} = c = U[0.001, 0.01]$ para cualquier $i \in I$ y $j \in J$. El costo de mantenimiento por unidad y por unidad de tiempo en el DC j es $h_j = h = U[0.25, 1.25]$ para cualquier $j \in J$. El costo de ordenar en el DC j es $S_j = S = U[100, 500]$ para cualquier $j \in J$, y el tiempo de entrega es $L_j = U[0.2, 0.6]$ para cualquier $j \in J$.

4.3. Resultados

En primera instancia se intentó resolver el modelo convexo β-CQLIP de forma directa con solvers como Knitro y Bonmin; sin embargo, para instancias de mayor tamaño surgieron problemas de convergencia numérica y consumo excesivo de recursos, lo que impidió obtener soluciones óptimas sin exceder el tiempo límite.

Se continúa Para poder comparar los resultados del modelo β-CQLIP resuelto mediante los métodos de descomposición Outer Aproximation (OA) y Branch & Cut (BC) se utilizarán 2 métricas: el tiempo computacional empleado en resolver una instancia y el GAP relativo, que se calculan a partir de la mejor solución factible obtenida (Inf_UB) y la mejor solución relajada del problema (LB), las cuales se obtienen del problema primal y maestro

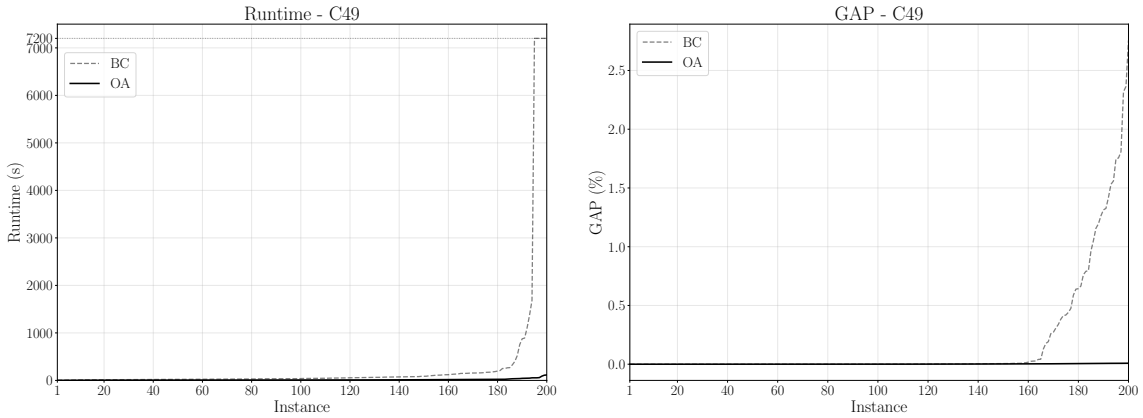
respectivamente.

$$GAP(\%) = \frac{|Inf\_UB - LB|}{Inf\_UB} \cdot 100. \tag{4.1}$$

Por otro lado, se realiza un perfil de rendimiento (Dolan and Moré, 2002) que permite un análisis global de los resultados. Representa gráficamente la proporción de instancias resueltas en un factor del mejor tiempo de solución. Se define el tiempo de solución t_{ps} de una instancia p resuelta por el método s y t_p^* el mejor tiempo de solución para la instancia p. Entonces el algoritmo de rendimiento es representado por:

$$\rho_s(\tau) = \frac{\text{Número de instancias: } t_{ps} \leq \tau t_p^*}{\text{Número total de instancias}}. \tag{4.2}$$

Para todos los resultados se ordenan las instancias de forma ascendente para ambas métricas, para tener una visualización de la cantidad de instancias que se resolvieron y su calidad de solución, los gráficos se presentan a continuación:

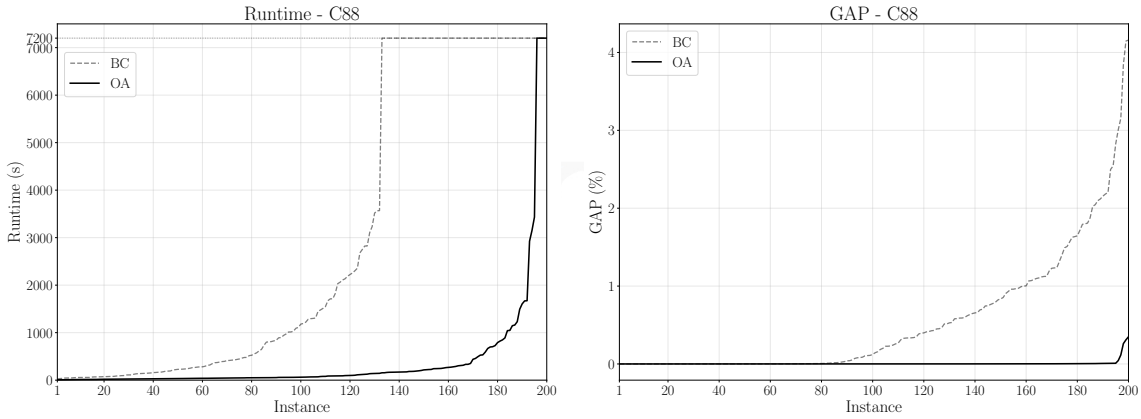


(a) Runtime configuración 49 nodos. **(b)** Gap relativo configuración 49 nodos.

Figura 4.1: Resultados GAP y Runtime para la configuración de 49 nodos.
Fuente: Elaboración propia.

Para la configuración de 49 nodos, con Outer Approximation todas las instancias (200) fueron resueltas antes del tiempo límite de 2 horas (100 %) y ninguna mostró un GAP superior a la tolerancia $\epsilon = 10^{-4}$ (0 % de las instancias). El GAP máximo observado en OA fue $8.5 \cdot 10^{-3}$ %. En Branch & Cut, 6 instancias (3.0 % del total de instancias) no fueron resueltas por tiempo límite, 35 instancias (17.5 %) presentaron GAP mayor que ϵ de las cuales 29 instancias (14.5 %) lo hicieron sin alcanzar el tiempo límite. El GAP máximo

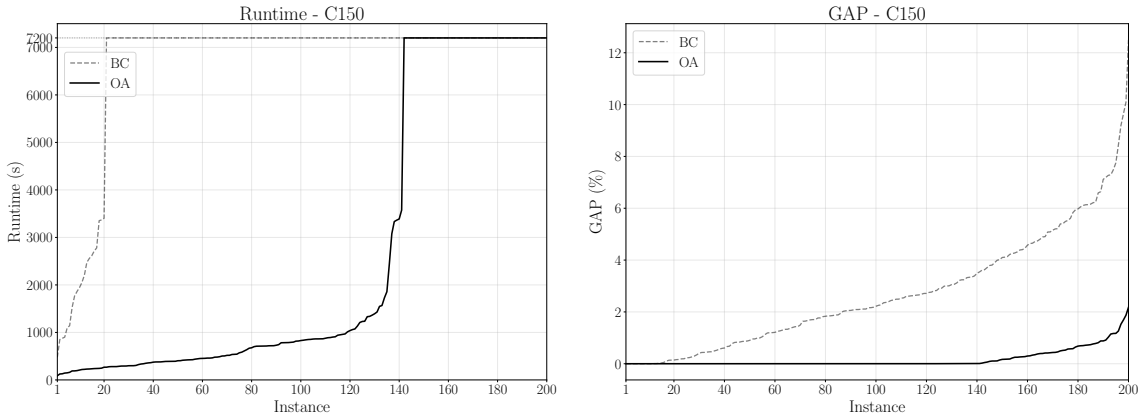
observado en BC fue 2.76 %.



(a) Runtime configuración 88 nodos. **(b)** Gap relativo configuración 88 nodos.

Figura 4.2: Resultados GAP y Runtime para la configuración de 88 nodos.
Fuente: Elaboración propia.

Para la configuración de 88 nodos, en Outer Approximation 5 instancias (2.5 %) no fueron resueltas antes del tiempo límite y 4 instancias (2.0 %) presentaron GAP mayor que ϵ, con un GAP máximo en OA de 0.35 %. En Branch & Cut, 68 instancias (34.0 %) no fueron resueltas por tiempo límite y 104 instancias (52.0 %) exhibieron GAP mayor que ϵ de las cuales 45 instancias (22.5 %) lo hicieron sin alcanzar el tiempo límite. El GAP máximo en BC fue 4.1590 %.



(a) Runtime configuración 150 nodos. **(b)** Gap relativo configuración 150 nodos.

Figura 4.3: Resultados GAP y Runtime para la configuración de 150 nodos.
Fuente: Elaboración propia.

Para la configuración de 150 nodos, en Outer Approximation 59 instancias no fueron resueltas antes del tiempo límite de 2 horas, y 55 instancias mostraron un GAP superior

a la tolerancia de 10^{-4}, alcanzando un GAP máximo de 2.18 %. En Branch & Cut, 180 instancias no fueron resueltas por tiempo límite y 184 presentaron GAP superior a la tolerancia, con un GAP máximo de 12.50 %.

Los resultados se resumen en la siguiente tabla comparativa que cuenta la cantidad de instancias que cumplen superan tiempo máximo de resolución o cuales :

Config.	Método	$t \geq t_{max}$	$GAP > \epsilon$	BC: $GAP > \epsilon \wedge t < t_{max}$	Inst. resueltas (%)	GAP_{max}(%)
49	OA	0	0	–	100.0	0.0085
49	BC	6	35	29	97.0	2.7566
88	OA	5	4	–	97.5	0.3489
88	BC	68	104	45	66.0	4.1590
150	OA	59	55	–	70.5	2.1760
150	BC	180	184	7	10.0	12.4978

Tabla 4.1: Resumen de métricas por configuración y método.

Para analizar el desempeño computacional de los métodos evaluados se presenta a continuación los gráficos de rendimiento. Estos gráficos muestran la proporción de instancias resueltas en función del factor respecto al mejor tiempo, entregando una curva creciente que permite comparar los métodos. Se elige un umbral de $\tau = 50$ para poder ver con claridad la tendencia respecto de las proporciones de cada método, es decir, el factor de holgura respecto del mejor tiempo para cada instancia.

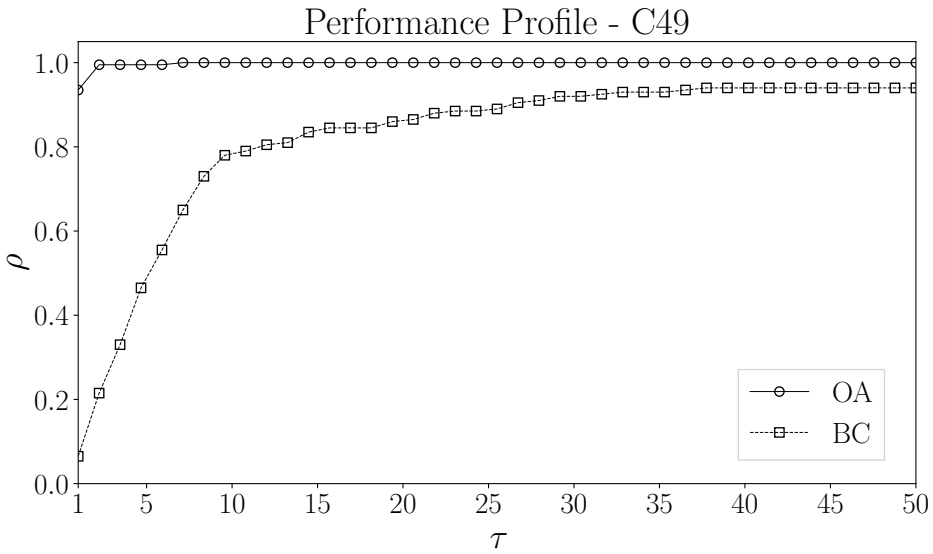


Figura 4.4: Perfil Rendimiento 49 nodos.
Fuente: Elaboración propia.

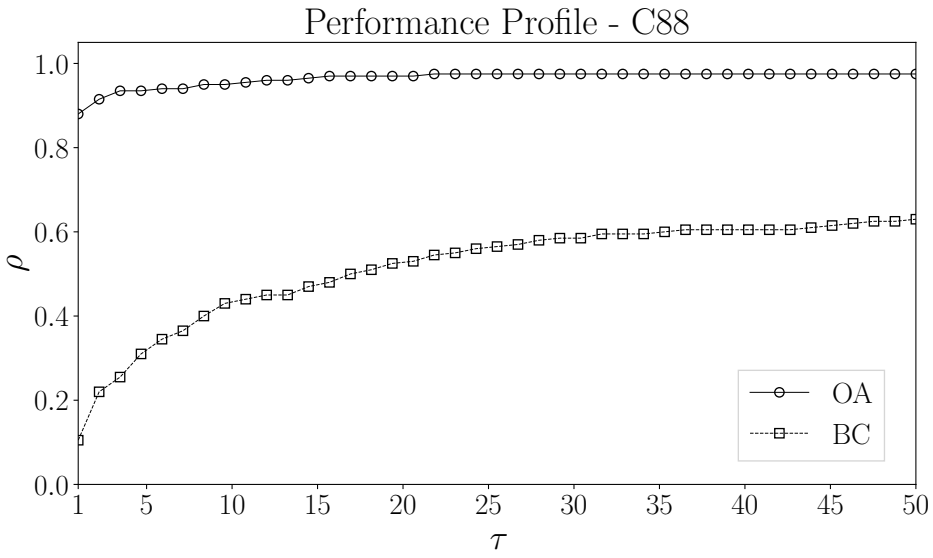


Figura 4.5: Perfil Rendimiento 88 nodos.
Fuente: Elaboración propia.

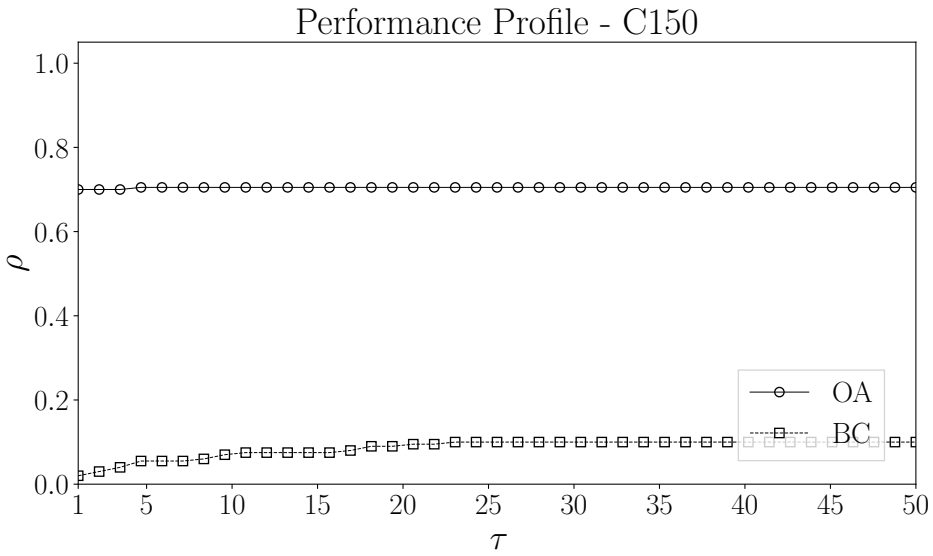


Figura 4.6: Perfil Rendimiento 150 nodos.
Fuente: Elaboración propia.

Se observa una clara predominancia de Outer Approximation por sobre Branch & Cut. En particular, cuando $\tau = 1$ se aprecia la proporción de instancias en las que cada método fue el más rápido, y en todas las configuraciones Outer Approximation presenta un porcentaje mayor de instancias resueltas más rápidamente. El parámetro τ puede interpretarse como un umbral de holgura respecto del mejor tiempo observado: a medida que τ aumenta, la

curva muestra la proporción de instancias que realmente logró resolver cada método dentro de ese factor del mejor tiempo. Por ejemplo, para la configuración de 49 nodos, cuando $\tau = 10$ Branch & Cut alcanza solo un 80 % de instancias resueltas, mientras que Outer Approximation ya resuelve el 100 %.

La siguiente tabla presenta la proporción de cada componente de costo respecto del costo total:

Config.	FC (%)	TC (%)	OC (%)	HC (%)	BO (%)	HC sin BO (%)
49	27.84	56.21	4.32	11.64	0.78	10.86
88	27.37	57.17	4.39	11.07	0.78	10.29
150	18.25	67.22	4.15	10.39	0.76	9.63

Tabla 4.2: Proporción de costos por sobre el total.

Esta tabla resume la participación relativa de los distintos componentes del costo total en cada configuración. Como se mencionó con anterioridad, el costo total se compone del fixed cost, transportation cost, ordering cost, holding cost, y backorders, que se incluyen dentro del holding cost al reflejar el costo de pedidos no atendidos oportunamente.

Las principales diferencias se pueden observar en cuanto al costo fijo de instalación y costo de transporte, ya que en términos de proporción, el resto de los costos no muestran gran diferencia. Se observa que a mayor cantidad de nodos, el costo de instalar se vuelve muy relevante, ya que puede ser más costosa la instalación de nuevos centros de distribución que destinar los recursos para su transporte.

Por otro lado, el porcentaje asociado a los backorders resulta muy bajo en comparación con las demás partidas de costo, lo cual se explica porque la restricción de nivel de servicio fill-rate limita la acumulación de pedidos pendientes y obliga al modelo a mantener un alto cumplimiento de la demanda (Escalona et al., 2024). En consecuencia, los atrasos tienen una participación despreciable dentro del costo total y no alteran de manera significativa la estructura general del sistema, ni las decisiones de localización o inventario.

5 | Conclusiones

En este trabajo se cumplieron los objetivos planteados, obteniendo resultados relevantes tanto en capacidad de resolución como en el análisis comparativo de métodos para el problema estudiado. Respecto a la formulación del modelo, se muestra una formulación que preserva la convexidad, por lo que es candidata de ser abordada mediante otros métodos para resolver MINLP.

En cuanto a las contribuciones de este trabajo, se muestra una mejora respecto a trabajos previos como el presentado en Escalona et al. (2024), ya que se consiguió resolver instancias de hasta 150 nodos, superando el límite de 88 nodos de la publicación.

En la evaluación comparativa entre Branch & Cut (B&C) y Outer Approximation (OA), los resultados no confirman la ventaja esperada de B&C. A pesar de que estudios previos (Melo et al., 2022) que sostienen la eficacia de B&C en problemas similares, en este caso el procedimiento de separación en el problema primal resulta costoso en tiempo. Además, los cortes generados pueden estar eliminando soluciones factibles relevantes debido a la naturaleza combinatoria del problema y a limitaciones numéricas/computacionales, lo cual puede aumentar los tiempos totales de resolución.

Hay que notar que para cada CD_j, existen las curvas $g(v_j, Q_j, C1_j, C2_j)$ y $b(v_j, Q_j, C2_j)$ que son aquellas que se quieren aproximar. Pero al resolver el problema de forma iterativa, solo se están aproximando cuando un centro es instalado, es decir, se calcula un conjunto de cortes de aproximación cada vez que $X_j = 1$, dejando aquellos centros no instalados de lado. Esto genera una oscilación sistemática de ir añadiendo dichos cortes. Para Outer Approximation por sí solo esto no genera mayor inconveniente porque resuelve el problema maestro y subproblema siempre a optimalidad por lo que convergerá eventualmente, ya que añade iterativamente los cortes para las soluciones enteras. Pero para Branch & Cut entra en

juego la parte fraccionaria de las soluciones, es decir, en su resolución encuentra soluciones relajadas del problema maestro en las cuales se buscan añadir cortes de factibilidad y optimalidad, pero estos no aportan información relevante lo cual retrasa su búsqueda, y por ende su tiempo de resolución. Es decir, Outer Approximation dentro del esquema del Branch & Cut no es bueno porque, cuando no hay cortes en los centros de distribución, genera la oscilación mencionada y en los que si hay, no están mejorando la relajación al no aportar información, es decir, no propone una buena relajación del problema porque los cortes añadidos son de mala calidad y en su búsqueda de mejorar la solución, tomará más tiempo. Esto podría solucionarse añadiendo de forma previa cortes de buena calidad que acoten el problema para que no se pierda tiempo en su búsqueda, en otras palabras, mejorar la calidad de su partida en caliente (warmstart), que muchos solvers actuales permiten esa opción.

Por otra parte, en cuanto a los resultados se refuerza el supuesto de trabajo de que, bajo las restricciones de fill-rate consideradas, los backorders resultan despreciables desde la perspectiva de la formulación y los resultados numéricos obtenidos.

A pesar de las limitaciones del B&C en este estudio, la calidad de las soluciones obtenidas no fueron del todo mal: tomando a OA como el método de referencia, el gap máximo observado fue de 2.1 % en el peor de los casos en la configuración de nodos más grande, lo que indica soluciones cercanas al óptimo bajo las condiciones evaluadas.

Líneas futuras: conviene desarrollar un enfoque de Branch & Cut específico para este tipo de problemas que controle mejor la generación y aplicación de cortes previos, o bien explorar cortes alternativos y métodos mixtos. Entre las posibilidades se sugieren: métodos híbridos de aproximación interna y externa, cortes pre-calculados basados en ECP similares a los empleados en trabajos sobre variantes relacionadas como en Escalona et al. (2026). Además, en este trabajo se consideró la demanda con distribución Normal; en investigaciones futuras conviene evaluar escenarios con demanda Gamma u otras distribuciones, o bien desarrollar un enfoque genérico que abarque distintas familias de distribución.

Bibliografía

Aljunaidi, A. and Ankrak, S. (2014). The application of lean principles in the fast moving consumer goods (FMCG). *J. Oper. Supply Chain Manag.*, 7(2):1–25.

Andersen, E. D. and Andersen, K. D. (2000). *The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm*, pages 197–232. Springer US, Boston, MA.

Andersson, J. and Melchiors, P. (2001). A two-echelon inventory model with lost sales. *International Journal of Production Economics*, 69(3):307–315.

Atamtürk, A., Berenguer, G., and Shen, Z.-J. (2012). A conic integer programming approach to stochastic joint location-inventory problems. *Operations Research*, 60(2):366–381.

Axsäter, S. (2006). A simple procedure for determining order quantities under a fill rate constraint and normally distributed lead-time demand. *European journal of operational research*, 174(1):480–491.

Axsäter, S. (2015). *Inventory control*, volume 225. Springer.

Axsäter, S. and Zhang, W.-F. (1999). A joint replenishment policy for multi-echelon inventory control. *International Journal of Production Economics*, 59:243–250.

Bala, M. and Kumar, D. (2011). Supply chain performance attributes for the fast moving consumer goods industry. *Journal of transport and supply chain management*, 5(1):23–38.

Balinski, M. L. (1964). On finding integer solutions to linear programs. Technical report, Proceedings of the I.B.M. Scientific Computing Symposium on Combinatorial Problems.

Belotti, P., Berthold, T., Gally, T., Gottwald, L., and Pólik, I. (2025). Solving minlps to global optimality with fico® xpress global. *Optimization*, 0(0):1–19.

Belotti, P., Lee, J., Liberti, L., Margot, F., and Wächter, A. (2009). Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4-5):597–634.

Berman, O., Krass, D., and Tajbakhsh, M. (2012). A coordinated location-inventory model. *European Journal of Operational Research*, 217:500–508.

Bonami, P., Biegler, L. T., Conn, A. R., Cornuéjols, G., Grossmann, I. E., Laird, C. D., Lee, J., Lodi, A., Margot, F., Sawaya, N., et al. (2008). An algorithmic framework for convex mixed integer nonlinear programs. *Discrete optimization*, 5(2):186–204.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brown, R. G. (1967). *Decision rules for inventory management*. New York, Holt, Rinehart and Winston [1967].

Byrd, R. H., Nocedal, J., and Waltz, R. A. (2006). K nitro: An integrated package for nonlinear optimization. *Large-scale nonlinear optimization*, pages 35–59.

Chung, K.-J., Ting, P.-S., and Hou, K.-L. (2009). A simple cost minimization procedure for the (q,r) inventory system with a specified fixed cost per stockout occasion. *Appl. Math. Model.*, 33(5):2538–2543.

Daskin, M., Coullard, C., and Shen, Z. (2002). An inventory-location model: Formulation, solution algorithm and computational results. *Annals of Operations Research*, 110:83–106.

Daskin, M. S. (2011). *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons.

Daskin, M. S. (2013). *Network and discrete location*. Wiley-Blackwell, Hoboken, NJ, 2 edition.

Diabat, A., Abdallah, T., and Henschel, A. (2015). A closed-loop location-inventory problem with spare parts consideration. *Computers & Operations Research*, 54:245–256.

Dolan, E. D. and Moré, J. J. (2002). Benchmarking optimization software with performance profiles. *Mathematical programming*, 91:201–213.

Duran, M. A. and Grossmann, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming*, 36(3):307–339.

Escalona, P., Angulo, A., Brotcorne, L., Fortz, B., and Tapia, P. (2024). Fill-rate service level constrained distribution network design. *International Transactions in Operational Research*, 31(1):5–28.

Escalona, P., Araya, D., Simpson, E., Ramirez, M., and Stegmaier, R. (2021). On the shortage control in a continuous review (Q, r) inventory policy using α_l service-level. *RAIRO-Operations Research*, 55(5):2785–2806.

Escalona, P., Brotcorne, L., Angulo, A., Wolf, N., and Mora, C. (2026). The uncapacitated joint location-inventory problem with shortage costs: a selective extended cutting plane algorithm. *International Transactions in Operational Research*.

Escalona, P., Marianov, V., Ordónez, F., and Stegmaier, R. (2018). On the effect of inventory policies on distribution network design with several demand classes. *Transportation Research Part E: Logistics and Transportation Review*, 111:229–240.

Escalona, P., Ordóñez, F., and Marianov, V. (2015). Joint location-inventory problem with differentiated service levels using critical level policy. *Transportation Research Part E: Logistics and Transportation Review*, 83:141–157.

Estellés-Miguel, S., Cardós, M., Albarracín, J. M., and Palmer, M. E. (2014). Design of a continuous review stock policy. In *Annals of Industrial Engineering 2012*, pages 139–146. Springer London, London.

Estelles-Miguel, S., Cardós, M., Albarracín-Guillem, J., and Palmer-Gato, M. (2014). *Design of a Continuous Review Stock Policy*, pages 139–146. Elsevier.

Floudas, C. A. (1995). *Nonlinear and mixed-integer optimization: fundamentals and applications*. Oxford University Press.

González-Rodríguez, B., Ossorio-Castillo, J., González-Díaz, J., González-Rueda, Á. M., Penas, D. R., and Rodríguez-Martínez, D. (2023). Computational advances in polynomial optimization: RAPOSa, a freely available global solver. *J. Glob. Optim.*, 85(3):541–568.

Hojny, C., Besançon, M., Bestuzheva, K., Borst, S., Chmiela, A., Dionísio, J., Eifler, L., Ghannam, M., Gleixner, A., Göß, A., Hoen, A., van der Hulst, R., Kamp, D., Koch, T., Kofler, K., Lentz, J., Maher, S. J., Mexi, G., Mühmer, E., Pfetsch, M. E., Pokutta, S., Serrano, F., Shinano, Y., Turner, M., Vigerske, S., Walter, M., Weninger, D., and Xu, L. (2025). The SCIP Optimization Suite 10.0. Technical report, Optimization Online.

Lewis (1998). *Demand forecasting and inventory control*. John Wiley & Sons, Nashville, TN, 1998 edition.

Lin, Y. and Schrage, L. (2009). The global solver in the lindo api. *Optimization Methods and Software*, 24(4-5):657–668.

Liu, K., Zhou, Y., and Zhang, Z. (2010). Capacitated location model with online demand pooling in a multi-channel supply chain. *European Journal of Operational Research*, 207(1):218–231.

Manatkar, R. P., Karthik, K., Kumar, S. K., and Tiwari, M. K. (2016). An integrated inventory optimization model for facility location-allocation problem. *Int. J. Prod. Res.*, 54(12):3640–3658.

Melo, W., Fampa, M., and Raupp, F. (2020). An overview of minlp algorithms and their implementation in muriqui optimizer. *Annals of Operations Research*, 286(1):217–241.

Melo, W., Fampa, M., and Raupp, F. (2022). Two linear approximation algorithms for convex mixed integer nonlinear programming. *Annals of Operations Research*, 316(2):1471–1491.

Miranda, P. and Garrido, R. (2004). Incorporating inventory control decision into a strategic distribution network design model with stochastic demand. *Transportation Research Part E*, 40:183–207.

Miranda, P. and Garrido, R. (2009). Inventory service-level optimization within distribution network design problem. *International Journal of Production Economics*, 122:276–285.

Miranda, P. A. and Garrido, R. A. (2006). A simultaneous inventory control and facility location model with stochastic capacity constraints. *Networks and Spatial Economics*, 6(1):39–53.

Miranda, P. A. and Garrido, R. A. (2008). Valid inequalities for lagrangian relaxation in an inventory location problem with stochastic capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(1):47–65.

Nasiri, G. R., Kalantari, M., and Karimi, B. (2021). Fast-moving consumer goods network design with pricing policy in an uncertain environment with correlated demands. *Comput. Ind. Eng.*, 153(106997):106997.

Octeract Team (2024). Octeract engine. `https://octeract.gg/octeract-engine/`.

Ozsen, L., Coullard, C., and Daskin, M. (2008). Capacitated warehouse location model with risk pooling. *Naval Research Logistics*, 55:295–312.

Peterson, R. and Silver, E. A. (1979). Decision systems for inventory management and production planning. *(No Title)*.

Quesada, I. and Grossmann, I. E. (1992). An lp/nlp based branch and bound algorithm for convex minlp optimization problems. *Computers & chemical engineering*, 16(10-11):937–947.

Sahinidis, N. V. (1996). BARON: A general purpose global optimization software package. *J. Glob. Optim.*, 8(2):201–205.

Schneider, H. (1981). Effect of service-levels on order-points or order-levels in inventory models. *The International Journal of Production Research*, 19(6):615–631.

Shahabi, M., Unnikrishnan, A., Jafari-Shirazi, E., and Boyles, S. D. (2014). A three level location-inventory problem with correlated demand. *Transportation Research Part B: Methodological*, 69:1–18.

Shen, Z., Coullard, C., and Daskin, M. (2003). A joint location-inventory model. *Transportation Science*, 37:40–55.

Shen, Z.-J. M. (2005). A multi-commodity supply chain design problem. *IIE Transactions*, 37:753–762.

Shu, J., Teo, C.-P., and Shen, Z.-J. M. (2005). Stochastic transportation-inventory network design problem. *Operations Research*, 53(1):48–60.

Silver, E. A., Pyke, D. F., Peterson, R., et al. (1998). *Inventory management and production planning and scheduling*, volume 3. Wiley New York.

Snyder, L., Daskin, M., and Teo, C.-P. (2007). The stochastic location model with risk pooling. *European Journal of Operational Research*, 179:1221–1238.

Stollsteimer, J. (1961). *The Effect of Technical Change and Output Expansion on the Optimum Number, Size, and Location of Pear Marketing Facilities in a California Pear Producing Region*. University of California, Berkeley.

Tapia-Ubeda, F. J., Miranda, P. A., and Macchi, M. (2018). A generalized benders decomposition based algorithm for an inventory location problem with stochastic inventory capacity constraints. *European Journal of Operational Research*, 267(3):806–817.

Van Horenbeek, A., Buré, J., Cattrysse, D., Pintelon, L., and Vansteenwegen, P. (2013). Joint maintenance and inventory optimization systems: A review. *International Journal of Production Economics*, 143(2):499–508. Focusing on Inventories: Research and Applications.

Wang, L. and Chen, H. (2022). Optimization of a stochastic joint replenishment inventory system with service level constraints. *Comput. Oper. Res.*, 148(106001):106001.

Waters, D. (2003). *Inventory Control and Management*. John Wiley & Sons, Chichester, England, 2 edition.

You, F. and Grossmann, I. (2008). Mixed-integer nonlinear programming models and algorithms for large-scale supply chain design with stochastic inventory management. *Industrial & Engineering Chemistry Research*, 47:7802–7817.

Zhang, H. (1998). A note on the convexity of Service-Level measures of the (r, q) system. *Manage. Sci.*, 44(3):431–432.

Zipkin, P. (1986). Inventory service-level measures: convexity and approximation. *Management Science*, 32(8):975–981.

A | Glosario

Conjuntos	
I	Conjunto de tiendas retail minoristas, indexado por $i = 1, \ldots, \|I\|$.
J	Conjunto de centros de distribución candidatos, indexado por $j = 1, \ldots, \|J\|$.
Parámetros	
μ_i	Demanda media por unidad de tiempo del minorista i.
σ_i^2	Variación de la demanda por unidad de tiempo del minorista i.
f_j	Costo fijo por unidad de tiempo para instalar un centro de distribución en la ubicación j.
d_{ij}	Costo de transporte por unidad desde DC j hasta el minorista i.
a_j	Costo de transporte por unidad desde la planta hasta el centro de distribución j.
S_j	Costo del pedido en el centro de distribución j.
h_j	Costo de mantenimiento por unidad y unidad de tiempo en DC j.
$\bar{\beta}$	Parámetro establecido para restricción de fill-rate.
L_j	Plazo de entrega fijo en unidad de tiempo desde el proveedor hasta el centro de distribución j.
$\hat{d}_{ij}$	Costo total de transporte por unidad de tiempo definido como $\hat{d}_{ij} := (a_j + d_{ij})\mu_i$
Variables	
X_j	1 si el centro de distribución es instalado en el sitio j, 0 en otro caso.
Y_{ij}	1 si el centro de distribución j abastece tienda minorista i, 0 en otro caso.
Q_j	Cantidad de pedido fija en centro de distribución candidato j.
r_j	Punto de reorden en centro de distribución candidato j.
v_j	Inventario de seguridad en centro de distribución candidato j. Definido como $v_j := r_j - L_j \sum_{i \in I} \mu_i Y_{ij}$.

B | Reformulaciones

Fillrate service level constraint - reformulation

$$\hat{\beta}\left(v_j, Q_j, C2_j\right) = 1 - \frac{\sqrt{L_j}C2_j}{Q_j}\left[G\left(\frac{v_j}{\sqrt{L_j}C2_j}\right) - G\left(\frac{v_j + Q_j}{\sqrt{L_j}C2_j}\right)\right]$$

$$b(v_j, Q_j, C2_j) = -\sqrt{L_j}C2_j(1-\bar{\beta}) + \frac{L_jC2_j^2}{Q_j}\left[G\left(\frac{v_j}{\sqrt{L_j}C2_j}\right) - G\left(\frac{v_j + Q_j}{\sqrt{L_j}C2_j}\right)\right]$$

Backorders - reformulation

$$\hat{B}(v_j, Q_j, C2_j) = L_j\frac{C2_j^2}{Q_j}\left[H\left(\frac{v_j}{\sqrt{L_j}C2_j}\right) - H\left(\frac{v_j + Q_j}{\sqrt{L_j}C2_j}\right)\right]$$

C | Pseudo código - OA

Algorithm 1 OA Algorithm for JLIP with shortage costs and full-backorders

$k = 0$
$(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}) \leftarrow$ solve UFLP
Set $LB^{(0)} = Z^*_{UFLP}$
$(\mathbf{v}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{C1}^{(0)}, \mathbf{C2}^{(0)}), Z_P^{(0)} \leftarrow$ solve $P(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)})$
Set ínf $UB = UB^{(0)} = Z_P^{(0)}$
while $\frac{\text{ínf } UB - LB^{(k)}}{\text{ínf } UB} > \epsilon$ **do**
    $k = k + 1$
    $(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}), LB^{(k)} = Z^*_{RMP} \leftarrow$ solve RMP
    $(\mathbf{v}^{(k)}, \mathbf{Q}^{(k)}, \mathbf{C1}^{(k)}, \mathbf{C2}^{(k)}), UB^{(k)} = Z_P^{(k)} \leftarrow$ solve $P(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$
    if ínf $UB > UB^{(k)}$ **then**
        ínf $UB = UB^{(k)}$
    end if
end while
Set $r_j = v_j + L_j \sum_{i \in I} \mu_i Y_{ij} \quad \forall j \in J : X_j = 1$, and $r_j = 0 \quad \forall j \in J : X_j = 0$
Return $(\mathbf{X}, \mathbf{Y}, \mathbf{r}, \mathbf{Q})$, ínf UB

D | Pseudo código - B & C

Algorithm 2 Branch & Cut Algorithm for JLIP with shortage costs and full-backorders

$k = 0$
$(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}), Z^*_{UFLP} \leftarrow$ Solve UFLP
Set $LB^{(0)} = Z^*_{UFLP}$
$(\mathbf{v}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{C1}^{(0)}, \mathbf{C2}^{(0)}), Z^*_P, Cut^{(0)} \leftarrow$ Solve $P(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)})$
Set $\inf UB = UB^{(0)} = Z^*_P$
while $GAP^{(k)} = \frac{\inf UB - LB^{(k)}}{\inf UB} > \epsilon$ **do**
    $k = k + 1$
    Let $(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}), Z^{(k)}_{RMP}$ feasible integer solution and objective value of $RMP(\mathcal{K}_r^{(k-1)})$
    $(\mathbf{v}^{(k)}, \mathbf{Q}^{(k)}, \mathbf{C1}^{(k)}, \mathbf{C2}^{(k)}), Z^*_P, Cut^{(k)} \leftarrow$ Solve $P(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$
    if $Z^{(k)}_{RMP} > \inf UB$ **then**
        $LB^{(k)} = LB^{(k-1)}$
        $UB^{(k)} = \inf UB$
    else
        $LB^{(k)} = Z^{(k)}_{RMP}$
        $UB^{(k)} = Z^*_P$
        if $UB^{(k)} \leq \inf UB$ **then**
            $\inf UB = UB^{(k)}$
        end if
    end if
end while