



UNIVERSIDAD AUTÓNOMA METROPOLITANA

MAESTRÍA EN CIENCIAS

MATEMÁTICAS APLICADAS E INDUSTRIALES

*Modelos Matemáticos para Análisis de Demanda en Transporte*

PROYECTO DE INVESTIGACIÓN III

AVANCE DE ESCRITURA CORRESPONDIENTE AL TRIMESTRE 13–P

MARÍA VICTORIA CHÁVEZ HERNÁNDEZ

ASESOR:

LORENZO HÉCTOR JUÁREZ VALENCIA

MÉXICO D. F., FEBRERO 2014



# Dedicatoria

*A mis padres  
Alicia Hernández Cortés y Joel Chávez Martínez.*



# Agradecimientos

A través de estas líneas quiero expresar mi más sincero agradecimiento a todas las personas que con su apoyo hicieron posible la realización de este trabajo de investigación.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su apoyo y patrocinio para mis estudios de maestría, los cuales concluyen con la realización de éste proyecto de tesis.

Al Dr. Michael Florian y a Yolanda Noriega por sus valiosas enseñanzas, sugerencias, cuestionamientos y hospitalidad durante mi estancia en Montreal, así como por estar al pendiente de mi trabajo y responder mis dudas.

Al Dr. Héctor Juárez por el excelente asesoramiento que recibí en todas y cada una de las etapas del presente trabajo, por atender todas mis dudas, por revisar este manuscrito y por la confianza que tuvo en mí para desarrollar esta tesis y llevarla a buen término.

A mis sinodales el Dr. David Romero y al Dr. Joaquín Delgado por revisar el presente trabajo, así como también por sus importantes sugerencias, sus atinados comentarios y observaciones.

A Ana Fernández por tenerme la paciencia para enseñarme a usar EMME, por orientarme en detalles técnicos del programa, pero también por todas sus exposiciones, las cuales me ayudaron mucho para abordar el tema.

Al Ing. Pablo Torres por permitirme usar las instalaciones del metro y su licencia de EMME mientras estaba aprendiendo, también por todas sus valiosas observaciones durante las reuniones y su detalladas explicaciones de muchos conceptos.

Y cuando uno termina de agradecer a las instituciones y a las personas que nos guían durante el trabajo, no queda más que agradecer a todos aquéllos que nos apoyaron moralmente.

Por supuesto quiero agradecerle a toda mi familia, a mis papás a quiénes además les dedico el presente trabajo, pero además también quiero agradecerle a mis hermanos y hermanas a quienes nombro en orden par que no se sientan: Miguel, Alis, Lupita, Joelillo, Sarita y Fede

por echarme porras y estar siempre conmigo. Y ya que estamos en materia de familia, también agradezco a mis hermosos sobrinos y ahijados por alegrarme la vida con sus ocurrencias y por considerarme su tía y/o madrina favorita.

Y bueno, también es bueno tener ratos de ocio de vez en cuando, así que les agradezco a mis amigos Moc, Pau, Pablo, Checo, Isa y Fer por sus visitas que me distraían y me servían para recordarme que no estoy tan lejos de casa. A mis amigos y compañeros de maestría, Juan Luis, Victor, Marco, Marlene, Adals, Hector, Male, Lety, Raquel, Alfonso, y Liz por hacer mi paso por la maestría mucho más ameno y por todas esas idas a las pizzas y comidas con Margarito que disfrutamos juntos. A Gaby, Daniel y Miguel por sus consejos, experiencias, y ocurrencias que me hicieron disfrutar siempre de cada conversación y/o viaje que tuvimos juntos. Por supuesto no puedo olvidar a Jorge, quien me orientó enormemente en mis primeros trámites dentro de la UAM y por ser un gran amigo. A Lulú por todas sus atenciones, paseos juntas, salidas al café (las cuales extraño) y por orientarme para disfrutar mi estancia en Montreal.

Por último, pero no menos importante, a mi galán Sergio Gutiérrez a quien adoro y admiro por su carácter y paciencia, por se aguantador y cuidar de mis hijos mientras me fui a disfrutar la vida en Montreal... quise decir, mientras estuve haciendo mi estancia de investigación; pero también por soportar mis neurosis y ayudarme a tomarme la vida con menos seriedad y más calma.

También agradezco a todos aquellos que no pongo explícitamente en estas líneas, compañeros, maestros y familiares, porque con la ayuda de todos ustedes disfruté enormemente mi paso por la MCMAI. Me alegro y doy gracias a la vida por haberlos encontrado a todos ustedes en mi camino.

# Índice general

Dedicatoria	III
Agradecimientos	V
Nomenclatura	XI
Introducción	XIII
<b>1. Asignación de tránsito</b>	<b>1</b>
1.1. El modelo básico sin congestión . . . . .	2
1.2. Asignación con congestión . . . . .	9
<b>2. Estimación de matrices de demanda</b>	<b>11</b>
2.1. Modelos para estimación de matrices . . . . .	12
2.1.1. Modelo gravitacional . . . . .	12
2.1.2. Modelo de máxima entropía . . . . .	13
2.2. Métodos de balanceo de matrices . . . . .	16
2.2.1. Método biproportional estándar . . . . .	16
2.2.2. Método biproportional con cotas superiores . . . . .	17
2.2.3. Método triproporcional . . . . .	21
<b>3. Métodos de conteo</b>	<b>23</b>
3.1. Minimización de información . . . . .	24
3.2. Modelo de entropía considerando el flujo . . . . .	26
3.3. Mínimos cuadrados generalizados . . . . .	26
<b>4. Métodos iterativos para estimación de demanda</b>	<b>29</b>
4.1. El modelo de Spiess y el método de máximo descenso . . . . .	30
4.1.1. El modelo de Spiess . . . . .	30
4.1.2. El método de descenso . . . . .	31
4.1.3. Máximo descenso en el modelo que agrega la diferencia de demanda .	34
4.2. El método de gradiente conjugado multiplicativo . . . . .	35
4.2.1. Introducción del algoritmo de gradiente conjugado . . . . .	36

4.2.2.	Aspectos para construir del algoritmo multiplicativo . . . . .	37
4.2.3.	El algoritmo de gradiente conjugado multiplicativo . . . . .	39
4.2.4.	Aplicación del algoritmo de gradiente conjugado multiplicativo al problema de estimación de matrices de demanda . . . . .	40
<b>5.</b>	<b>Resultados</b>	<b>43</b>
5.1.	La raíz del error cuadrático medio y el coeficiente de correlación . . . . .	43
5.2.	Ejemplo de aplicación a la red de la ciudad de Winnipeg . . . . .	44
5.2.1.	Resultados con el método de balanceo biproporcional . . . . .	45
5.2.2.	Resultados con los métodos de descenso para ajuste de matrices . . . . .	48
<b>6.</b>	<b>Conclusiones y trabajo a futuro</b>	<b>61</b>
<b>A.</b>	<b>Algoritmos</b>	<b>65</b>
A.1.	Algoritmo de asignación para una red simple . . . . .	65
A.2.	Método biproporcional estándar . . . . .	66
A.3.	Método biproporcional con cotas superiores . . . . .	66
A.4.	Método triproporcional . . . . .	68
A.5.	Método de gradiente conjugado . . . . .	68



# Índice de figuras

1.1.	Red simple con dos centroides y cuatro líneas de tránsito. . . . .	5
1.2.	Estrategia óptima. Tiempo esperado de viaje 27.75 minutos. . . . .	7
1.3.	Volumen de pasajeros sobre cada segmento de tránsito. . . . .	8
1.4.	Resultados de una asignación de tránsito estándar. . . . .	9
2.1.	Ejemplo de una red pequeña con 9 nodos. . . . .	11
2.2.	Red de dos centroides con tres líneas de transporte. . . . .	20
5.1.	Red de transporte público de la ciudad de Winnipeg, Manitoba, [20]. . . . .	45
5.2.	Demanda esperada vs demanda inicial. . . . .	46
5.3.	Demanda esperada vs demanda balanceada, 5 iteraciones. . . . .	47
5.4.	Conteos en 112 arcos de la red. . . . .	48
5.5.	Dispersión inicial de los volúmenes de arco. . . . .	49
5.6.	Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los arcos y $\alpha = 0.5$ . . . . .	50
5.7.	Volúmenes observados vs volúmenes asignados, para máximo descenso y gradiente conjugado considerando conteos en los arcos y $\alpha = 0.5$ . . . . .	51
5.8.	Conteos en 136 segmentos de la red. . . . .	52
5.9.	Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y $\alpha = 0.5$ . . . . .	53
5.10.	Volúmenes asignados vs volúmenes observados, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y $\alpha = 0.5$ . . . . .	54
5.11.	Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y un factor de penalización de $k = 100$ . . . . .	56
5.12.	Volúmenes observados vs volúmenes asignados, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y un factor de penalización de $k = 100$ . . . . .	57
5.13.	Dispersión de demanda y de volúmenes para gradiente conjugado considerando conteos en los segmentos y un factor de penalización de $k = 1000$ con 54 iteraciones. . . . .	58

5.14. Dispersión de demanda y de volúmenes para gradiente conjugado considerando conteos en los segmentos y un factor de penalización de  $k = 10000$  con 53 iteraciones. . . . . 59

# Nomenclatura

$A$	Conjunto de aristas de la red.
$A_i^+$	Conjunto de aristas de salida desde el nodo $i \in N$ .
$A_i^-$	Conjunto de aristas de entrada desde el nodo $i \in N$ .
$\bar{A}$	Conjunto de aristas $a \in A$ que definen una estrategia.
$\bar{A}^*$	Conjunto de aristas $a \in A$ que definen una estrategia óptima.
$A^m$	Conjunto de aristas donde se tienen conteos del modo de transporte $m$ .
$a$	Elemento del conjunto de aristas $A$ .
$C$	Costo generalizado total que perciben todos los usuarios de la red.
$c_{pq}$	Costo generalizado al viajar del nodo $p$ al nodo $q$ .
$D_q$	Número de viajes que terminan en el nodo $q$ .
$E$	Entropía de un sistema.
$f_a$	Frecuencia efectiva de la arista $a \in A$ .
$\mathcal{G}$	Red generalizada de tránsito.
$G_a$	Función de distribución de tiempos de espera.
$G$	Matriz de demanda conocida a priori.
$G_{pq}$	$p$ -ésimo renglón y $q$ -ésima columna de la matriz $G$ .
$g$	Matriz de demanda que se desea calcular.
$g_i$	Demanda de pasajeros en el nodo $i \in N - q$ al nodo $q$ .
$g_{pq}$	$p$ -ésimo renglón y $q$ -ésima columna de la matriz $g$ .
$g_{pq}^m$	Demanda del nodo $p$ al nodo $q$ para el modo de transporte $m$ .
$h_s$	Flujo sobre la estrategia $s \in S$
$I$	Intervalo de costos.
$M$	Conjunto de todos los modos de transporte de la red.
$m$	Elemento del conjunto $M$ .
$N$	Conjunto de todos los nodos de la red.
$O_p$	Número de viajes que inician en el nodo $p$ .
$P$	Conjunto de todos los nodos origen.
$P(A_i^+)$	Probabilidad de que la arista $a$ sea servida primero.
$p$	Elemento del conjunto $P$ .
$\pi_s$	Probabilidad de la ruta $s$ .
$\pi_{pq}^a$	Probabilidad del estado $pq$ en el arco $a$ .
$Q$	Conjunto de todos los nodos destino.

$q$	Elemento del conjunto $Q$ .
$R_I$	Número de viajes asociado al intervalo $I$ .
$S$	Conjunto de todas las estrategias posibles.
$s$	Conjunto no vacío de líneas atractivas o estrategia.
$s_a(v_a)$	Función de costo en la arista $a$ .
$T_s(v)$	Tiempo de tránsito esperado para la estrategia $s$ .
$t_a$	Tiempo de viaje sobre la arista $a$ .
$\tau$	Tiempo mínimo.
$u_i$	Variable del problema dual que representa el tiempo total esperado de viaje del nodo $i$ al destino $q$ .
$V_0$	Conjunto de flujos factibles.
$V_a$	Volúmenes observados en el arco $a$ .
$V_i^d(v)$	Conjunto de flujos de equilibrio locales.
$V_{pq}^a$	Fracción de flujo observado del nodo $p$ al nodo $q$ sobre la arista $a$ .
$v_a$	Volumen de pasajeros en el arco $a$ .
$v_i$	Volumen esperado de pasajeros en el nodo $i \in N$ .
$W(A_i^+)$	Tiempo de espera combinado de las aristas $a \in A_i^+$ .
$w_i$	Tiempo total de espera para todos los viajeros en el nodo $i \in N$ .

# Introducción

En las sociedades actuales, los problemas de transporte son cada vez más importantes, sobre todo en las grandes ciudades, y es de suma importancia la planeación a mediano y largo plazo con el objeto de proporcionar un servicio más eficiente. Para ello es necesario estudiar y comprender el funcionamiento de la red de transporte utilizando las herramientas adecuadas. En particular, los modelos matemáticos de asignación de tránsito son una herramienta que permite entender cómo los usuarios del transporte público utilizan la red de transporte para viajar de sus distintos orígenes a sus diferentes destinos. Matemáticamente, una red de transporte se puede representar por medio de un conjunto de nodos y aristas. Los nodos pueden dividirse en nodos centroides, que son las zonas donde se origina o termina un viaje, y nodos simples que son los puntos donde los medios de transporte hacen paradas o intersecciones de dos o más aristas. Las aristas son los caminos de los que dispone el usuario durante su viaje, a cada arista se le asocia una función que modela el flujo sobre la misma. En las aristas también están definidos los segmentos de tránsito, los cuales representan a las diferentes líneas de transporte que pasan por dicha arista, por lo tanto el número de segmentos siempre será mayor o igual que el número de aristas.

En las grandes ciudades, suele ocurrir que algunos servicios de tránsito se saturan al grado de que los pasajeros no pueden abordar el primer vehículo que llega a su punto de espera, en estos casos es necesario modelar tanto la congestión de pasajeros dentro de los vehículos como los tiempos de espera crecientes para abordar un vehículo. Muchos modelos de elección de ruta no consideran el aumento en el tiempo de espera y generalmente solo imponen restricciones de capacidad, lo cual ocasiona que se sobrestime la oferta de servicio que pueden proporcionar algunas líneas. Por lo tanto, aparte de modelar los tiempos de espera crecientes, sería útil determinar cuándo la demanda no puede ser satisfecha por el servicio, independientemente de la elección de ruta. En el capítulo 1 se explican de manera muy breve algunos conceptos relativos a la red, la noción de estrategia y se introduce el modelo simple de asignación, así como el modelo más general que toma en cuenta la congestión y los límites de capacidad de las unidades de transporte.

Una vez realizada una asignación de tránsito, es importante saber de qué manera aprovechar la información obtenida para posibles estudios a futuro de tal manera que las nuevas asignaciones sean menos costosas que las anteriores y que permita hacer una buena planeación de transporte a futuro. Para tener una buena planeación de transporte es necesario

conocer datos de campo; esto para que los resultados obtenidos sean lo más realistas posible. Estos datos se pueden obtener a base de encuestas y otra clase de estudios que resultan ser muy complejos y costosos. Los datos de campo obtenidos para un proceso de asignación sólo serán útiles durante un corto periodo de tiempo, esto se debe a los cambios a los que están sujetas las ciudades y a su rápido crecimiento debido a la apertura de nuevos centros de entretenimiento o nuevas fuentes de trabajo. Si en el futuro se quisiera hacer un nuevo análisis de la red de transporte, entonces sería necesario obtener nuevamente los datos de entrada, lo cual requeriría una nueva inversión monetaria y mucho trabajo de campo. Para evitar hacer nuevos estudios completos sobre la demanda y los flujos sobre la red, existen algunas técnicas que permitan hacer aproximaciones a los datos más recientes haciendo uso de los datos conocidos anteriormente y utilizando solamente una cantidad pequeña pero significativa de datos nuevos.

Uno de los elementos más importantes en el proceso de planeación de transporte es la matriz de demanda, también denominada como matriz origen–destino, la cual representa el flujo entre cada centroide origen y cada centroide destino de la red de transporte. La matriz de demanda no es sólo uno de los elementos más importantes, sino también uno de los más difíciles de obtener, ya que dichos datos no pueden obtenerse a base de observaciones directas. Comprender el concepto de demanda de transporte no es fácil, ya que un viaje no representa un fin en sí mismo y las personas viajan para satisfacer sus necesidades o para llevar a cabo ciertas actividades, es por esto que la demanda no sólo depende del día en que se desea hacer el estudio, sino también del horario. Por otra parte, también es necesario considerar el factor del comportamiento humano, el cual introduce aleatoriedad e incertidumbre en el proceso de elección de rutas.

La elección de una representación adecuada de la demanda de transporte consiste en un intercambio entre la complejidad del modelo y de la precisión de los datos. Por un lado, el número total de viajes en la zona de interés, podría utilizarse como un indicador de la demanda de transporte, sin embargo el uso práctico de dicha información es limitado. Por otro lado, una descripción detallada de cada viaje, incluyendo el origen, el destino, todas las paradas intermedias, la hora exacta, el propósito del viaje, etc., proporcionaría información suficientemente completa, pero la viabilidad de recoger estos datos es bastante dudosa, especialmente para grandes áreas de estudio. Por otra parte, incluso si estuviese disponible toda esta información, sería difícil manejarla y probablemente no sería aceptable la amplitud de los errores de medición, por lo tanto, una representación razonable de la demanda debe estar entre estos dos extremos.

Una matriz de demanda se define como una tabla de dos dimensiones, cuyas filas y columnas representan cada zona del área de estudio. Una celda de la matriz se refiere por lo tanto a un par origen–destino en particular, y contiene el número total de personas que llevan a cabo este viaje, el cual debe estimarse a partir de los datos. Existen modelos de balanceo de matrices que consisten en actualizar las matrices ya existentes y considerando el número total de viajes que se originan en una zona  $O_p$  y el número total de viajes que tienen como destino

la zona  $D_q$ . Uno de los métodos simples es el método de balanceo biproportional, utilizado desde los años 40's por Deming y Stephan, [7], posteriormente conocido como método RAS usado por Stone en 1962 para estimar matrices de insumo-producto, [31], en 1970 conocido como el método de Furnes [16] o método de Fratar [21], el cual se basa en tasas de crecimiento y factores de balanceo. Este método puede verse con más detalle en la sección 2.2.1. Si al método de Fratar se le agrega una condición más referente a los costos generalizados de viaje se obtiene el método triproporcional, ver sección 2.2.3. Aunque estos modelos son muy interesantes desde el punto de vista teórico, en la práctica han tenido relativamente poca importancia, debido al gran tiempo de cálculo y a los requisitos de almacenamiento que surgen en implementaciones prácticas y que limitan estos enfoques a problemas de tamaños muy pequeños.

Existen otros modelos de demanda de transporte, los cuales se derivan de las leyes de la física. El más conocido es el gravitacional [4], el cual es una analogía al modelo de la ley de gravitación de Newton. Otro de los modelos más conocidos es el modelo de la entropía, derivado del segundo principio de la termodinámica [35] y [36]. La formulación de estos modelos puede verse en las secciones 2.1.1 y 2.1.2.

Contar el tráfico en una ciudad no es fácil; esto se puede hacer por ejemplo, observando las placas de los vehículos en los estacionamientos, de esta manera conocer su procedencia y tal vez, con permiso de los dueños, conocer no sólo el origen del vehículo, sino también la ruta que siguió para llegar a ese destino. Otra de las maneras para contar el tráfico es mediante sensores en las calles o cámaras que identifiquen el número de vehículos que utilizan ciertos arcos de la red. Por otra parte, esta tesis está enfocada a estimar matrices de demanda para cuando se tienen conteos de tránsito, sin embargo, estos datos son aún más difíciles de obtener. Una forma de contar el tránsito podría ser por medio de los torniquetes en las estaciones del metro o los sensores con los que cuentan algunos autobuses, pero esta información, sobre todo en los autobuses distaría mucho de ser realista, ya que en muchas ocasiones los usuarios no suben o bajan del autobús por la puerta destinada para dicha opción. Otra manera más eficiente, y tal vez más costosa de contar el tránsito, podría ser mediante observadores que aborden ciertos vehículos de transporte y que vayan registrando el volumen de pasajeros sobre cada segmento de tránsito.

Para hacer uso de la información obtenida mediante conteos, se han desarrollado un gran número de investigaciones que se enfocan en los diferentes métodos para estimar matrices O-D. Uno de estos métodos es el de mínimos cuadrados generalizados, el cual consiste en minimizar tanto el cuadrado de la diferencia entre la matriz conocida a priori y la matriz calculada, como en el cuadrado de la diferencia de los volúmenes observados y los volúmenes asignados, ver sección 3.3. Existe también el método de minimización de la información, el cual sugiere el uso de una matriz O-D que agrega la menor cantidad de información posible a la información obtenida en los conteos; este método puede verse en la sección 3.1 de manera más detallada. Una tercera técnica de los métodos de conteo, se basa en el modelo de maximización de entropía, pero considerando la restricción de los flujos observados, ver sección 3.2.

En el capítulo 4, se estudia el método de máximo descenso multiplicativo de Spiess [29], para resolver el problema de ajuste de demanda usando el modelo de mínimos cuadrados, el método se plantea de tal forma que la matriz resultante conserva la estructura de la matriz conocida a priori. Siguiendo la idea de Spiess, en el mismo capítulo se propone por primera vez el uso del método de gradiente conjugado para el ajuste de demanda, donde también se propone plantear el problema con un término de penalización, es decir, se considera que el término más importante a minimizar es la diferencia entre la demanda a priori y la demanda ajustada, entonces se propone agregar a la función objetivo el término correspondiente a la diferencia de volúmenes como una restricción y se penaliza. El sentido de la penalización es minimizar la diferencia de demandas hasta donde la tolerancia del error lo permita, y minimizar la diferencia de volúmenes sólo hasta cierto punto. Las contribuciones más importantes de este trabajo de tesis se encuentran en este capítulo, los cuales son: la introducción del método de gradiente conjugado para el ajuste de la demanda y un nuevo modelo de proyección utilizando mínimos cuadrados, y que está basado en la penalización de los volúmenes medidos.

Existen varios paquetes de software en el mercado para resolver problemas de asignación de tráfico y tránsito para implementar posibles escenarios futuros que ayuden a prevenir cualquier clase de contingencia, uno de estos paquetes es EMME4 [20], el cual fue desarrollado por la compañía INRO, la cual surge de la colaboración entre profesores y alumnos de la Universidad de Montreal, en particular, de Michael Florian, Heinz Spiess y posteriormente Yolanda Noriega, quienes han contribuido con diversos trabajos y artículos referentes a problemas de transporte urbano. Además de que EMME4 es una herramienta muy útil para hacer asignaciones de tráfico y tránsito, ya que toma en cuenta diversos factores que influyen en el proceso de asignación, también es posible hacer estimaciones de matrices O-D. Entre los métodos que se han implementado en dicho paquete se encuentran el de Fratar y el de mínimos cuadrados mencionados anteriormente. Para estimar una matriz O-D con el método de mínimos cuadrados, en EMME4 se utiliza el método de máximo descenso [29], el cual ha dado buenos resultados para algunas redes de transporte donde se ha usado (Winnipeg y Montreal entre otras).

En el capítulo 5 se muestran las comparaciones de los métodos de máximo descenso y de gradiente conjugado, considerando primeramente que se conocen los volúmenes en algunos arcos de la red y posteriormente que se conocen los volúmenes sobre algunos segmentos. También se muestran los resultados cuando se considera la función objetivo penalizada y se toman como factores de penalización 100, 1000 y 10000, mostrando la efectividad del algoritmo de gradiente conjugado aplicado al nuevo modelo con penalización. Finalmente, en el capítulo 6, se establecen las conclusiones de este trabajo y posibles aspectos para el trabajo a futuro.



# Capítulo 1

## Asignación de tránsito

En la ingeniería del transporte los modelos de asignación se dividen en dos tipos: los modelos de tráfico y los de tránsito. Estos modelos tienen por objeto estudiar redes de transporte urbanas para describir, predecir y explicar la forma en cómo los conductores de automóviles particulares, en el caso de tráfico, y los usuarios del transporte público, en el caso de tránsito, utilizan las diferentes rutas y líneas de transporte disponibles para dirigirse a su destino. Por lo tanto los modelos de asignación son una herramienta muy valiosa que puede ayudar a diseñar una mejor planeación estratégica y a la toma de decisiones en políticas de operación, con el objeto de mejorar la eficiencia del sistema de transporte y el ahorro de recursos, [15].

En un modelo de asignación de tránsito se busca modelar la forma en cómo la demanda del transporte (pasajeros) se distribuye en las diferentes rutas y líneas de transporte disponibles, de tal manera que el costo total en el sistema sea mínimo. El costo se refiere a cantidad de tiempo, confort, tarifas, o una combinación de éstas, es decir es un costo generalizado. Esta asignación se logra cuando no existe incentivo para que el usuario cambie de ruta, lo cual se conoce como equilibrio de la red.

Es muy común que se quiera formular el problema de asignación como un problema de ruta más corta, sin embargo se debe considerar que en el problema del viajero, el automovilista tiende a elegir una sola ruta de un conjunto de rutas posibles, mientras que en una red de tránsito los viajeros tienden a elegir un conjunto de rutas posibles y permite que la ruta a tomar quede determinada por el vehículo que arribe primero a cada nodo donde él se encuentra.

Una red de tránsito se representa por medio de una gráfica fuertemente conexa llamada red generalizada de tránsito denotada por  $\mathcal{G} = (N; A)$ , donde  $N$  es el conjunto de nodos de interconexión y  $A$  es el conjunto de arcos o aristas que representan los tramos y segmentos de las líneas de tránsito, así como los caminos peatonales.

Además de la red de transporte es necesario conocer la demanda en la misma, representada por la matriz origen–destino  $G = \{G_{pq}\}$  cuyo tamaño depende del número de nodos, en donde  $G_{pq}$  denota el número de pasajeros que inician su viaje en el nodo  $p$  y lo terminan

en el nodo  $q$ .

En el proceso de asignación se supone que en cada nodo se conoce el tiempo de inter-arribo (“headways”) de los vehículos de cada línea que sirve a ese nodo y que además se conoce la tasa de arribo de pasajeros. Con estos datos es posible conocer distribución del tiempo de espera de un vehículo de una línea dada, el tiempo promedio de espera combinado para el arribo del primer vehículo y la probabilidad de cada línea para arribar primero al nodo.

Considérese un usuario que va de un nodo origen  $p$  a un nodo destino  $q$ . Es posible que el usuario tenga una ruta definida o bien que tenga que tomar una decisión de como utilizar las diferentes líneas para llegar a su destino. En cualquier caso sus decisiones están basadas en estrategias que le reporten el menor costo. En 1984, Spiess y Florian introdujeron un modelo de asignación de tránsito basado en el concepto de estrategia y estrategia óptima, [27] y [30]. Una estrategia puede incluir subconjuntos de líneas si el viajero tuviese más información de la red, por ejemplo, la línea que será servida más pronto en cada nodo. Si hubiese más información, como la cantidad necesaria del tiempo de espera en un nodo ó información sobre el tiempo de viaje de otros vehículos, las estrategias podrían ser aún más complicadas. En este trabajo se supone que la única información disponible es el tiempo de espera de cada línea. Se dice que una estrategia es factible si la gráfica definida por sus líneas atractivas no contiene ciclos. Es posible utilizar los tiempos de viaje en cada línea, el tiempo de espera promedio en cada nodo y las probabilidades de línea, para calcular el tiempo total esperado de viaje de  $p$  a  $q$  en cada estrategia, cualquiera de ellas que minimice el tiempo total esperado de viaje se denomina estrategia óptima. Una vez seleccionado el conjunto de estrategias, se busca asignar los volúmenes de pasajeros a cada una de ellas, de tal forma que su tiempo de viaje sea mínimo.

## 1.1. El modelo básico sin congestión

El objetivo del modelo es minimizar el tiempo total esperado de viaje y se construye como una suma ponderada del tiempo de espera más el tiempo de viaje más el tiempo de caminata. En el modelo básico el conjunto de todas las estrategias factibles es el conjunto de elección de los viajeros, además, se considera que cada componente de viaje incluye un tiempo constante de viaje a bordo de un vehículo, así como una distribución de tiempos de espera. Antes de construir dicho modelo introducimos la siguiente notación::

- $A_i^+(A_i^-)$ : conjunto de aristas de salida (entrada) desde el nodo  $i \in N$ .
- $t_a \geq 0$ : tiempo de viaje sobre la arista  $a$ .
- $G_a$ : función de distribución del tiempo de espera, es decir:

$$G_a(x) = \text{probabilidad} \{ \text{tiempo de espera sobre la arista } a \leq x \}$$

- $W(A_i^+)$ : tiempo de espera promedio para el arribo del primer vehículo que sirve cualquiera de las aristas  $a \in A_i^+$ . A éste se le denomina el tiempo de espera combinado de las aristas  $a \in A_i^+$ .
- $P_a(A_i^+)$ : probabilidad de que la arista  $a \in A_i^+$  sea servida primero (de entre las aristas  $A_i^+$ ). Por conveniencia se define  $P_a(A_i^+) = 0$  si  $a \in A - A_i^+$ .
- $\bar{A}$ : Conjunto de aristas  $a \in A$  que definen una estrategia.
- $v_a$ : volumen esperado de pasajeros sobre la arista  $a \in A$ .
- $v_i$ : volumen esperado de pasajeros en el nodo  $i \in N$ .
- $g_i$ : demanda de pasajeros del nodo  $i \in N - \{q\}$  al nodo  $q$ .

Retomado la construcción del modelo básico, obsérvese que dada una estrategia  $\bar{A}$  y las demandas  $g_i$ , se realiza una asignación en la red dando lugar a los volúmenes de arista (arco)  $v_a$ . Por otro lado, el volumen de pasajeros  $v_i$  en un nodo  $i \in N$  es la suma de los volúmenes en todas las aristas de llegada y de la demanda en ese nodo. Es decir, se satisface la siguiente relación de balanceo de flujos:

$$v_i = \sum_{a \in A_i^-} v_a + g_i, \quad \forall i \in N \quad (1.1)$$

El volumen  $v_i$  de viajeros acumulados en el nodo  $i$  se distribuye sobre las aristas salientes de acuerdo a sus probabilidades de arista bajo la estrategia  $\bar{A}$ :

$$v_a = P_a(A_i^+) v_i, \quad a \in A_i^+, \quad i \in N \quad (1.2)$$

Una estrategia óptima  $\bar{A}^*$  minimiza la suma del tiempo total de viaje sobre las aristas más el tiempo total de espera sobre los nodos. Tomando en cuenta esto y la relación de balanceo de flujo (1.1), el modelo de optimización general toma la siguiente forma:

$$\text{mín} \sum_{a \in A} t_a v_a + \sum_{i \in N} W(\bar{A}_i^+) v_i \quad (1.3)$$

$$\text{sujeta a: } v_i \geq 0, \quad (1.4)$$

que se complementa con las restricciones (1.1) y (1.2). Como caso especial, se puede considerar una distribución de tiempos de espera para cada arista como un parámetro positivo  $f_a$  denominado la frecuencia de la arista. Con esta suposición, se derivan expresiones para el tiempo promedio de espera combinado y las probabilidades de arista:

$$W(\bar{A}_i^+) = \frac{\alpha}{\sum_{a \in \bar{A}_i^+} f_a}, \quad 0 < \alpha \leq 1 \quad (1.5)$$

$$P_a(\bar{A}_i^+) = \frac{f_a}{\sum_{b \in \bar{A}_i^+} f_b}, \quad a \in \bar{A}_i^+ \quad (1.6)$$

La expresión (1.5) para el tiempo de espera se basa en evidencia que se obtuvo de simulaciones, [6]. Para los diferentes valores de  $\alpha$  se consideran los siguientes casos, [6], [10], [26]:

- $\alpha = 1$ : Corresponde a una distribución exponencial de tiempos de inter-arribo de vehículos, con media  $1/f_a$ , además de una tasa uniforme de arribo de pasajeros en los nodos.
- $\alpha = 1/2$ : Corresponde a un servicio regular (tiempos de inter-arribo de vehículos constante) donde el usuario espera en promedio la mitad del tiempo de inter-arribo de los vehículos.
- $0.5 < \alpha < 1$ : Se estará modelando el tiempo de espera de un servicio irregular, cuando el vehículo tarda más tiempo de lo esperado.
- $\alpha < 0.5$ : Corresponde a un modelo donde los usuarios conocen los tiempos de inter-arribo de los vehículos y por lo tanto arriban al nodo al mismo tiempo que los vehículos.

Como las probabilidades de arista (1.6) son independientes de las unidades en que se especifica  $f_a$ , es posible escalar las frecuencia por el factor  $1/f_a$ . Por lo tanto, sin pérdida de generalidad, se puede suponer que  $\alpha = 1$  para construir el modelo.

Como puede observarse el problema (1.3) con las restricciones (1.1) y (1.2) es no lineal. Sin embargo, es posible transformar este problema sustituyendo en el segundo sumando el tiempo de espera de todos los viajeros en cada nodo  $i$ , [30]:

$$\omega_i = \frac{v_i}{\sum_{a \in A_i^+} \chi_a f_a} \quad (1.7)$$

obteniendo el siguiente problema equivalente de programación lineal:

$$\text{mín} \sum_{a \in A} t_a v_a + \sum_{i \in N} \omega_i, \quad (1.8)$$

$$\text{sujeta a:} \quad \sum_{a \in A_i^+} v_a - \sum_{a \in A_i^-} v_a = g_i, \quad i \in N \quad (1.9)$$

$$v_a \leq f_a \omega_i, \quad a \in A_i^+, \quad i \in N \quad (1.10)$$

$$v_a \geq 0, \quad a \in A \quad (1.11)$$

En [30] se demuestra formalmente que este problema es equivalente al problema no lineal previo.

En la práctica no se resuelve directamente el problema lineal (1.8)–(1.11), sino su formulación dual, debido a que es mucho más fácil de resolver y permite calcular la estrategia óptima en forma eficiente utilizando programación dinámica. Esto último hace posible su aplicación a redes de gran tamaño como la del Valle de México, [10]. En el problema dual se maximiza el tiempo total esperado de viaje  $u_i$  desde cada nodo  $i$  al nodo destino  $q$ :

$$\text{máx} \sum_{i \in N} g_i u_i \quad (1.12)$$

sujeto a ciertas restricciones de las variables duales asociadas. Una vez calculada la estrategia óptima se pueden asignar fácilmente los volúmenes sobre cada arco. Para mayores detalles se puede consultar [10] y [30].

El modelo lineal no toma en cuenta la congestión que se presenta en las horas de mayor demanda, como las horas pico, ni tampoco la capacidad limitada de los vehículos de transporte. A pesar de ello este modelo es muy útil cuando se utiliza como elemento básico en la construcción y solución de modelos más generales.

Para fijar ideas, considérese el siguiente ejemplo: Una red de transporte con cuatro nodos, donde dos de ellos son centroides. El flujo ocurre en una sola dirección, del nodo  $O$  al nodo  $D$ . La red de la figura 1.1 cuenta con cuatro líneas de transporte y se conocen los headways de cada una de ellas así como su tiempo de viaje. Se busca minimizar el tiempo de viaje del nodo  $O$  al nodo  $D$ .

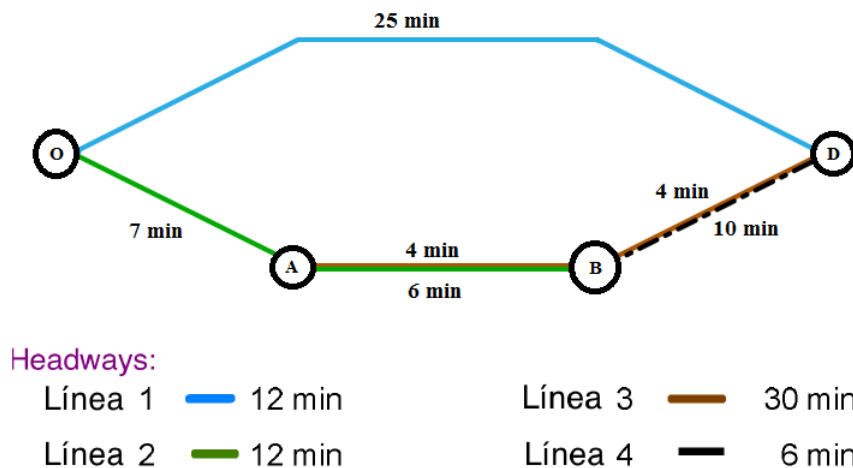


Figura 1.1: Red simple con dos centroides y cuatro líneas de tránsito.

El conjunto de estrategias que se tienen para esta pequeña red es el siguiente:

Nodo	Líneas atractivas (Línea → Nodo de salida)	Tiempo de espera (minutos)	Probabilidades de Línea			
			1	2	3	4
O	1→D	6.0	1.00	–	–	–
O	2 → A	6.0	–	1.00	–	–
O	2 → B	6.0	1.00	–	–	–
O	1 → D, 2 → A	3.0	0.50	0.50	–	–
O	1 → D, 2 → B	3.0	0.50	0.50	–	–
A	2 → B	6.0	–	1.00	–	–
A	3 → B	15.0	–	–	1.00	–
A	3 → D	15.0	–	–	1.00	–
A	2 → B, 3 → B	4.3	–	0.71	0.29	–
A	2 → B, 3 → D	4.3	–	0.71	0.29	–
B	3 → D	15.0	–	–	1.00	–
B	4 → D	3.0	–	–	–	1.00
B	3→D, 4→D	2.5	–	–	0.17	0.83

Tabla 1.1: Conjunto de 75 estrategias.

Una forma de resolver el problema es encontrando la ruta más corta, en este caso la solución sería tomar la línea 1 en  $O$  y bajarse en  $D$ , la cual tiene un costo de 31 minutos (25 minutos de viaje + 6 minutos de espera). Es posible encontrar un menor tiempo de viaje si se distribuyen los viajeros en las diferentes líneas de transporte. Por ejemplo, que en el nodo  $O$  la mitad de los pasajeros tomen la línea azul para descender en el nodo  $D$  y la otra mitad tome la línea 2 para descender en el nodo  $B$ ; asimismo, los pasajeros que llegan al nodo  $B$  se distribuyen de la siguiente forma para continuar su trayecto: el 8% toma la línea 3 y el otro 42% toma la línea 4. Entonces, en este caso el tiempo de tránsito puede calcularse de la siguiente manera:

- El tiempo de espera en los nodos se calcula utilizando la suma de las frecuencias de las líneas atractivas que pasan por ese nodo:

$$t_O = \frac{1/2}{1/12+1/12} = 3 \quad \text{Tiempo de espera en el nodo } O$$

$$t_B = \frac{1/2}{1/30+1/6} = \frac{5}{2} \quad \text{Tiempo de espera en el nodo } B$$

- El tiempo de viaje es simplemente el tiempo a bordo de cada vehículo, en este caso, los tiempos para cada línea están marcados en la figura 1.1.
- La probabilidad de abandonar el nodo  $i$  utilizando cierta línea se calcula dividiendo la frecuencia de esa línea entre la frecuencia combinada de las líneas atractivas en ese

nodo.

$$\pi_1^O = \frac{1/12}{1/12+1/12} = \frac{1}{2} \quad \text{Probabilidad de abandonar el nodo } O \text{ tomando la línea 1.}$$

$$\pi_2^O = \frac{1/12}{1/12+1/12} = \frac{1}{2} \quad \text{Probabilidad de abandonar el nodo } O \text{ tomando la línea 2.}$$

$$\pi_3^B = \frac{1/30}{1/30+1/6} = \frac{1}{6} \quad \text{Probabilidad de abandonar el nodo } B \text{ tomando la línea 3.}$$

$$\pi_4^B = \frac{1/6}{1/30+1/6} = \frac{5}{6} \quad \text{Probabilidad de abandonar el nodo } O \text{ tomando la línea 1.}$$

- Entonces, el tiempo de viaje será la suma del tiempo a bordo por la probabilidad de línea, por el volumen de pasajeros en cada arco:

$$t = 25 \frac{1}{2} \frac{100}{100} + (7 + 6) \frac{1}{2} \frac{100}{100} + 4 \frac{1}{6} \frac{50}{100} + 10 \frac{5}{6} \frac{50}{100} = 23.5$$

- El tiempo total se calcula sumando los tiempos de espera multiplicados por el volumen de pasajeros en cada nodo, más el tiempo de viaje, dando como resultado:

$$T = t_{espera} + t_{viaje} = \left( 3 \frac{100}{100} + 2.5 \frac{50}{100} \right) + 23.5 = 4.25 + 23.5 = 27.75$$

Por lo tanto, el tiempo de viaje es de 27.75 min, el cual es menor que el encontrado con la ruta más corta. Esta estrategia se ilustra en la figura 1.2.

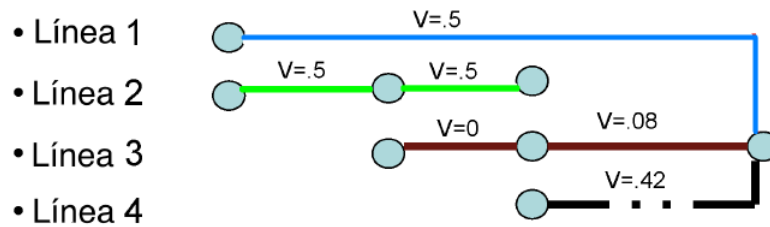


Figura 1.2: Estrategia óptima. Tiempo esperado de viaje 27.75 minutos.

El algoritmo para resolver este problema de asignación de tránsito fue implementado en MATLAB, obteniendo primero el conjunto de líneas atractivas y posteriormente asignando los volúmenes en cada arista. Los detalles de este algoritmo se pueden consultar en el apéndice A.1.

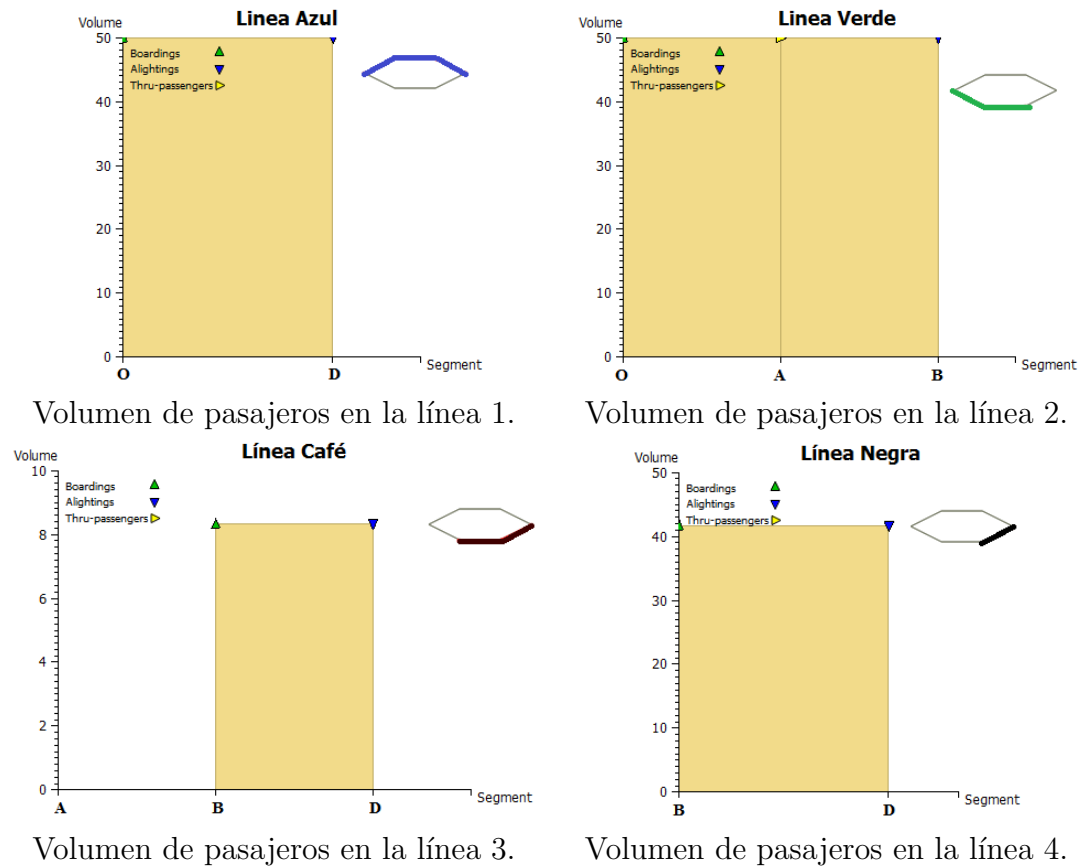


Figura 1.3: Volumen de pasajeros sobre cada segmento de tránsito.

En suma, se obtiene lo que ya se había ilustrado en la Figura 1.2, esto es, se asigna el 50 % de pasajeros a la línea 1 en el nodo  $O$  los cuales continúan su viaje en esa línea hasta llegar al nodo destino  $D$ . El otro 50 % de pasajeros fue asignado a la línea 2, la cual pasa por los nodos  $A$  y  $B$ , ningún pasajero desciende en  $A$  y todos llegan al nodo  $B$ . La línea 3 que va de  $A$  a  $D$  pasando por  $B$ , no lleva pasajeros en su primer segmento mientras que en el segundo, se asigna el 8 % de los pasajeros. A la línea 4, en su único segmento, se le asigna el 42 % de los pasajeros. Se puede ver información un poco más detallada de la asignación en el Logbook de EMME, como se muestra en la figura 1.4. La demanda total que se tenía en este ejemplo es de 100 pasajeros, la cual fue asignada en su totalidad. El número de abordajes es de 150, esto es 100 abordajes en el nodo  $O$  (50 línea 1 y 50 línea 2) más 50 abordajes en el nodo  $B$  (8 línea café y 42 línea negra). También se tiene que 50 pasajeros tomaron solamente una línea durante todo su viaje y que los otros 50 pasajeros tomaron dos líneas, esto hace un promedio de 1.5 líneas por pasajero. En la gráfica de resultados 1.4 se aprecia la demanda total a cada zona, en este caso, sólo se tiene un destino con una demanda de 100 pasajeros. También se muestra como tiempo promedio el tiempo de la estrategia óptima calculado anteriormente.



**Standard transit assignment**

Summary for all zones

total demand	assigned demand	not assigned demand	total boardings
100.0	100.0	0.0	150.0

avg lines per passenger	mean time	cpu time
1.5	27.75	0.01

Summary statistics per zone



Figura 1.4: Resultados de una asignación de tránsito estándar.

## 1.2. Asignación con congestionamiento

En esta parte se toma en cuenta cómo la congestión afecta los tiempos de espera promedio y los tiempos de viaje y, por lo tanto, los flujos (distribución de los volúmenes de pasajeros sobre los arcos). El modelo está basado en resultados de trabajos previos, [5] y [6]. En el último trabajo se extienden los resultados para obtener una nueva caracterización de los flujos de equilibrio, la cual permite formular un problema de optimización equivalente en términos de una función de holgura, y que se anula en el equilibrio. Esta nueva formulación del modelo permite trabajar con tiempos de viaje dependientes del flujo y es una generalización de modelos de equilibrio en redes de tránsito basados en estrategias. El enfoque permite obtener un algoritmo que se ha aplicado exitosamente en redes de tránsito de gran tamaño.

La congestión sobre los arcos se modela introduciendo funciones (de congestión) que se definen como la suma de un costo fijo  $t_a^0$  más una función de demora  $d_a(v_a)$ , es decir:

$$t_a(v_a) = t_a^0[1 + d_a(v_a)], \quad \text{con } d_a(0) = 0$$

Las funciones de demora  $d_a(x)$  son funciones no–negativas, continuas y crecientes que modelan la incomodidad en vehículos congestionados. Los tipos de funciones más utilizados son: funciones BPR (Bureau of Public Roads) y las funciones cónicas [28]. Debido a que los costos de viaje dependen de los volúmenes, los modelos resultantes ya no pueden ser lineales y solo

es posible apelar a ciertos principios para el planteamiento de los problemas como el principio de Wardrop [34], el cual afirma que, para todos los pares origen–destino las estrategias que llevan flujo son de costo generalizado mínimo y las que no llevan flujo son de costo mayor o igual al mínimo.

Por otro lado, una demanda excesiva puede provocar que ciertos pasajeros decidan no abordar el primer vehículo debido a la capacidad limitada de los mismos. Conforme los segmentos de tránsito se congestionan los niveles de comodidad disminuyen y los tiempos de espera aumentan. En este tipo de situaciones los tiempos de espera pueden modelarse con fórmulas de colas de estado estacionario que toman en cuenta la capacidad residual de los vehículos, así como el número de abordajes y de descensos. Una forma de modelar este fenómeno es multiplicar el headway original, sin congestión, por un factor para obtener un headway percibido o ajustado:

$$headway_{percibido} = headway_{original} \frac{1}{1 - \left( \frac{\text{subidas}}{\text{capacidad residual}} \right)^\beta}$$

donde  $\beta$  es un parámetro positivo menor que uno. Estos headways dan lugar a frecuencias de línea, denominadas frecuencias efectivas, que dependen del volumen en los arcos. En este caso la decisión óptima de un pasajero puede ser afectada por las decisiones de otros, por lo que es posible que haya más de una estrategia óptima. En trabajos sucesivos sobre el problema de líneas comunes [5] y [6], se extiende el modelo de equilibrio de tránsito para incluir tanto el congestionamiento dentro de los vehículos como los tiempos de espera crecientes. En esta versión del modelo asignación de tránsito, la caracterización de equilibrio en términos de las condiciones de Wardrop produce un problema mucho más complicado. En particular, se demuestra que un flujo de tránsito es de equilibrio si éste minimiza la siguiente función:

$$Gap(v) = \sum_{q \in Q} \left[ \sum_{a \in A} t_a(v) v_a + \sum_{i \in N} \omega_i - \sum_{i \in N} g_i u_i \right] \quad (1.13)$$

complementada con restricciones análogas a las del modelo lineal, y cuyo mínimo global es cero. Es decir, el tiempo total de tránsito menos el tiempo sobre las estrategias más cortas es igual a cero en el óptimo. Este es un problema considerablemente más difícil, debido a que no tiene una formulación equivalente en términos de un problema diferenciable de optimización convexa. Por esta razón, se utiliza el algoritmo de promedios sucesivos, el cual es un algoritmo iterativo de tipo heurístico que permite acercarse al óptimo mediante la solución de un problema lineal en cada iteración y el promediado de las soluciones sucesivas obtenidas. La función *Gap* permite monitorear el acercamiento al óptimo global y sirve como criterio de paro, para más detalles consultar [5] y [6].

## Capítulo 2

# Estimación de matrices de demanda

Considérese la red de la figura 2.1. La red está conformada por nueve nodos, de los cuales supondremos que los nodos del 1 al 5 son centriodos y los nodos del 6 al 9 son regulares.

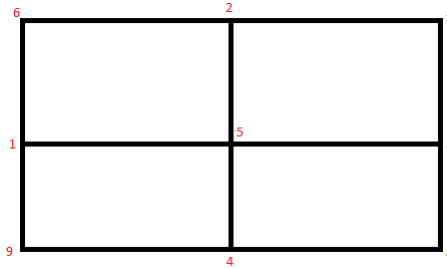


Figura 2.1: Ejemplo de una red pequeña con 9 nodos.

Una matriz de demanda asociada a dicha red puede ser la siguiente:

$$M = \begin{pmatrix} 0 & 2 & 4 & 7 & 3 \\ 10 & 0 & 3 & 5 & 1 \\ 2 & 2 & 0 & 6 & 8 \\ 5 & 3 & 8 & 0 & 5 \\ 6 & 9 & 5 & 3 & 0 \end{pmatrix}$$

en donde las entradas de la matriz  $M = \{m_{i,j}\}$  indican la cantidad de pasajeros que se transportan del nodo  $i$  al nodo  $j$ , por ejemplo, se tiene una demanda de 10 pasajeros que van del nodo 2 al nodo 1. Es importante conocer la matriz de demanda para una red de transporte ya que con ella será posible hacer asignaciones de tránsito y de esta manera conocer el flujo de personas sobre los arcos para poder mejorar el sistema de transporte o prevenir situaciones de contingencia.

En la actualidad, para obtener una matriz de demanda de transporte, es necesario hacer algunos estudios y encuestas para generar dicha información. En México, el organismo

que se encarga de realizar dichos estudios es el Instituto Nacional de Estadística y Geografía (INEGI), [19]. Estos datos son procesados y enviados al organismo de vialidad correspondiente, encargado de formular y conducir el desarrollo integral del transporte, así como planear y operar las vialidades en la zona de interés. En ocasiones, dado el inmenso trabajo que se debe hacer para obtener una matriz de demanda, los datos resultantes son liberados años más tarde, lo cual provoca que al momento que se quiere hacer un estudio del impacto de un posible escenario en una red de transporte, los resultados obtenidos no serán actuales.

En este capítulo se estudiarán los modelos y técnicas para estimar matrices de demanda haciendo uso de información conocida a priori.

## 2.1. Modelos para estimación de matrices

Existen maneras de hacer uso de la información obtenida años antes, combinándola con información más reciente y de esta manera tener una mejor aproximación de la matriz de demanda actual. Una de estas maneras es el balanceo de matrices, en el cual se considera el número total de pasajeros que inician y terminan su viaje en cada una de las zonas, el método consiste en encontrar parámetros, tales que al multiplicarlos por la matriz de demanda anterior, se genere una nueva matriz de demanda que coincida con los conteos en las zonas donde se origina y termina el tránsito. Algunos de estos modelos de ajuste de demanda de transporte se han derivado de las leyes de la física, de los cuales los mejor conocidos son el modelo gravitacional [4] y el modelo de máxima entropía [35] y [36]. En este capítulo se presentan sus derivaciones matemáticas y se muestra que éstas son equivalentes.

### 2.1.1. Modelo gravitacional

Considérese el conjunto de orígenes, denotado por  $P$ , y el conjunto de destinos, denotado por  $Q$ . La entrada de la matriz que corresponde a la  $p$ -ésima fila ( $p \in P$ ) y la  $q$ -ésima columna ( $q \in Q$ ) se denota por  $g_{pq}$ . Se supone que se conocen estimaciones anteriores de la matriz O-D a partir de trabajo de campo y que este conjunto de datos se utiliza para formar la matriz a priori  $G$ . El número total de viajes que salen del origen  $p$  se denota por  $O_p$ , mientras que el número total de los viajes que llegan al destino  $q$  se denota por  $D_q$ . El modelo gravitacional se construye usando una analogía a la ley de gravitación universal de Newton, en el cual se supone que el número de viajes  $g_{pq}$  entre el origen  $p$  y el destino  $q$  es proporcional al número  $O_p$  de personas que abandonan  $p$  y al número de personas  $D_q$  que llegan a  $q$ , y que es inversamente proporcional al cuadrado del costo generalizado  $c_{pq}$  al viajar de  $p$  a  $q$ , esto es

$$g_{pq} = \alpha \frac{O_p D_q}{c_{pq}^2}, \quad p \in P \quad q \in Q \quad (2.1)$$

Este modelo se puede generalizar introduciendo una función que depende del costo

$$g_{pq} = \alpha O_p D_q f(c_{pq}) \quad p \in P \quad q \in Q \quad (2.2)$$

En donde la función de disuasión  $f(c_{pq})$  puede tener las siguientes formas:

- Exponencial:  $f(c_{pq}) = e^{-\beta c_{pq}}$ .
- Polinomial:  $f(c_{pq}) = c_{pq}^{-\eta}$ .
- Combinación de funciones:  $f(c_{pq}) = c_{pq}^{\eta} e^{-\beta c_{pq}}$ .

En dichas funciones los parámetros  $\beta$  y  $\eta$  deberán ser calculados dependiendo del contexto.

### 2.1.2. Modelo de máxima entropía

Otro de los modelos teóricos que se encuentra en la literatura con mayor frecuencia es el llamado modelo de la entropía, en el cual se supone que para una determinada matriz O–D, cada micro estado correspondiente es equiprobable y esto ocurre cuando la entropía del sistema es máxima. Para el caso de balanceo de matrices O–D, los micro estados pueden interpretarse como los posibles pares origen destino que utilizan cierto arco de la red, y en este modelo, cada par origen destino es igual de probable que otro. La entropía de un sistema está dada por:

$$E(g) = \frac{\left( \sum_{pq \in PQ} g_{pq} \right)!}{\prod_{pq \in PQ} g_{pq}!} \quad (2.3)$$

en donde  $pq \in PQ$  significa que  $p \in P$  y  $q \in Q$ , y así se entenderá de ahora en adelante. Como no se conoce más información, como por ejemplo los costos de viaje, entonces la matriz O–D más probable se obtiene por:

$$\max_{g_{pq}} \frac{\left( \sum_{pq \in PQ} g_{pq} \right)!}{\prod_{pq \in PQ} g_{pq}!} \quad (2.4)$$

Además debe cumplir las condiciones de conservación y de no negatividad:

$$\sum_{q \in Q} g_{pq} = O_p \quad p \in P \quad (2.5)$$

$$\sum_{p \in P} g_{pq} = D_q \quad q \in Q \quad (2.6)$$

$$g_{pq} \geq 0. \quad pq \in PQ \quad (2.7)$$

Definiendo

$$E' = \log E = \log \left( \sum_{pq \in PQ} g_{pq} \right)! - \sum_{pq \in PQ} \log g_{pq}!$$

y usando la aproximación de Stirling [9], se obtiene:

$$E' = \log \left( \sum_{pq \in PQ} g_{pq} \right)! - \sum_{pq \in PQ} (g_{pq} \log g_{pq} - g_{pq})$$

Como el primer término permanece constante, es posible omitirlo en el proceso de maximización, reduciendo el problema a:

$$\max_{g_{pq}} E'' = - \sum_{pq \in PQ} (g_{pq} \log g_{pq} - g_{pq}) \quad (2.8)$$

sujeto a las condiciones (2.5)–(2.7). Como  $f(g_{pq}) = g_{pq} - g_{pq} \log g_{pq}$  es una función cóncava, y la suma de funciones cóncavas es cóncava, entonces la función objetivo (2.8) es cóncava. Además lo es de manera estricta, es por esto que el problema (2.8) sujeto a (2.5)–(2.7) tiene solución única.

Si se conoce información de la matriz a priori  $G_{pq}$ , esta información puede ser incluida de la siguiente manera:

$$\max_{g_{pq}} E'' = - \sum_{pq \in PQ} (g_{pq} \log(g_{pq}/G_{pq}) - g_{pq}) \quad (2.9)$$

en donde  $G_{pq}$  es el número de viajes de  $p$  a  $q$  de acuerdo a la información a priori.

Hasta este punto, no se ha considerado el impacto que tienen los costos de viaje en la demanda, esto puede hacerse agregando la siguiente restricción:

$$\sum_{pq \in PQ} g_{pq} c_{pq} = C \quad (2.10)$$

En esta última igualdad,  $C$  es el costo generalizado total que perciben todos los usuarios del sistema. Formulando el Lagrangiano del nuevo problema se tiene:

$$\begin{aligned} L(g) = & - \sum_{pq \in PQ} (g_{pq} \log(g_{pq}/G_{pq}) - g_{pq}) + \sum_{p \in P} \lambda_{1p} \left( O_p - \sum_{q \in Q} g_{pq} \right) \\ & + \sum_{q \in Q} \lambda_{2q} \left( D_q - \sum_{p \in P} g_{pq} \right) + \lambda_3 \left( C - \sum_{pq \in PQ} g_{pq} c_{pq} \right) \end{aligned}$$

Las condiciones de primer orden son:

$$\frac{\partial L(g)}{\partial g_{pq}} = - \log \frac{g_{pq}}{G_{pq}} - \lambda_{1p} - \lambda_{2q} - \lambda_3 c_{pq} = 0$$

Despejando  $g_{pq}$  se obtiene:

$$g_{pq} = G_{pq} e^{-\lambda_{1p} - \lambda_{2q} - \lambda_3 c_{pq}}$$

$$= G_{pq} e^{-\lambda_{1p}} e^{-\lambda_{2q}} e^{-\lambda_3 c_{pq}}$$

Denotando  $a_p = e^{-\lambda_{1p}}$  y  $b_q = e^{-\lambda_{2q}}$  y  $\alpha = a_p b_q$ , cuando se utiliza la información a priori contenida en  $G$ , se obtiene el modelo gravitacional con una función de disuasión exponencial:

$$g_{pq} = G_{pq} a_p b_q e^{-\lambda_3 c_{pq}}$$

Es posible obtener un modelo gravitacional con función de disuasión polinomial, en lugar de exponencial, si se reemplaza la condición (2.10) por:

$$\sum_{pq \in PQ} g_{pq} \log c_{pq} = C$$

En este caso, descartando la información a priori, el Lagrangiano y las condiciones de primer orden son las siguientes:

$$\begin{aligned} L(g) &= - \sum_{pq \in PQ} (g_{pq} \log(g_{pq}) - g_{pq}) + \sum_{p \in P} \lambda_{1p} \left( O_p - \sum_{q \in Q} g_{pq} \right) \\ &\quad + \sum_{q \in Q} \lambda_{2q} \left( D_q - \sum_{p \in P} g_{pq} \right) + \lambda_3 \left( C - \sum_{pq \in PQ} g_{pq} \log c_{pq} \right) \\ \frac{\partial L(g)}{\partial g_{pq}} &= - \log g_{pq} - \lambda_{1p} - \lambda_{2q} - \lambda_3 \log c_{pq} = 0 \end{aligned}$$

Despejando nuevamente  $g_{pq}$  se obtiene:

$$g_{pq} = e^{-\lambda_{1p}} e^{-\lambda_{2q}} c_{pq}^{-\lambda_3}$$

El cual es el modelo gravitacional pero con una función de disuasión polinomial:

$$g_{pq} = a_p b_q c_{pq}^{-\lambda_3}$$

Aunque el modelo gravitacional y el modelo de entropía se basan en diferentes supuestos, se ha demostrado que ambos modelos son equivalentes matemáticamente.

Por último, si en lugar de  $c_{pq}$  se toma  $c_{pq} - \log c_{pq}$  en (2.10) y formulando el Lagrangiano del nuevo problema se obtiene:

$$\begin{aligned} L(g) &= - \sum_{pq \in PQ} (g_{pq} \log(g_{pq}/G_{pq}) - g_{pq}) + \sum_{p \in P} \lambda_{1p} \left( O_p - \sum_{q \in Q} g_{pq} \right) \\ &\quad + \sum_{q \in Q} \lambda_{2q} \left( D_q - \sum_{p \in P} g_{pq} \right) + \lambda_3 \left[ C - \sum_{pq \in PQ} g_{pq} (c_{pq} - \log c_{pq}) \right] \end{aligned}$$

Las condiciones de primer orden son:

$$\frac{\partial L(g)}{\partial g_{pq}} = -\log \frac{g_{pq}}{G_{pq}} - \lambda_{1p} - \lambda_{2q} - \lambda_3 (c_{pq} - \log c_{pq}) = 0$$

Despejando  $g_{pq}$  se obtiene:

$$\begin{aligned} g_{pq} &= G_{pq} e^{-\lambda_{1p} - \lambda_{2q} - \lambda_3 c_{pq}} c_{pq}^{\lambda_3} \\ &= G_{pq} e^{-\lambda_{1p}} e^{-\lambda_{2q}} e^{-\lambda_3 c_{pq}} c_{pq}^{\lambda_3} \end{aligned}$$

Denotando  $a_p = e^{-\lambda_{1p}}$  y  $b_q = e^{-\lambda_{2q}}$ , se obtiene el modelo gravitacional con una función de disuasión combinada:

$$g_{pq} = G_{pq} a_p b_q e^{-\lambda_3 c_{pq}} c_{pq}^{\lambda_3}$$

## 2.2. Métodos de balanceo de matrices

Considérese el problema de determinar una matriz origen destino  $g$  con las siguientes restricciones para la matriz  $g$ :

$$\sum_{q \in Q} g_{pq} = O_p, \quad \forall p \in P \quad (2.11)$$

$$\sum_{p \in P} g_{pq} = D_q, \quad \forall q \in Q \quad (2.12)$$

Existen métodos muy simples para actualizar la matriz O–D, los cuales hacen uso del número real de viajes que se originan en cada zona  $O_p$  y el número efectivo de viajes que terminan en cada zona  $D_q$ . Estos métodos, además de considerar las restricciones de conservación (2.11) y (2.12), pueden incluir también otras condiciones como cotas superiores de las entradas de la matriz O–D.

### 2.2.1. Método biproporcional estándar

El método estándar de balanceo biproporcional de matrices, también conocido como método RAS [31], método de Fratar [21] o método de Furness [16] consiste en encontrar  $a_p$  y  $b_q$  para cada origen  $p \in P$  y destino  $q \in Q$  tales que, al multiplicarse por los correspondientes renglones y columnas de la matriz conocida a priori  $G_{pq}$ , los totales marginales de la matriz resultante  $g_{pq}$  corresponden a los orígenes  $O_p$  y los destinos  $D_q$ . Esto puede expresarse en las siguientes ecuaciones:

$$g_{pq} = a_p b_q G_{pq} \quad p \in P, \quad q \in Q, \quad (2.13)$$

$$\sum_{q \in Q} g_{pq} = O_p, \quad p \in P \quad (2.14)$$

$$\sum_{p \in P} g_{pq} = D_q, \quad q \in Q \quad (2.15)$$

$$g_{pq} \geq 0 \quad p \in P, \quad q \in Q, \quad (2.16)$$



Siguiendo a Florian [13], el problema anterior también puede expresarse como un problema de optimización convexa y también puede encontrarse en la literatura como problema de transporte de máxima entropía. El problema puede escribirse como:

$$\min \sum_{pq \in PQ} g_{pq} (\ln g_{pq} - \ln G_{pq} - 1) \quad (2.17)$$

sujeto a las condiciones (2.14)–(2.16), el cual corresponde a la ecuación (2.9).

Definiendo  $\alpha_p$  y  $\beta_q$  como las variables duales asociadas a las restricciones (2.14) y (2.15) y de las condiciones de optimalidad de Kuhn–Tucker se obtiene la siguiente expresión:

$$\begin{aligned} \cancel{g_{pq}} / \cancel{g_{pq}} + \ln g_{pq} - \ln G_{pq} - \cancel{\lambda} + \alpha_p + \beta_q &= 0 \\ \Rightarrow \ln g_{pq} &= \ln G_{pq} - \alpha_p - \beta_q \\ g_{pq} &= e^{-\alpha_p - \beta_q} G_{pq}, \quad pq \in PQ \end{aligned} \quad (2.18)$$

recordando que  $pq \in PQ$  significa  $p \in P$  y  $q \in Q$ . Considerando  $a_p = e^{-\alpha_p}$  y  $b_q = e^{-\beta_q}$ ,  $pq \in PQ$ , el problema anterior corresponde al problema planteado en la ecuación (2.13). Además, la formulación dual del problema (2.17) se puede escribir como el siguiente problema de optimización sin restricciones:

$$\min_{\alpha, \beta} \sum_{pq \in PQ} G_{pq} e^{-\alpha_p - \beta_q} + \sum_{p \in P} \alpha_p O_p + \sum_{q \in Q} \beta_q D_q \quad (2.19)$$

Si el coeficiente de proporción  $\alpha$  de la ecuación (2.2) se separase en dos factores de balanceo  $\alpha = a_p b_q$ , y considerando las restricciones (2.11) y (2.12), el problema se puede resolver con el método de Furnes.

El modelo estándar de balanceo biproporcional de matrices se caracteriza por ser factible cuando  $\sum_{p \in P} O_p = \sum_{q \in Q} D_q$  y  $G_{pq} > 0$  para todo  $pq \in PQ$ . Es fácil ver que para este problema se pueden combinar las ecuaciones (2.13) y (2.14) para despejar  $a_p$  y balancear los orígenes, posteriormente combinar las ecuaciones (2.13) y (2.15) para despejar  $b_q$  y balancear los destinos. El algoritmo de solución se incluye en el apéndice A.2, también se pueden consultar detalladamente los aspectos de unicidad de la solución y convergencia del método en [1].

### 2.2.2. Método biproporcional con cotas superiores

En esta sección se introduce una extensión del modelo definido en la sección anterior, en donde se imponen cotas superiores  $U_{pq}$  a los elementos de la matriz resultante  $g_{pq}$ . La

formulación del problema extendido es la siguiente:

$$\min \sum_{pq \in PQ} g_{pq} (\ln g_{pq} - \ln G_{pq} - 1) \quad (2.20)$$

$$\text{sujeto a: } \sum_{q \in Q} g_{pq} = O_p, \quad p \in P \quad (2.21)$$

$$\sum_{p \in P} g_{pq} = D_q, \quad q \in Q \quad (2.22)$$

$$g_{pq} \leq U_{pq}, \quad pq \in PQ \quad (2.23)$$

Utilizando nuevamente  $\alpha_p$  y  $\beta_q$  como las variables duales asociadas a las restricciones (2.21) y (2.22) y agregando la nueva variable dual  $\mu_{pq}$  asociada a la condición de la cota superior (2.23), se pueden escribir las condiciones de Kuhn–Tucker del problema anterior como:

$$\ln g_{pq} - \ln G_{pq} + \alpha_p + \beta_q + \mu_{pq} = 0, \quad pq \in PQ \quad (2.24)$$

$$\mu_{pq}(U_{pq} - g_{pq}) = 0, \quad pq \in PQ. \quad (2.25)$$

La condición de optimalidad (2.24) puede reescribirse como:

$$g_{pq} = G_{pq} e^{-\alpha_p - \beta_q - \mu_{pq}}, \quad pq \in PQ \quad (2.26)$$

Aplicando la condición de holgura complementaria (2.25) para la variable dual  $\mu_{pq}$  y distinguiendo los casos cuando  $\mu_{pq} = 0$  y  $\mu_{pq} > 0$ , la ecuación (2.26) se convierte en:

$$g_{pq} = \begin{cases} G_{pq} e^{-\alpha_p - \beta_q} = a_p b_q G_{pq} & \text{si } \mu_{pq} = 0 \\ U_{pq} & \text{Otro caso} \end{cases} \quad (2.27)$$

la cual puede escribirse de manera más concisa como:

$$g_{pq} = \min\{G_{pq} e^{-\alpha_p - \beta_q}, U_{pq}\} = \min\{a_p b_q G_{pq}, U_{pq}\}, \quad pq \in PQ \quad (2.28)$$

Nótese que la formulación anterior no contiene explícitamente las variables duales  $\mu_{pq}$ .

Formulando el Lagrangiano del problema dual se obtiene:

$$\min_{\alpha, \beta, \mu} -L(\alpha, \beta, \mu) = \sum_{pq \in PQ} G_{pq} e^{-\alpha_p - \beta_q - \mu_{pq}} + \sum_{p \in P} \alpha_p O_p + \sum_{q \in Q} \beta_q D_q + \sum_{pq \in PQ} \mu_{pq} U_{pq} \quad (2.29)$$

$$\mu \geq 0$$

y es fácil observar que a excepción de la no–negatividad de  $\mu_{pq}$ , el problema dual es esencialmente un problema sin restricciones. Para este tipo de problemas, se sabe que el método de descenso en direcciones canónicas converge a la solución óptima [22]. Por lo tanto se puede utilizar este método para resolver el problema dual (2.29) y encontrar los valores óptimos de las variables duales. Si el problema primal (2.20) es factible, los valores óptimos de las

variables duales  $\alpha_p$  y  $\beta_q$  se sustituyen en la ecuación (2.28) para obtener la solución óptima del problema primal.

Aplicar el método de descenso en las direcciones canónicas al problema (2.29) significa resolver cíclicamente las condiciones de optimalidad de primer orden respecto a las variables duales correspondientes. Esto implica encontrar los ceros de las primeras derivadas parciales de la función objetivo dual respecto a las variables duales.

$$\frac{\partial L(\alpha, \beta, \mu)}{\partial \alpha_p} = - \sum_{q \in Q} G_{pq} e^{-\alpha_p - \beta_q - \mu_{pq}} + O_p = 0, \quad p \in P, \quad (2.30)$$

$$\frac{\partial L(\alpha, \beta, \mu)}{\partial \beta_q} = - \sum_{p \in P} G_{pq} e^{-\alpha_p - \beta_q - \mu_{pq}} + D_q = 0, \quad q \in Q, \quad (2.31)$$

$$\frac{\partial L(\alpha, \beta, \mu)}{\partial \mu_{pq}} = -G_{pq} e^{-\alpha_p - \beta_q - \mu_{pq}} + U_{pq} = 0, \quad \forall \mu_{pq} > 0. \quad (2.32)$$

Observando que (2.32) es equivalente a (2.28), se pueden combinar (2.30) y (2.32) en una sola condición más simple:

$$\sum_{q \in Q} \min\{a_p b_q G_{pq}, U_{pq}\} = O_p, \quad p \in P \quad (2.33)$$

y las condiciones (2.31) y (2.32) se pueden combinar como:

$$\sum_{p \in P} \min\{a_p b_q G_{pq}, U_{pq}\} = D_q, \quad q \in Q \quad (2.34)$$

La introducción de las cotas superiores influye en la factibilidad del problema de la misma manera en que influyen los ceros en la matriz  $G_{pq}$ . Las condiciones necesarias de factibilidad son:

$$\sum_{q \in Q} U_{pq} \geq O_p \quad p \in P \quad (2.35)$$

$$\sum_{p \in P} U_{pq} \geq D_q \quad q \in Q \quad (2.36)$$

El algoritmo de solución se basa en resolver iterativamente las ecuaciones (2.33) y (2.34).

Para resolver el problema de balanceo biproporcional de matrices con cotas superiores (2.20) se propone utilizar un algoritmo que consta de dos partes. Por un lado se realiza el balanceo respecto a los coeficientes  $a_p, b_q$ , asumiendo que es posible encontrar  $x$  que resuelve el problema del tipo  $\sum_i \min\{x f_i, u_i\} = T$ , ver algoritmo 1 del apéndice A.3. En el algoritmo 2 del apéndice A.3 se considera el subproblema de encontrar el factor  $x$ , ordenando los elementos  $i$  de acuerdo a los cocientes  $u_i/f_i$  y revisando ordenadamente en los elementos si el valor óptimo  $x$  se encuentra entre dos coeficientes consecutivos  $u_i/f_i$ . En esta sección se

agregó una cota superior al método biproporcional, pero esto puede aplicarse también a otro tipo de métodos, aunque esto ya no se muestra en este trabajo.

La naturaleza multiplicativa de los métodos vistos anteriormente, implica que las entradas de la matriz O–D que son inicialmente cero permanecerán así durante el balanceo, aunque si se desea, esto se puede cambiar asignando valores pequeños llamados semillas a dichas entradas para permitirles que aumente su valor. Para ejemplificar estos métodos considérese una red de transporte en dos cuadras de cualquier ciudad representada en la figura 2.2.

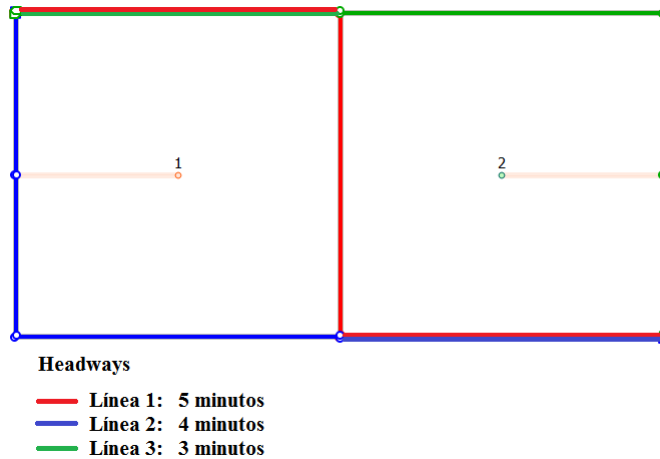


Figura 2.2: Red de dos centroides con tres líneas de transporte.

Supóngase que mediante una serie de encuestas se obtuvo una matriz de demanda  $G$  y el total de viajes originados y terminados en cada nodo.

$$G = \begin{pmatrix} 0 & 280 \\ 179 & 0 \end{pmatrix}; \quad O = \begin{pmatrix} 300 \\ 150 \end{pmatrix}; \quad D = \begin{pmatrix} 150 & 300 \end{pmatrix}$$

Como puede observarse, la suma de renglones de  $G$  no corresponde a los orígenes ni la suma de las columnas de  $G$  corresponde a los destinos, esto quiere decir que la matriz  $G$  no está balanceada. Aplicando el algoritmo para balanceo biproporcional de matrices con cotas superiores, apéndice A.3, considerando como cota la suma total de orígenes y destinos, en dos iteraciones se obtiene:

$$a = (1.0714 \quad 0.8380); \quad b = (1 \quad 1); \quad g = \begin{pmatrix} 0 & 300 \\ 150 & 0 \end{pmatrix}$$

Como se ve, la suma de los renglones y de las columnas de la matriz  $g$  si corresponden a los orígenes y destinos, respectivamente, conocidos en los nodos. Esto quiere decir que la matriz encontrada  $g$  representa de manera más precisa la demanda entre los dos nodos de la red, esto quiere decir que hay 300 personas en el nodo 1 que desean ir al nodo 2, y que 150 personas del nodo 2 desean ir al nodo 1. En el capítulo 6, se muestra este mismo método aplicado a la red de transporte de la ciudad de Winnipeg.

### 2.2.3. Método triproporcional

Otro de los métodos simples es el método tri-proporcional, el cual incluye información sobre los costos de viaje además de las restricciones (2.14) y (2.15). Los datos de distribución de costos proporcionan el número total de viajes correspondientes a cada intervalo de costo agregado. Esto puede hacerse agregando la siguiente restricción:

$$\sum_{p \in P, q \in Q} g_{pq} \delta_{pq}^I = R_I \quad \forall I, \quad (2.37)$$

en donde  $R_I$  es el número de viajes asociado al intervalo  $I$  y  $\delta_{pq}^I$  es 1 si el costo de  $p$  a  $q$  está en el intervalo  $I$ , y es cero en otro caso. El procedimiento de este método es similar al del método de Fratar, consiste en ajustar cada una de las restricciones de manera sucesiva e iterar hasta alcanzar la convergencia o alcanzar un número máximo de iteraciones. Se considera que este método es un poco mejor que el de Fratar, pero sólo cuando se conoce la distribución de costos de viaje, además se supone que esta información es independiente de las variaciones de demanda; es por esto que este método resulta útil pero sólo cuando se buscan aproximaciones a corto plazo. El algoritmo correspondiente se incluye en el apéndice A.4.



# Capítulo 3

## Métodos de conteo

Los modelos y métodos introducidos en el capítulo anterior, aunque son muy sencillos y fáciles de realizar computacionalmente, tienen el inconveniente de que necesitan información que es difícil de obtener en la práctica. Para poder aplicarlos, es necesario conocer el número de viajes  $O_p$  que salen de cada origen  $p$  y el número de viajes  $D_q$  que llegan a cada destino  $q$  de la red. En una red de gran tamaño es prácticamente imposible conocer, o al menos estimar, esta información de manera confiable, a menos que se tengan políticas de actualización de la información en forma periódica, lo cual no sucede en México. Por esta razón, en este trabajo se opta por estudiar modelos y métodos en los que la información que se requiere se pueda obtener en forma realista y económica. A continuación se introducen una clase de modelos basados en la observación de flujos sobre un conjunto de arcos o segmentos de la red de transporte.

Las técnicas de estimación de matrices basadas en la observación de flujos han sido ampliamente desarrolladas, principalmente por el bajo costo que requiere la obtención de dichos datos. Todos ellos se basan en las ecuaciones de asignación

$$v_a = \sum_{pq \in PQ} V_{pq}^a g_{pq} \quad (3.1)$$

en donde  $g = \{g_{pq}\}$ , y  $g_{pq}$  es el elemento del renglón  $p$  y la columna  $q$  de la matriz de demanda,  $V_{pq}^a$  es la proporción del volumen de pasajeros que van del nodo  $p$  al nodo  $q$  y que utilizan el arco  $a$ ,  $v_a$  es el volumen total sobre el arco  $a$ .

Es importante aclarar que la estimación de la matriz de demanda utilizando conteos dependerá del método utilizado para el proceso de asignación. Generalmente se consideran dos tipos de modelos, en uno se supone que la fracción de volumen  $V_{pq}^a$  no depende de la congestión en los arcos o segmentos de tráfico y se consideran asignaciones del tipo “todo o nada”, en el otro se consideran los efectos de la congestión y las restricciones de capacidad basándose en los principios de conservación de Wardrop [34] y la transformación de Beckman [2]. A continuación se presentarán tres modelos basados en la observación de flujos en un conjunto predeterminado de arcos o segmentos de la red de transporte. En los primeros dos

modelos se supone que  $V_{pq}^a$  no depende de la congestión y que tampoco se puede obtener directamente de la matriz O–D. Estos son el modelo de minimización de la información (sección 3.1) y el modelo de entropía (sección 3.2). El último modelo se introduce en la sección 3.3, éste es un modelo de proyección basado en mínimos cuadrados generalizados, el cual se estudia y desarrolla de manera exhaustiva en el Capítulo 4.

### 3.1. Minimización de información

En este método se sugiere el uso de una matriz que agregue la menor cantidad de información o incertidumbre posible a la información obtenida en los conteos. De acuerdo a la distribución multinomial, la información contenida en un conjunto de  $N$  observaciones, donde el estado  $k$  es observado  $n_k$  número de veces, se define como:

$$I = -\log \left( N! \prod_k \frac{\pi_k^{n_k}}{n_k!} \right) \quad (3.2)$$

en donde  $\pi_k$  es la probabilidad de observar el estado  $k$ . En este caso, el estado sería el par O–D  $pq$ , que se obtiene de observar a un usuario de la red que va de  $p$  a  $q$ . El número de observaciones  $n_{pq}^a$  del estado  $pq$  en el arco  $a$  es el siguiente:

$$n_{pq}^a = g_{pq} V_{pq}^a, \quad pq \in PQ \quad (3.3)$$

en donde  $V_{pq}^a$  es la fracción de volumen observado de  $p$  a  $q$  en el arco  $a$ , esta cantidad está entre 0 y 1. La probabilidad del estado  $pq$   $\pi_{pq}^a$  está dada por:

$$\pi_{pq}^a = \frac{G_{pq} V_{pq}^a}{\sum_{pq \in PQ} G_{pq} V_{pq}^a} \quad (3.4)$$

Sustituyendo (3.3) y (3.4) en (3.2), se obtiene la información  $I_a$  contenida en  $v_a$  observaciones sobre el arco  $a$ , (ver [33]):

$$I_a(g) = -\log v_a! \prod_{pq \in PQ} \frac{\left( \frac{G_{pq} V_{pq}^a}{S^a} \right)^{g_{pq} V_{pq}^a}}{(g_{pq} V_{pq}^a)!} \quad (3.5)$$

en donde  $S^a = \sum_{pq \in PQ} G_{pq} V_{pq}^a$ . Por las propiedades de los logaritmos la ecuación (3.5) se puede expresar como:

$$I_a(g) = \sum_{pq \in PQ} \log(g_{pq} V_{pq}^a)! - \log v_a! - \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{G_{pq} V_{pq}^a}{S^a} \right)$$



Usando la aproximación de Stirling y haciendo uso de (3.1) se obtiene:

$$I_a(g) = \sum_{pq \in PQ} [g_{pq} V_{pq}^a \log(g_{pq} V_{pq}^a) - \cancel{g_{pq} V_{pq}^a}] - v_a \log v_a + \cancel{v_a} + \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{S^a}{G_{pq} V_{pq}^a} \right)$$

$$I_a(g) = \sum_{pq \in PQ} g_{pq} V_{pq}^a \log(g_{pq} V_{pq}^a) - \sum_{pq \in PQ} g_{pq} V_{pq}^a \log v_a + \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{S^a}{G_{pq} V_{pq}^a} \right)$$

Factorizando se tiene:

$$I_a(g) = \sum_{pq \in PQ} g_{pq} V_{pq}^a \left[ \log(g_{pq} V_{pq}^a) - \log v_a + \log \left( \frac{S^a}{G_{pq} V_{pq}^a} \right) \right]$$

Utilizando nuevamente las propiedades de los logaritmos queda:

$$I_a(g) = \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right)$$

La información total está dada por:

$$I(g) = \sum_{a \in A} I_a(g_{pq}) = \sum_{a \in A} \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right) \quad (3.6)$$

El problema se reduce a minimizar esta información bajo las restricciones (3.1). El Lagrangiano asociado a este problema es el siguiente:

$$L(g) = \sum_{a \in A} \left[ \sum_{pq \in PQ} g_{pq} V_{pq}^a \log \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right) + \lambda_a \left( \sum_{pq \in PQ} g_{pq} V_{pq}^a - v_a \right) \right]$$

en donde  $\lambda_a$  es el multiplicador de Lagrange correspondiente al arco  $a$ . Las condiciones de primer orden son las siguientes:

$$\begin{aligned} \frac{\partial L(g)}{\partial g_{pq}} &= \sum_{a \in A} \left[ V_{pq}^a + V_{pq}^a \log \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right) + \lambda_a V_{pq}^a \right] = 0 \\ \Rightarrow \log \prod_{a \in A} \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right)^{V_{pq}^a} &= - \sum_{a \in A} (1 + \lambda_a) V_{pq}^a \\ \Rightarrow \prod_{a \in A} \left( \frac{S^a g_{pq}}{v_a G_{pq}} \right)^{V_{pq}^a} &= e^{-\sum_{a \in A} (1 + \lambda_a) V_{pq}^a} = \prod_{a \in A} e^{-(1 + \lambda_a) V_{pq}^a} \\ \Rightarrow \left( \frac{g_{pq}}{G_{pq}} \right)^{\sum_{a \in A} V_{pq}^a} &= \left[ \prod_{a \in A} \frac{v_a}{S^a} e^{-(1 + \lambda_a)} \right]^{V_{pq}^a} \end{aligned}$$

Despejando  $g_{pq}$  finalmente se obtiene:

$$g_{pq} = G_{pq} \left[ \prod_{a \in A} \frac{v_a}{S^a} e^{-(1 + \lambda_a)} \right]^{V_{pq}^a / \sum_{a \in A} V_{pq}^a}$$

### 3.2. Modelo de entropía considerando el flujo

Este modelo, como su nombre lo dice, se basa en el modelo de entropía visto en la sección 2.1.2. La función (2.8) se maximiza considerando la restricción de flujo (3.1). El Lagrangiano asociado a este problema es:

$$L(g) = - \sum_{pq \in PQ} (g_{pq} \log g_{pq} - g_{pq}) - \sum_{a \in A} \lambda_a \left( \sum_{pq \in PQ} g_{pq} V_{pq}^a - v_a \right)$$

en donde  $\lambda_a$  es el correspondiente multiplicador de Lagrange para el arco  $a$ . Las condiciones de primer orden son las siguientes:

$$\frac{\partial L(g_{pq})}{\partial g_{pq}} = -\log g_{pq} - \sum_{a \in A} \lambda_a V_{pq}^a = 0$$

Despejando  $g_{pq}$  se obtiene:

$$g_{pq} = \prod_{a \in A} e^{-\lambda_a V_{pq}^a} \quad (3.7)$$

Es posible extender el modelo anterior haciendo uso de la información conocida a priori, esto puede hacerse reemplazando la ecuación (2.3) por:

$$E(g, G) = \left( \sum_{pq \in PQ} g_{pq} \right)! \prod_{pq \in PQ} \left[ \frac{1}{g_{pq}} \left( \frac{G_{pq}}{\sum_{pq \in PQ} G_{pq}} \right)^{g_{pq}} \right] \quad (3.8)$$

Aplicando logaritmos, utilizando la aproximación de Stirling y resolviendo el problema sujeto a la restricción (3.1) se obtiene:

$$g_{pq} = G_{pq} \prod_{a \in A} \left[ \left( \sum_{pq \in PQ} G_{pq} \right)^{\frac{1}{m}} e^{-\lambda_a} \right]^{V_{pq}^a} \quad (3.9)$$

en donde  $m$  es el número de arcos considerados en el conteo.

### 3.3. Mínimos cuadrados generalizados

El método de mínimos cuadrados generalizados consiste en la minimización de la distancia entre la matriz conocida a priori  $G$  y la matriz calculada  $g$ , y entre los volúmenes observados  $V_a$  y los asignados  $v_a$ . La función objetivo que se desea minimizar es:

$$\min_g Z(g) = \sum_{pq \in PQ} w_{pq}^G (g_{pq} - G_{pq})^2 + \gamma \sum_{a \in A} w_a^V (v_a - V_a)^2 \quad (3.10)$$

en donde  $A$  es el conjunto de arcos donde se realizan los conteos,  $v_a$  es el volumen asignado con la matriz  $g$ ,  $V_a$  es el volumen observado en el arco  $a$ ,  $w_{pq}^G$  es la confianza relativa que se tiene en los datos  $G_{pq}$ ,  $w_a^V$  es la confianza que se tiene en el valor  $V_a$ , y  $\gamma$  es el peso relativo global de los conteos comparados con los datos de la matriz a priori  $G$ . Los valores de  $\{v_a\}_{a \in A}$  se determinan a partir de los valores de  $g$  y la ecuación de asignación (3.1). Los parámetros  $w_{pq}^G$ ,  $w_a^V$  y  $\gamma$  se calculan a partir de la matriz de varianza–covarianza de los datos. De manera más general, se puede denotar por  $\vec{b} = (b_i)_{i=1}^m$  al vector de  $m$  observaciones,  $\vec{x} = (x_i)_{i=1}^n$  al vector de  $n$  variables a estimar, y

$$A : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$\vec{x} \rightarrow A\vec{x},$$

la relación lineal entre las variables y las observaciones. El problema de mínimos cuadrados generalizados se puede definir como:

$$\min_{\vec{x} \in \mathbb{R}^n} \frac{1}{2} \|W(A\vec{x} - \vec{b})\|_2^2 \quad (3.11)$$

en donde  $W$  es una matriz de  $m \times m$  que contiene los errores de medición. Típicamente se considera a  $W$  como la matriz de varianza–covarianza de las observaciones  $\vec{b}$ . Para mayores detalles de este tipo de modelos se puede consultar [3].

Este tipo de modelos se pueden resolver numéricamente utilizando métodos iterativos de descenso. Actualmente existen métodos muy efectivos para resolver problemas del tipo (3.11) como son los métodos de gradiente conjugado y los métodos de subespacios de Krylov. Sorpresivamente, nos hemos dado cuenta que estos métodos han sido muy poco usados en el ámbito de la ingeniería del transporte, y en algunos casos se utiliza solamente el método básico de descenso máximo (steepest descent, en inglés). Lo anterior nos motiva a explorar la aplicación del método de gradiente conjugado en este tipo de problemas. Finalmente, queremos mencionar que los métodos vistos en las secciones anteriores de este capítulo pueden servir para generar un valor inicial para el método iterativo, aunque en este trabajo no se utiliza esta recomendación, puesto que pensamos que es más natural tomar como valor de comienzo la demanda conocida ‘a priori’, como se muestra en el capítulo siguiente.



# Capítulo 4

## Métodos iterativos para estimación de demanda

Como se mencionó en la sección 3.3, el modelo de mínimos cuadrados generalizados consiste en la minimización de la distancia entre la matriz conocida a priori  $G$  y la matriz calculada  $g$ , y entre los volúmenes observados  $V_a$  y los asignados  $v_a$ . En este capítulo se estudiarán modelos simplificados, en donde se supondrá que  $\omega_{pq}^G = 1$  y  $\omega_a^V = 1$  en la ecuación (3.10). Noriega y Florian [24] consideran además, un promedio ponderado entre la diferencia de volúmenes y la diferencia de demanda de tal manera que la función objetivo que se desea minimizar es:

$$\text{mín } Z(g) = \frac{\alpha}{2} \sum_{a \in \bar{A}} (v(g)_a - V_a)^2 + \frac{1 - \alpha}{2} \sum_{pq \in PQ} (g_{pq} - G_{pq})^2 \quad (4.1)$$

$$\text{Sujeto a: } v(g) = \text{assign}(g) \quad \text{y } g_{pq} \geq 0 \quad (4.2)$$

en donde  $\bar{A}$  es el conjunto de arcos donde se realizan los conteos,  $v_a$  es el volumen asignado con la matriz  $g$  y  $V_a$  es el volumen observado en el arco  $a$ , y  $0 < \alpha \leq 1$ . La función  $\text{assign}(g)$  se utiliza para indicar los volúmenes que resultan de una asignación de tránsito con la matriz de demanda  $g$ .

Aunque la función objetivo de la ecuación (4.1) depende tanto de los volúmenes como de la demanda, no debe perderse de vista que lo que se desea minimizar es la diferencia de demandas y que el término correspondiente a los volúmenes se puede ver como una restricción de igualdad aproximada de volúmenes, es decir, se quiere que los volúmenes calculados  $v(g)_a$  sean tan cercanos a los volúmenes medidos  $V_a$  como sea posible. Por tal razón, pensamos que el problema de mínimos cuadrados puede incluir como término principal el cuadrado de la diferencia de las matrices en la función objetivo e incluir la diferencia de los volúmenes como una restricción que se debe satisfacer tanto como sea posible. Para garantizar la existencia y unicidad de una solución, así como la aplicación inmediata de los métodos iterativos tradicionales, una de las técnicas más simples consiste en penalizar la restricción de igualdad y

agregarla a la función objetivo obteniendo, en nuestro caso, el siguiente problema:

$$\min Z(g) = \frac{1}{2} \sum_{pq \in PQ} (g_{pq} - G_{pq})^2 + \frac{k}{2} \sum_{a \in \bar{A}} (v(g)_a - V_a)^2 \quad (4.3)$$

$$\text{Sujeto a: } v(g) = \text{assign}(g) \quad \text{y } g_{pq} \geq 0 \quad (4.4)$$

en donde  $k$  es un coeficiente de penalización. El efecto que resulta de este procedimiento es que, a mayor penalización (mayor valor de  $k$ ), se exige menor diferencia en el término penalizado. Es posible que las mediciones de los volúmenes en algunos arcos sea más confiables o precisas que en otros, en cuyo caso es posible utilizar diferentes parámetros de penalización  $k_a$  para los diferentes arcos  $a$ , dependiendo del nivel de confianza en los mismos. Sin embargo, en el presente trabajo solo consideraremos el caso en que el parámetro  $k$  toma el mismo valor para todos los arcos, dejando el estudio más general para un posible trabajo futuro.

En el resto del capítulo se estudiará primero un modelo alternativo muy simple, debido a Spiess [29], ya que los métodos propuestos para resolver nuestro modelo están inspirados en la metodología propuesta para resolver dicho problema. Posteriormente se adaptará la metodología a nuestro problema y, finalmente, la se generalizará para poder aplicar el método de gradiente conjugado multiplicativo.

## 4.1. El modelo de Spiess y el método de máximo descenso

### 4.1.1. El modelo de Spiess

Spiess [29] propone minimizar solamente la distancia entre los volúmenes observados y los asignados, el cual corresponde a tomar  $\alpha = 1$  en el modelo (4.1). Este modelo es muy simple y consiste en resolver:

$$\min Z(\mathbf{g}) = \frac{1}{2} \sum_{a \in A} (v_a(\mathbf{g}) - V_a)^2 \quad (4.5)$$

$$\text{Sujeto a: } v(g) = \text{assign}(g) \quad \text{y } g_{pq} \geq 0 \quad (4.6)$$

En la expresión anterior  $V_a$  son los volúmenes observados en un conjunto de arcos  $A$  de la red, y los volúmenes  $v_a(\mathbf{g})$  son los resultantes de hacer una asignación con la matriz de demanda  $\mathbf{g} = \{g_{pq}\}_{gp \in PQ}$ . El modelo de asignación “*assign*( $\mathbf{g}$ )” que se utiliza en (4.6) debe corresponder a un problema de optimización convexa. Vale la pena aclarar que el modelo (4.5)–(4.6) fue propuesto antes del modelo (4.1)–(4.2) y se estudiará primero, ya que es conveniente tomarlo como punto de partida para generar los métodos utilizados en la solución del modelo más general, el cuál constituye el tema central de este trabajo.

En esta sección la palabra “asignación” se refiere a una asignación de equilibrio, en donde se supone que se cuenta con un conjunto de funciones no decrecientes de costo en las aristas

$t_a(v_a)$  de todos los arcos de la red  $a \in A$  que aseguran la convexidad del modelo. Este tipo de problemas de asignación de equilibrio han sido estudiados ampliamente y se pueden resolver de manera eficiente, ya sea por aproximación lineal sucesiva [14], [11], o por el método Partan [12]. Dado que el problema de estimar la matriz, tal como se formula en (4.5), es altamente indeterminado, se suele admitir un número infinito de soluciones óptimas. Por supuesto, en el contexto de planificación real, se espera que la matriz resultante se parezca lo más posible a la matriz inicial, ya que contiene información estructural importante en los movimientos origen–destino.

### 4.1.2. El método de descenso

Existen varios métodos iterativos de descenso (también llamados de tipo gradiente) para encontrar el mínimo de una función diferenciable, suponiendo que se conoce un entorno del punto donde ocurre el mínimo. Estos métodos parten de un valor inicial, a partir de la cual se encuentran una sucesión de valores (descendentes), y que se espera sean cada vez más cercanos al mínimo. Los métodos de descenso se distinguen entre sí por la dirección en la que se escoge descender en cada iteración, y por el tamaño del paso que asegure un buen descenso. Uno de los métodos más sencillos e intuitivos es el método de descenso máximo (steepest descent), en donde se toma como dirección de descenso la opuesta al gradiente, por ser la de máximo decrecimiento de la función en cada punto. Este método aplicado al problema (4.5)–(4.6), cuando se toma directamente el gradiente con respecto a la variable  $g$  se escribe de la siguiente manera:

$$g_{pq}^{l+1} = \begin{cases} G_{pq} & \text{para } l = 0, \\ g_{pq}^l - \lambda^l \left[ \frac{\partial Z(g)}{\partial g_{pq}} \right]_{g_{pq}^l} & \text{para } l = 1, 2, \dots \end{cases} \quad (4.7)$$

en donde el tamaño del paso  $\lambda^l$  debe tomarse lo suficientemente pequeña para asegurar un descenso en cada iteración, es decir para que  $Z(g^{l+1}) \leq Z(g^l)$ . Si el gradiente se expresa en las variables  $g$  como en la ecuación (4.7), esto implica que los cambios en la matriz de demanda se miden de un modo absoluto. Esto implicaría que los pares O–D con  $g_{pq} = 0$  también se verán afectados en el ajuste. Para una aproximación más realista el gradiente deberá basarse en los cambios en la demanda, esto puede escribirse como:

$$g_{pq}^{l+1} = \begin{cases} G_{pq} & \text{para } l = 0, \\ g_{pq}^l \left( 1 - \lambda^l \left[ \frac{\partial Z(g)}{\partial g_{pq}} \right]_{g_{pq}^l} \right) & \text{para } l = 1, 2, \dots \end{cases} \quad (4.8)$$

Nótese que cuando se utiliza el gradiente relativo, el algoritmo se vuelve multiplicativo. Por lo tanto, un cambio en la demanda es proporcional a la demanda en la matriz inicial y, en particular, los ceros se conservarán en el proceso, [29].

Para completar el algoritmo (4.8), es necesario calcular el gradiente en términos de los flujos de arco, para lo cual es necesario establecer la relación entre los flujos de arco y los

flujos de ruta, como se indica a continuación.

Sea  $S_{pq}$  el conjunto de rutas usadas para cada par O-D,  $pq \in PQ$ , y  $h_s$  el vector correspondiente al flujo de ruta  $s \in S_{pq}$ . El volumen de arista puede expresarse como:

$$v_a(\mathbf{g}) = \sum_{pq \in PQ} \sum_{s \in S_{pq}} \delta_{as} h_s, \quad a \in A \quad (4.9)$$

en donde  $\delta_{as}$  es la matriz de incidencia entre flujos de arco y flujos de ruta, definida por

$$\delta_{as} = \begin{cases} 0 & \text{si } a \notin s \text{ (la arista } a \text{ no se encuentra en la ruta } s) \\ 1 & \text{si } a \in s \text{ (la arista } a \text{ está contenida en la ruta } s) \end{cases} \quad (4.10)$$

Usando las probabilidades de ruta en lugar de los flujos de ruta

$$\pi_s = \frac{h_s}{g_{pq}}, \quad s \in S_{pq}, \quad pq \in PQ \quad (4.11)$$

la ecuación (4.9) puede reescribirse de la siguiente manera:

$$v_a(\mathbf{g}) = \sum_{pq \in PQ} g_{pq} \sum_{s \in S_{pq}} \delta_{as} \pi_s, \quad a \in A \quad (4.12)$$

En cada iteración del algoritmo de descenso, introducido por Spiess [29], se utiliza como dirección de descenso un vector  $\mathbf{d}$ , cuyas componentes son de la forma:

$$d_{pq} = -g_{pq} \frac{\partial Z(\mathbf{g})}{\partial g_{pq}}, \quad pq \in PQ. \quad (4.13)$$

en donde es posible calcular la derivada parcial usando la regla de la cadena, obteniendo:

$$\frac{\partial Z(\mathbf{g})}{\partial g_{pq}} = \frac{\partial Z(\mathbf{g})}{\partial v_a} \frac{\partial v_a}{\partial g_{pq}} = \sum_{a \in A} \frac{\partial v_a(\mathbf{g})}{\partial g_{pq}} (v_a(\mathbf{g}) - V_a), \quad pq \in PQ \quad (4.14)$$

Asumiendo que las probabilidades sobre cada ruta son constantes, de la ecuación (4.12) se obtiene:

$$\frac{\partial v_a}{\partial g_{pq}} = \sum_{s \in S_{pq}} \delta_{as} \pi_s, \quad a \in A, \quad pq \in PQ. \quad (4.15)$$

Sustituyendo la ecuación (4.15) en la ecuación (4.14) se obtiene:

$$\frac{\partial Z(\mathbf{g})}{\partial g_{pq}} = \sum_{a \in A} \sum_{s \in S_{pq}} \delta_{as} \pi_s (v_a - V_a) = \sum_{s \in S_{pq}} \pi_s \sum_{a \in A} \delta_{as} (v_a - V_a), \quad pq \in PQ \quad (4.16)$$

Para implementar el método de máximo descenso (4.8), es necesario dar un tamaño de paso adecuado  $\lambda^l$ . Si se eligen pasos muy pequeños se asegura que  $Z(\mathbf{g})$  disminuya en cada



iteración, sin embargo se requiere un gran número de pasos para llegar al mínimo. Por el contrario, si se eligen tamaños de paso muy grandes, es posible que  $Z(\mathbf{g})$  vaya creciendo y se pierda la convergencia. Así, el tamaño de paso óptimo  $\lambda^*$  dada una demanda  $\mathbf{g}$  puede calcularse resolviendo el siguiente sub-problema de minimización unidimensional.

$$\min_{\lambda} \phi(\lambda) = Z \left( g_{pq} \left( 1 - \lambda \frac{\partial Z(\mathbf{g})}{\partial g_{pq}} \right) \right) \quad (4.17)$$

$$\text{sujeto a: } \lambda \frac{\partial Z(\mathbf{g})}{\partial g_{pq}} \leq 1, \quad \forall pq \in PQ, \text{ con } g_{pq} > 0 \quad (4.18)$$

Por tal motivo, es de interés evaluar los volúmenes de arco con la “nueva matriz de demanda”  $\mathbf{g} + \lambda \mathbf{d}$ . De acuerdo a (4.12) se tiene:

$$v_a(\mathbf{g} + \lambda \mathbf{d}) = \sum_{pq \in PQ} (g_{pq} + \lambda d_{pq}) \sum_{s \in S_{pq}} \delta_{as} \pi_s, \quad \text{en donde ahora } \pi_s = \frac{h_s}{g_{pq} + \lambda d_{pq}} \quad (4.19)$$

Esta última expresión se puede simplificar si se calcula  $\pi_s$  como en (4.11), pues en este caso se podría escribir:

$$v_a(\mathbf{g} + \lambda \mathbf{d}) = v_a(\mathbf{g}) + \lambda v_a(\mathbf{d}), \quad (4.20)$$

debido a que  $\pi_s$  sería constante (no dependería de  $\lambda \mathbf{d}$ ). La simplificación anterior es factible debido a que, en un algoritmo de descenso, la diferencia de la demanda entre iteraciones sucesivas tiende a cero cerca del óptimo. De cualquier forma, la aproximación anterior también puede pensarse como el término de primer orden de la expansión de Taylor de  $v_a(\mathbf{g} + \lambda \mathbf{d})$  respecto de  $\lambda$ , con un error cuadrático proporcional a  $\|\lambda \mathbf{d}\|^2$ . Utilizando (4.20) es fácil estimar la variación de los volúmenes con respecto al paso  $\lambda$ , (suponiendo conocidas  $\mathbf{g}$  y  $\mathbf{d}$ ):

$$v'_a(\lambda) = \frac{dv_a}{d\lambda} \approx v_a(\mathbf{d}). \quad (4.21)$$

Para estimar el paso óptimo  $\lambda^*$  en el algoritmo de descenso máximo se necesita minimizar la función escalar

$$\phi(\lambda) = \frac{1}{2} \sum_{a \in A} (v_a(\mathbf{g} + \lambda \mathbf{d}) - V_a)^2 \approx \frac{1}{2} \sum_{a \in A} (v_a(\mathbf{g}) + \lambda v_a(\mathbf{d}) - V_a)^2. \quad (4.22)$$

Está claro que el mínimo se alcanza en donde la derivada respecto de  $\lambda$  es igual a cero

$$\phi'(\lambda) \approx \sum_{a \in A} (v_a(\mathbf{g}) + \lambda v_a(\mathbf{d}) - V_a) v_a(\mathbf{d}) = 0. \quad (4.23)$$

Despejando  $\lambda$  de la última ecuación, se obtiene la siguiente estimación para el valor óptimo del paso

$$\lambda^* \approx \frac{\sum_{a \in A} v_a(\mathbf{d})(V_a - v_a(\mathbf{g}))}{\sum_{a \in A} v_a(\mathbf{d})^2}. \quad (4.24)$$

el cual, para que sea una solución factible, debe cumplir la restricción (4.18).

Nótese que con la restricción  $\lambda \frac{\partial Z(\mathbf{g})}{\partial g_{pq}} \leq 1$  y la fórmula iterativa de Spiess (4.8) se garantiza que la solución  $g_{pq} \geq 0$ . Con las ecuaciones anteriores, se tienen los resultados necesarios para resolver el problema de ajuste (4.5) utilizando el método de descenso máximo con un gradiente relativo (4.8).

### 4.1.3. Máximo descenso en el modelo que agrega la diferencia de demanda

Si se desea mantener la estructura de la matriz O-D durante el ajuste, es importante considerar el segundo término de la función objetivo (4.1):

$$Z(\mathbf{g}) = \frac{1}{2} \sum_{pq \in PQ} (g_{pq} - G_{pq})^2 + \frac{k}{2} \sum_{a \in A} (v_a(\mathbf{g}) - V_a)^2. \quad (4.25)$$

en donde  $\mathbf{G} = \{G_{pq}\}_{pq \in PQ}$  es la matriz conocida “a priori”,  $\mathbf{g}$  es la demanda a calcular, y  $k > 0$  es un factor de penalización. De esta forma, el gradiente de la función objetivo consistirá en dos partes, la primera correspondiente a la diferencia de volúmenes dada por la ecuación (4.16), y la segunda dada por la diferencia en la demanda. Por lo tanto, el gradiente en este caso está formado por las siguientes derivadas parciales:

$$\frac{\partial Z(\mathbf{g})}{\partial g_{pq}} = (g_{pq} - G_{pq}) + k \sum_{s \in S_{pq}} \pi_s \sum_{a \in A} \delta_{as} (v_a - V_a), \quad pq \in PQ \quad (4.26)$$

Para garantizar que las entradas de la matriz  $g_{pq} = 0$  permanezcan sin modificaciones, es conveniente multiplicar la dirección de descenso  $-\nabla Z(\mathbf{g})$  por la demanda  $\mathbf{g}$ , obteniendo:

$$\mathbf{d} = -\mathbf{g} \cdot * \nabla Z(\mathbf{g}),$$

en donde el símbolo  $\cdot *$  indica el producto entrada por entrada, es decir:

$$d_{pq} = -g_{pq} \left[ (g_{pq} - G_{pq}) + k \sum_{s \in S_{pq}} \pi_s \sum_{a \in A} \delta_{as} (v_a - V_a) \right], \quad pq \in PQ \quad (4.27)$$

Para calcular el tamaño óptimo del paso en el algoritmo de descenso es necesario minimizar la siguiente función escalar:

$$\phi(\lambda) = Z(\mathbf{g} + \lambda \mathbf{d}) \approx \frac{1}{2} \sum_{pq \in PQ} (g_{pq} + \lambda d_{pq} - G_{pq})^2 + \frac{k}{2} \sum_{a \in A} (v_a(\mathbf{g}) + \lambda v_a(\mathbf{d}) - V_a)^2 \quad (4.28)$$

$$\text{sujeto a: } \lambda \mathbf{d} \leq 1, \quad \forall pq \in PQ, \text{ con } g_{pq} > 0 \quad (4.29)$$

La estimación de la derivada respecto de  $\lambda$  es claramente:

$$\phi'(\lambda) \approx \sum_{pq \in PQ} (g_{pq} + \lambda d_{pq} - G_{pq}) d_{pq} + k \sum_{a \in A} (v_a(\mathbf{g}) + \lambda v_a(\mathbf{d}) - V_a) v_a(\mathbf{d}) \quad (4.30)$$

Igualando a cero la expresión anterior y despejando  $\lambda$ , se obtiene el tamaño óptimo del paso:

$$\lambda^* \approx \frac{\sum_{pq \in PQ} d_{pq}(G_{pq} - g_{pq}) + k \sum_{a \in A} v_a(\mathbf{d})(V_a - v_a(\mathbf{g}))}{\sum_{pq \in PQ} d_{pq}^2 + k \sum_{a \in A} v_a(\mathbf{d})^2}. \quad (4.31)$$

El problema puede verse de manera más general para diversos modos de transporte, esto tiene la ventaja de que se pueden ajustar varias matrices para diferentes modos de transporte de manera simultánea, [24]. Considérese el problema de ajustar matrices para diversos modos de transporte  $M$ . La matriz de demanda del nodo  $p$  al nodo  $q$  para el modo de transporte  $m \in M$  se denota por  $g_{pq}^m$ .  $\bar{A}^m \subset A$  es el conjunto de arcos donde se tienen conteos del modo  $m$ . En este caso, la función objetivo toma la siguiente forma:

$$Z(\mathbf{g}) = \frac{1}{2} \sum_{m \in M} \sum_{pq \in PQ} (g_{pq}^m - G_{pq}^m)^2 + \frac{k}{2} \sum_{m \in M} \sum_{a \in \bar{A}^m} (v_a^m(\mathbf{g}) - V_a^m)^2 \quad (4.32)$$

y de manera análoga a los pasos anteriores, se obtienen la dirección de descenso:

$$d_{pq}^m = -g_{pq}^m \left[ (g_{pq}^m - G_{pq}^m) + k \sum_{s \in S_{pq}^m} \pi_s^m \sum_{a \in \bar{A}^m} \delta_{as}^m (v_a^m - V_a^m) \right], \quad pq \in PQ, \quad m \in M \quad (4.33)$$

por lo que el tamaño de paso óptimo para cada modo de transporte se estima por medio de:

$$\lambda^{m*} \approx \frac{\sum_{pq \in PQ} d_{pq}^m (G_{pq}^m - g_{pq}^m) + k \sum_{a \in \bar{A}^m} v_a^m(\mathbf{d})(V_a^m - v_a^m(\mathbf{g}))}{\sum_{pq \in PQ} (d_{pq}^m)^2 + k \sum_{a \in \bar{A}^m} v_a^m(\mathbf{d})^2}. \quad (4.34)$$

Hasta este momento se ha desarrollado el método de máximo descenso para resolver el problema de ajuste de demanda, pero es posible utilizar otros métodos, como el de gradiente conjugado.

## 4.2. El método de gradiente conjugado multiplicativo

Es bien conocido que el método de descenso máximo puede llegar a ser muy ineficiente en la práctica, debido al problema de “zig-zagueo”, el cual ocurre principalmente con problemas mal condicionados, y que ocasiona que se requiera de un gran número de iteraciones para acercarse al óptimo. El método de Newton es mucho más eficiente en esos casos, sin embargo es muy caro debido a la necesidad de evaluar Hessianos y resolver sistemas de ecuaciones lineales en cada iteración. Un método intermedio, en el que no ocurren ninguna de estas desventajas es el método de gradiente conjugado. El método de gradiente conjugado puede adaptarse para evitar la evaluación de Hessianos, a la vez que reduce el número de iteraciones,

con un costo comparable al método de descenso máximo en cada iteración. En esta sección se introduce el algoritmo de gradiente conjugado para funciones cuadráticas y su generalización a funciones no cuadráticas y se explican algunas de sus propiedades. Al final se introduce una variante de gradiente conjugado que puede utilizarse para minimizar la función objetivo (4.25).

### 4.2.1. Introducción del algoritmo de gradiente conjugado

El método de gradiente conjugado se introdujo en 1952 por Hestenes y Stiefel, [17], y es considerado uno de los métodos de descenso más populares e importantes del siglo XX para problemas de gran escala [32]. En este método la dirección de descenso se genera como una combinación lineal de la dirección anterior y el gradiente actual, es decir  $\mathbf{d}^{l+1} = -\nabla f(\mathbf{x}^{l+1}) + \beta^l \mathbf{d}^l$ , donde  $\beta^l$  se calcula en términos de información conocida en cada iteración.

El método de gradiente conjugado y sus propiedades invariablemente se estudian para el caso especial de minimizar una función cuadrática, que en forma vectorial puede escribirse como:

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T Q \mathbf{x} - \mathbf{x}^T \mathbf{b}, \quad \text{con } \mathbf{x}, \mathbf{b} \in \mathbb{R}^n, \quad Q \in \mathbb{R}^{n \times n}, \quad (4.35)$$

en donde  $\mathbf{b}$  es dado y  $Q$  es una matriz simétrica y definida positiva (la Hessiana de  $f$ ). El algoritmo de gradiente conjugado básico para minimizar la función mutidimensional anterior se puede escribir de la siguiente manera:

1. **Inicialización:** Dado  $\mathbf{x}^0$ , se toma  $\mathbf{d}^0 = -\nabla f(\mathbf{x}^0)$ .
2. **Avance:** Para  $l \geq 0$ , suponiendo conocidos  $\mathbf{x}^l, \mathbf{d}^l$ , calcular  $\mathbf{x}^{l+1}, \mathbf{d}^{l+1}$  por medio de:
 
$$\mathbf{x}^{l+1} = \mathbf{x}^l + \lambda^l \mathbf{d}^l, \quad \text{en donde } \lambda^l \text{ minimiza } \phi(\lambda) = f(\mathbf{x}^l + \lambda \mathbf{d}^l) \quad (4.36)$$
3. **Prueba de convergencia:** Para  $0 < \epsilon \ll 1$ , si  $\nabla f(\mathbf{x}^{l+1}) \cdot \nabla f(\mathbf{x}^{l+1}) \leq \epsilon \nabla f(\mathbf{x}^0) \cdot \nabla f(\mathbf{x}^0)$ , ó bien si  $l \geq l_{max}$ , entonces parar y salir. En caso contrario hacer lo siguiente:
4. **Nueva dirección conjugada:**

$$\mathbf{d}^{l+1} = -\nabla f(\mathbf{x}^{l+1}) + \beta^l \mathbf{d}^l, \quad \text{en donde } \beta^l = \frac{(\nabla f(\mathbf{x}^{l+1}))^T Q \mathbf{d}^l}{(\mathbf{d}^l)^T Q \mathbf{d}^l} \quad (4.37)$$

Hacer  $l = l + 1$  e ir a 2.

Como la función es cuadrática, es posible calcular  $\lambda^l$  en forma exacta, obteniendo:

$$\lambda^l = \frac{-\nabla f(\mathbf{x}^l)^T Q \mathbf{d}^l}{(\mathbf{d}^l)^T Q \mathbf{d}^l}. \quad (4.38)$$

El algoritmo de gradiente conjugado también puede utilizarse para funciones que no sean cuadráticas, sustituyendo  $Q$  por la Hessiana de  $f$  evaluada en  $\mathbf{x}^{l+1}$  en (4.37). Sin embargo,

es ineficiente evaluar Hessianos, especialmente en problemas de gran escala. Incluso para funciones cuadráticas, en donde no es necesario evaluar Hessianos, es ineficiente multiplicar por la matrix  $Q$  para calcular los parámetros  $\lambda^l$  en (4.38) y  $\beta^l$  en (4.37). Por esta razón, para calcular dichos parámetros, generalmente se utilizan variantes en las que solamente se utilizan evaluaciones del gradiente de la función. Dichas variantes son más fáciles de obtener para el caso de funciones cuadráticas, debido a que en este caso  $\nabla f(x) = Q\mathbf{x} - \mathbf{b}$ , y se cumplen las siguientes propiedades, [23]:

**P1.**  $\nabla f(\mathbf{x}^{l+1}) = \nabla f(\mathbf{x}^l) + \lambda^l Q \mathbf{d}^l$  (al final de la sección se indica como obtenerla).

**P2.**  $\nabla f(\mathbf{x}^{l+1})$  es ortogonal a  $\nabla f(\mathbf{x}^l)$  para todo  $l$ .

**P3.**  $\nabla f(\mathbf{x}^{l+1})$  es ortogonal a  $\mathbf{d}^l$  para todo  $l$ .

Las variantes que comúnmente se utilizan en la práctica son:

**HS.** Fórmula de **Hestenes–Stiefel**, que se obtiene de despejar  $Q\mathbf{d}^l$  en P1 y sustituir en (4.37):

$$\beta_{l+1} = \frac{(\nabla f(\mathbf{x}^{l+1}))^T (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}{(\mathbf{d}^l)^T (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}$$

**PR.** Fórmula de **Polak–Ribière**, se obtiene de aplicar en el denominador de la expresión anterior la propiedad **P3** primero, y después utilizar que  $\mathbf{d}^l = -\nabla f(\mathbf{x}^l) + \beta_{l-1}\mathbf{d}^{l-1}$ :

$$\beta_{l+1} = \frac{(\nabla f(\mathbf{x}^{l+1}))^T (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}{(\nabla f(\mathbf{x}^l))^T \nabla f(\mathbf{x}^l)}$$

**FR.** Fórmula de **Fletcher–Reeves**, se obtiene de aplicar la propiedad **P2** a la fórmula anterior:

$$\beta_{l+1} = \frac{(\nabla f(\mathbf{x}^{l+1}))^T \nabla f(\mathbf{x}^{l+1})}{(\nabla f(\mathbf{x}^l))^T \nabla f(\mathbf{x}^l)}$$

### 4.2.2. Aspectos para construir del algoritmo multiplicativo

Para el caso de funciones cuadráticas se sabe que el número de iteraciones requeridas para llegar al óptimo es a lo más  $n$  (el número de variables), en aritmética exacta. En algunos casos, con un buen preconditionador, es posible acelerar dramáticamente la convergencia. Estas propiedades, y las mencionadas anteriormente, hacen del algoritmo de gradiente conjugado una muy buena opción, especialmente para problemas de gran escala. Por esta razón, se ha considerado la posibilidad de derivar un algoritmo del tipo gradiente conjugado y que sea adecuado para resolver el problema de minimización de la función objetivo (4.25).

Para derivar el algoritmo se mantendrá la notación de la sección anterior, es decir se sigue considerando el problema de encontrar el mínimo de una función cuadrática  $f(\mathbf{x})$  de la forma (4.35), con  $\mathbf{x} \in \mathbb{R}^n$ . Posteriormente, el algoritmo encontrado se adaptará al problema

asociado a (4.25). En esta ocasión, se supondrá que las componentes de  $\mathbf{x}$  (y de  $\mathbf{x}^l$  en cada iteración) son no negativas (factibles). Además, se quiere que las componentes cero de  $\mathbf{x}^l$  se mantengan como cero en las iteraciones posteriores. Por lo tanto, siguiendo la idea de Spiess, se propone la fórmula de iteración siguiente:

$$x_i^{l+1} = x_i^l + \lambda^l x_i^l d_i^l = x_i^l(1 + \lambda^l d_i^l), \quad l = 1, \dots, n. \quad (4.39)$$

cuya forma vectorial se puede escribir como:

$$\mathbf{x}^{l+1} = \mathbf{x}^l + \lambda^l \mathbf{d}_M^l, \quad (4.40)$$

en donde el vector  $\mathbf{d}_M^l$  tiene como componentes a  $x_i^l d_i^l$ , para  $i = 1, \dots, n$ . La letra  $M$  se utilizará para indicar que es un vector dirección modificado de la dirección  $Q$ -conjugada  $\mathbf{d} = \{d_i\}_{i=1}^n$  del algoritmo de gradiente conjugado original. El parámetro  $\lambda^l$  se calcula encontrando el mínimo de la función unidimensional  $\phi(\lambda) = f(\mathbf{x}^l + \lambda \mathbf{d}_M^l)$ .

En el algoritmo de gradiente conjugado, después de calcular  $\mathbf{x}^{l+1}$ , se calcula la nueva dirección conjugada utilizando la fórmula (4.37). Entonces, para utilizar la misma idea con  $\mathbf{d}_M^l$  en lugar de  $\mathbf{d}^l$ , se propone lo siguiente:

$$\mathbf{d}_M^{l+1} = -\nabla_M f(\mathbf{x}^{l+1}) + \beta^l \mathbf{d}_M^l, \quad \text{en donde las componentes de } \nabla_M f(\mathbf{x}^{l+1}) \text{ son } x_i^{l+1} \frac{\partial f(\mathbf{x}^{l+1})}{\partial x_i}. \quad (4.41)$$

El valor de  $\beta^l$  para que  $\mathbf{d}_M^l$  y  $\mathbf{d}_M^{l+1}$  sean  $Q$ -ortogonales ahora debe ser:

$$\beta^l = \frac{(\nabla_M f(\mathbf{x}^{l+1}))^T Q \mathbf{d}_M^l}{(\mathbf{d}_M^l)^T Q \mathbf{d}_M^l}. \quad (4.42)$$

Esta expresión para  $\beta^l$  difiere del valor en (4.37) en general, debido a que

$$\frac{(\nabla_M f(\mathbf{x}^{l+1}))^T Q \mathbf{d}_M^l}{(\mathbf{d}_M^l)^T Q \mathbf{d}_M^l} = \frac{(\nabla f(\mathbf{x}^{l+1}))^T Q_M^{l+1,l} \mathbf{d}^l}{(\mathbf{d}^l)^T Q_M^{l,l} \mathbf{d}^l},$$

en donde  $Q_M^{l+1,l} = \{x_i^{l+1} x_j^l q_{ij}\}$ , y  $Q_M^{l,l} = \{x_i^l x_j^l q_{ij}\}$ , siendo  $q_{ij}$  los coeficientes de la matrix  $Q$ .

Con el objeto de evitar el calculo del producto matriz por vector en (4.42), se tratará de hacer algo semejante a lo realizado para obtener la fórmula de Hestenes–Stiefel. Se parte de la expresión (4.40), la cual al multiplicar por  $Q$  y después restar  $\mathbf{b}$  se obtiene la siguiente relación:

$$\nabla f(\mathbf{x}^{l+1}) = \nabla f(\mathbf{x}^l) + \lambda^l Q \mathbf{d}_M^l \quad (4.43)$$

Despejando  $Q \mathbf{d}_M^l$  de esta igualdad y sustituyendo en (4.42), se obtiene el análogo a la fórmula de Hestenes–Stiefel para el cálculo de  $\beta^l$  en el nuevo algoritmo de gradiente conjugado

$$\beta^l = \frac{\nabla_M f(\mathbf{x}^{l+1}) \cdot (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}{\mathbf{d}_M^l \cdot (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}. \quad (4.44)$$

Se podría intentar derivar fórmulas análogas a las de Polak–Ribière y Fletcher–Reeves, pero en este caso se debe de ser cuidadoso, pues primero se debería demostrar los análogos a las propiedades **P2** y **P3**. Por el momento, se dejará para un posible análisis posterior.

### 4.2.3. El algoritmo de gradiente conjugado multiplicativo

Tomando en cuenta los aspectos estudiados en la sección anterior, se propone el siguiente algoritmo de gradiente conjugado para minimizar una función  $f(x)$ , no necesariamente cuadrática, y definida para  $\mathbf{x} \in \mathbb{R}_+^n$ . El algoritmo preserva las entradas cero del dato inicial  $\mathbf{x}^0$  a lo largo de las iteraciones. Recordar que el operador  $\nabla_M$  se define como en (4.41).

1. **Inicialización:** Dado  $\mathbf{x}^0$ , tomar  $\mathbf{d}_M^0 = -\nabla_M f(\mathbf{x}^0)$ .
2. **Avance:** Para  $l \geq 0$ , suponiendo conocidos  $\mathbf{x}^l$  y  $\mathbf{d}_M^l$ , calcular  $\mathbf{x}^{l+1}$  y  $\mathbf{d}_M^{l+1}$ , por medio de:
 
$$\mathbf{x}^{l+1} = \mathbf{x}^l + \lambda^l \mathbf{d}_M^l, \quad \text{en donde } \lambda^l \text{ minimiza } \phi(\lambda) = f(\mathbf{x}^l + \lambda \mathbf{d}_M^l). \quad (4.45)$$
3. **Prueba de convergencia:** Para  $0 < \epsilon \ll 1$ , si  $\|\nabla f(\mathbf{x}^{l+1})\|^2 \leq \epsilon \|\nabla f(\mathbf{x}^0)\|^2$ , ó bien si  $l \geq l_{max}$ , entonces parar y salir. En caso contrario hacer lo siguiente:

#### 4. Nueva dirección conjugada

$$\mathbf{d}_M^{l+1} = -\nabla_M f(\mathbf{x}^{l+1}) + \beta^l \mathbf{d}_M^l,$$

en donde:

$$\beta^l = \frac{(\nabla_M f(\mathbf{x}^{l+1}))^T (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}{(\mathbf{d}_M^l)^T (\nabla f(\mathbf{x}^{l+1}) - \nabla f(\mathbf{x}^l))}. \quad (4.46)$$

Hacer  $l = l + 1$  e ir a 2.

En base a lo anterior, se tienen las siguientes observaciones sobre el algoritmo:

- Por la forma en que se construyó el algoritmo, está claro que si el vector  $\mathbf{x}^0$  tiene algunas componentes igual a cero, dichas componentes se mantienen nulas en cada iteración.
- Es conveniente calcular  $\lambda^l$  en forma exacta, siempre que esto sea posible, ya que el algoritmo de gradiente conjugado multiplicativo puede ser sensible a la estimación de este parámetro. Cuando no sea posible calcular  $\lambda^l$  en forma exacta, el resultado dependerá del algoritmo de búsqueda de línea, del valor inicial para  $\lambda$  en dicho algoritmo y del intervalo de búsqueda.
- Es necesario hacer un análisis más exhaustivo de las propiedades del algoritmo de gradiente conjugado multiplicativo. Por ejemplo, es posible asegurar que, en este algoritmo, dos direcciones consecutivas  $\mathbf{d}_M^l$  y  $\mathbf{d}_M^{l+1}$  son  $Q$ -ortogonales, pero falta verificar si estos a su vez son  $Q$ -conjugados con los demás. De cualquier forma, debido al error de redondeo por la utilización de aritmética de punto flotante, la  $Q$ -ortogonalidad no se satisface de manera exacta y es posible que se degrade conforme avanzan las iteraciones. Si es necesario, es posible reinicializar el algoritmo.

#### 4.2.4. Aplicación del algoritmo de gradiente conjugado multiplicativo al problema de estimación de matrices de demanda

La función objetivo (4.25) tiene términos cuadráticos y se comporta bien cuando se le aplica el algoritmo de máximo descenso multiplicativo de Spiess, por lo que es natural pensar que se comportará bien con otros algoritmos iterativos de tipo gradiente. En particular, es posible aplicar el algoritmo de gradiente conjugado multiplicativo. La adaptación al problema de demanda es directa, dado que en lugar de la función  $f$  tenemos la función objetivo  $Z$ ; y en lugar del vector  $\mathbf{x}^l$  se tiene  $\mathbf{g}^l = \{g_{pq}^l\}_{pq \in PQ}$  (la matriz de demanda). Afortunadamente, para el problema de demanda el parámetro  $\lambda_l$  se puede estimar con suficiente precisión, como ya se demostró anteriormente. El subíndice  $M$  en los vectores de descenso  $\mathbf{d}_M^k$  se suprimirá con el objeto de simplificar la notación.

Haciendo un pequeño paréntesis, se puede recalcar que el método de gradiente conjugado está pensado para matrices Hessianas simétricas. Para el problema de demanda con la función objetivo (4.25) se tiene:

$$Q(\mathbf{g}) = \begin{pmatrix} \sum_{s \in S_{11}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{11}} \delta_{as} \pi_s + 1 & \sum_{s \in S_{11}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{12}} \delta_{as} \pi_s & \dots & \sum_{s \in S_{11}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{1n}} \delta_{as} \pi_s \\ \sum_{s \in S_{12}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{11}} \delta_{as} \pi_s & \sum_{s \in S_{12}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{12}} \delta_{as} \pi_s + 1 & \dots & \sum_{s \in S_{12}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{1n}} \delta_{as} \pi_s \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{s \in S_{nn}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{11}} \delta_{as} \pi_s & \sum_{s \in S_{nn}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{12}} \delta_{as} \pi_s & \dots & \sum_{s \in S_{nn}} \sum_{a \in A} \delta_{as} \sum_{s \in S_{nn}} \delta_{as} \pi_s + 1 \end{pmatrix}$$

y como se puede ver, la matriz  $Q(\mathbf{g})$  es simétrica.

Volviendo al algoritmo de gradiente conjugado, el valor y la dirección conjugada iniciales se pueden escoger como:

$$g_{pq}^0 = G_{pq} \quad \text{y} \quad d_{pq}^0 = -g_{pq}^0 \frac{\partial Z(\mathbf{g}^0)}{\partial g_{pq}}, \quad pq \in PQ. \quad (4.47)$$

La actualización de la demanda se realiza mediante las iteraciones:

$$g_{pq}^{l+1} = g_{pq}^l + \lambda^l d_{pq}^l, \quad pq \in PQ, \quad (4.48)$$

en donde el parámetro  $\lambda^l$  se estima encontrando el mínimo de la función  $\phi(\lambda) = Z(\mathbf{g}^l + \lambda \mathbf{d}^l)$ . De acuerdo a (4.24), este parámetro puede estimarse mediante la fórmula

$$\lambda^l \approx \frac{\sum_{pq \in PQ} d_{pq}^l (G_{pq} - g_{pq}^l) + k \sum_{a \in A} v_a(\mathbf{d}^l) (V_a - v_a(\mathbf{g}^l))}{\sum_{pq \in PQ} (d_{pq}^l)^2 + k \sum_{a \in A} v_a(\mathbf{d}^l)^2}. \quad (4.49)$$



La nueva dirección conjugada se calcula mediante una combinación del gradiente “más actual”  $-\nabla_M Z(\mathbf{g}^{l+1})$  y la dirección anterior  $\mathbf{d}^l$ , es decir:

$$d_{pq}^{l+1} = -g_{pq}^{l+1} \frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} + \beta^l d_{pq}^l, \quad pq \in PQ, \quad (4.50)$$

en donde el parámetro  $\beta^l$  se calcula para que las dos direcciones  $\mathbf{d}^l$  y  $\mathbf{d}^{l+1}$  sean conjugadas. Esto se logra por medio de la fórmula del tipo Hestenes–Stiefel (4.44)

$$\beta^l = \frac{(\nabla_M Z(\mathbf{g}^{l+1}))^T (\nabla Z(\mathbf{g}^{l+1}) - \nabla Z(\mathbf{g}^l))}{(\mathbf{d}^l)^T (\nabla Z(\mathbf{g}^{l+1}) - \nabla Z(\mathbf{g}^l))}, \quad (4.51)$$

es decir

$$\beta^l = \frac{\sum_{pq \in PQ} g_{pq}^{l+1} \frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} \left( \frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} - \frac{\partial Z(\mathbf{g}^l)}{\partial g_{pq}} \right)}{\sum_{pq \in PQ} d_{pq}^l \left( \frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} - \frac{\partial Z(\mathbf{g}^l)}{\partial g_{pq}} \right)}. \quad (4.52)$$

Con el objeto de simplificar el cálculo de  $\beta^l$  a continuación se establece la relación entre el gradiente de la función objetivo en dos iteraciones sucesivas. Utilizando la regla de la cadena sobre la función objetivo (4.25), se obtiene:

$$\frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} = (g_{pq}^{l+1} - G_{pq}) + k \sum_{a \in A} (v_a(\mathbf{g}^{l+1}) - V_a) \sum_{s \in S_{pq}} \delta_{as} \pi_s.$$

Para simplificar, las probabilidades de ruta  $\pi_s$  en la expresión anterior se evalúan en la demanda actual  $\mathbf{g}^l$  en lugar de la nueva demanda  $\mathbf{g}^{l+1}$ . Además, sustituyendo  $\mathbf{g}^{l+1} = \mathbf{g}^l + \lambda^l \mathbf{d}^l$  en la derecha de la igualdad anterior, se obtiene  $v_a(\mathbf{g}^{l+1}) = v_a(\mathbf{g}^l) + \lambda^l v_a(\mathbf{d}^l)$  y, por lo tanto:

$$\frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} = \frac{\partial Z(\mathbf{g}^l)}{\partial g_{pq}} + \lambda^l \left[ d_{pq}^l + k \sum_a v_a(\mathbf{d}^l) \sum_{s \in S_{pq}} \delta_{as} \pi_s \right]. \quad (4.53)$$

Con esta relación se puede evaluar la diferencia de gradientes en el cálculo de (4.51) ó (4.52), así como el gradiente de  $Z$  en la nueva demanda  $\mathbf{g}^{l+1}$ . Además, el cálculo del denominador en (4.52) se puede simplificar, debido a que:

$$\begin{aligned} \sum_{pq \in PQ} d_{pq}^l \left( \frac{\partial Z(\mathbf{g}^{l+1})}{\partial g_{pq}} - \frac{\partial Z(\mathbf{g}^l)}{\partial g_{pq}} \right) &= \lambda^l \sum_{pq \in PQ} d_{pq}^l \left[ d_{pq}^l + k \sum_{a \in A} v_a(\mathbf{d}^l) \sum_{s \in S_{pq}} \delta_{as} \pi_s \right] \\ &= \lambda^l \left[ \sum_{pq} (d_{pq}^l)^2 + k \sum_{a \in A} v_a(\mathbf{d}^l) \sum_{pq} d_{pq}^l \sum_{s \in S_{pq}} \delta_{as} \pi_s \right] \\ &= \lambda^l \left[ \sum_{pq \in PQ} (d_{pq}^l)^2 + k \sum_{a \in A} v_a(\mathbf{d}^l)^2 \right]. \end{aligned} \quad (4.54)$$

Observése que la última expresión es igual al producto de  $\lambda^l$  por el denominador de la expresión en (4.49). Por supuesto que, la expresión (4.54), también puede interpretarse como el numerador en la expresión (4.49). Por lo tanto, al calcular  $\beta^l$  en (4.52) no es necesario realizar el cálculo del denominador de nuevo. Además, la diferencia de las derivadas parciales en (4.49) se puede calcular utilizando (4.53).

Con los elementos que se introdujeron en este capítulo y el desarrollo detallado de los cálculos, se tienen todos los ingredientes necesarios para aplicar el algoritmo de gradiente conjugado multiplicativo al problema de estimación de demanda, basado en el modelo penalizado (4.25). En el siguiente capítulo se comparan los resultados obtenidos con este nuevo método con aquéllos que se obtienen con el método de descenso máximo de Spiess, para el caso de la red de transporte de la ciudad de Winnipeg en Canadá.

# Capítulo 5

## Resultados

En este capítulo se muestran los resultados de aplicar los métodos de descenso estudiados anteriormente a la red de transporte de la ciudad de Winnipeg. Para saber qué tan bien funciona cada método, se calculará la raíz del error cuadrático medio (RMSE) y el coeficiente de correlación ( $R^2$ ), los cuales se explica brevemente como obtenerlos a continuación.

### 5.1. La raíz del error cuadrático medio y el coeficiente de correlación

La raíz cuadrada del error cuadrático medio (RMSE) se utiliza frecuentemente para medir la diferencia entre una serie de valores obtenidos con un modelo y los valores observados. El RMSE es una buena medida de la precisión, pero solamente para comparar los errores de predicción entre modelos diferentes para una variable particular y no entre las variables, ya que es dependiente de la escala [18].

Supóngase que se tiene una muestra de datos observados  $V = (V_1, V_2, \dots, V_n)$  y que mediante un modelo se predicen los datos  $v = (v_1, v_2, \dots, v_n)$ . El RMSE de los  $n$  valores previstos  $v$  se calcula como la raíz cuadrada de la media de los cuadrados de las desviaciones:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (v_i - V_i)^2}{n}} \quad (5.1)$$

Otra forma de comparar resultados es mediante el coeficiente de correlación, denotado por  $R^2$  y se pronuncia R-cuadrada, el cual indica qué tan bien se ajustan los puntos de datos a una línea o curva. En estadística, se interpreta como una medida de qué tan bien los resultados observados se replican por el modelo [8]. Hay varias definiciones diferentes de  $R^2$ , las cuales sólo a veces son equivalentes, una de ellas incluye la regresión lineal.

$$f(x) = A + Bx$$

donde los coeficientes  $A$  y  $B$  se obtienen minimizando la suma residual de cuadrados.

Considérense nuevamente los datos  $v$  y  $V$ , el coeficiente  $R^2$  se calcula de la siguiente manera:

$$S_x = \sum_{i=1}^n V_i, \quad S_y = \sum_{i=1}^n v_i, \quad S_{xx} = \sum_{i=1}^n (V_i)^2, \quad S_{yy} = \sum_{i=1}^n (v_i)^2, \quad S_{xy} = \sum_{i=1}^n V_i v_i,$$

$$SS_y = S_{yy} - \frac{S_y S_y}{n}, \quad q = nS_{xx} - S_x S_x, \quad B = \frac{nS_{xy} - S_x S_y}{q}, \quad A = \frac{S_y - bS_x}{n},$$

$$SS_{yx} = S_{yy} + nA^2 + B^2 S_{xx} - 2(AS_y + BS_{xy} - ABS_x),$$

finalmente

$$R^2 = 1 - \frac{SS_{yx}}{SS_y} \quad (5.2)$$

con la propiedad de que:

$$RMSE = \sqrt{\frac{SS_{yx}}{n-2}} \quad (5.3)$$

Usando las fórmulas 5.2 y 5.3, se compararán los datos obtenidos con los métodos de descenso vistos en el capítulo anterior.

## 5.2. Ejemplo de aplicación a la red de la ciudad de Winnipeg

Considérese la red de transporte de la ciudad de Winnipeg (figura 5.1), la cual cuenta con las siguientes características:

906	nodos regulares.	4	tipos de vehículos de tránsito.
154	nodos centroides.	133	líneas de tránsito.
3005	arcos direccionales.	4347	segmentos de líneas de tránsito.
5	modos de transporte.		

Considérese una matriz  $M_1$ , la cual representa la demanda matutina de pasajeros en la red y que, mediante una asignación de tránsito, se obtienen los volúmenes en algunos de los arcos y algunos segmentos de tránsito de la red. La matriz  $M_1$  y las posiciones de los arcos en donde se obtienen los volúmenes fueron proporcionados por INRO [20]. También considérese la suma de orígenes  $O_p$  y la suma de destinos  $D_q$ . Dependiendo de la clase de datos que se conozcan de la red, se elegirá el método apropiado para resolver el problema de demanda. Es decir, si se tienen conteos del total de viajes que se originan en el nodo  $p$  y el total viajes que tienen como destino el nodo  $q$ , entonces se optará por utilizar el método de balanceo biproporcional y en caso de contar además con los costos de viaje, se optará por utilizar el método de balanceo triproporcional. Por otra parte, si los datos que se conocen son los conteos de volúmenes en arcos o segmentos de la red, entonces se utilizará alguno de los métodos de descenso.



Figura 5.1: Red de transporte público de la ciudad de Winnipeg, Manitoba, [20].

### 5.2.1. Resultados con el método de balanceo biproporcional

Supóngase que la matriz  $M_1$  es desconocida, y que en su lugar se tiene una matriz  $M_2$  conocida a priori, la cual se obtuvo de perturbar estocásticamente la matriz  $M_1$  para obtener entre el 70 y 100 % del valor inicial en cada una de sus entradas. Supóngase que se conocen también el total de viajes que se originan en el nodo  $p$  y el total de viajes que tienen como destino el nodo  $q$ ,  $O_p$  y  $D_q$ , estos totales corresponden a la suma de renglones y columnas de la matriz  $M_1$ , respectivamente. Por lo anterior es claro que la matriz  $M_2$  no está balanceada respecto a los  $O_p$  y  $D_q$  conocidos, es por esto que se le aplicará el método biproporcional de matrices para balancearla, esperando recuperar la matriz  $M_1$ . La dispersión de demanda inicial es la siguiente:

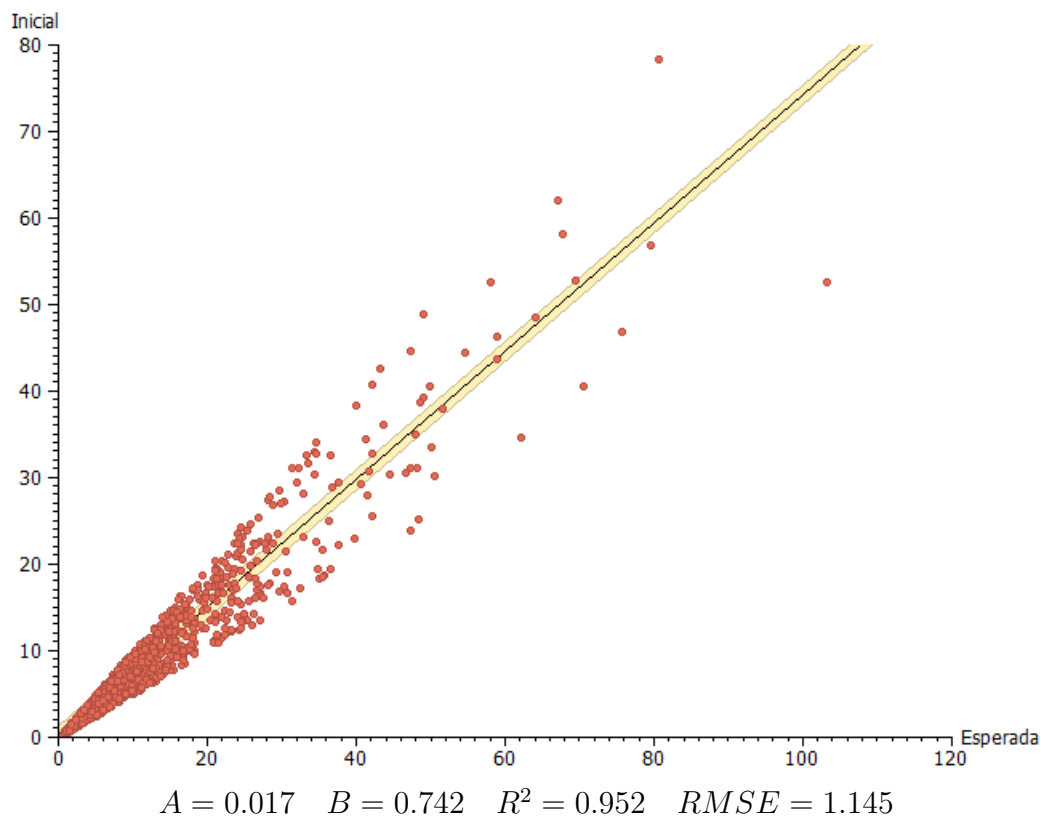


Figura 5.2: Demanda esperada vs demanda inicial.

En donde cada punto rojo de la gráfica de la figura 5.2, representa la demanda en un nodo de la red y tiene como coordenadas  $(M_{1\ pq}, M_{2\ pq})$ . Los coeficientes  $A$  y  $B$  representan, respectivamente la ordenada al origen y la pendiente de la recta de ajuste. El factor de correlación se representa con  $R^2$  y la raíz del error cuadrático medio con  $RMSE$ .

Al aplicar el algoritmo de balanceo biproporcional (ver apéndice A.2), después de 5 iteraciones se obtiene la siguiente gráfica de dispersión para la matriz balanceada:

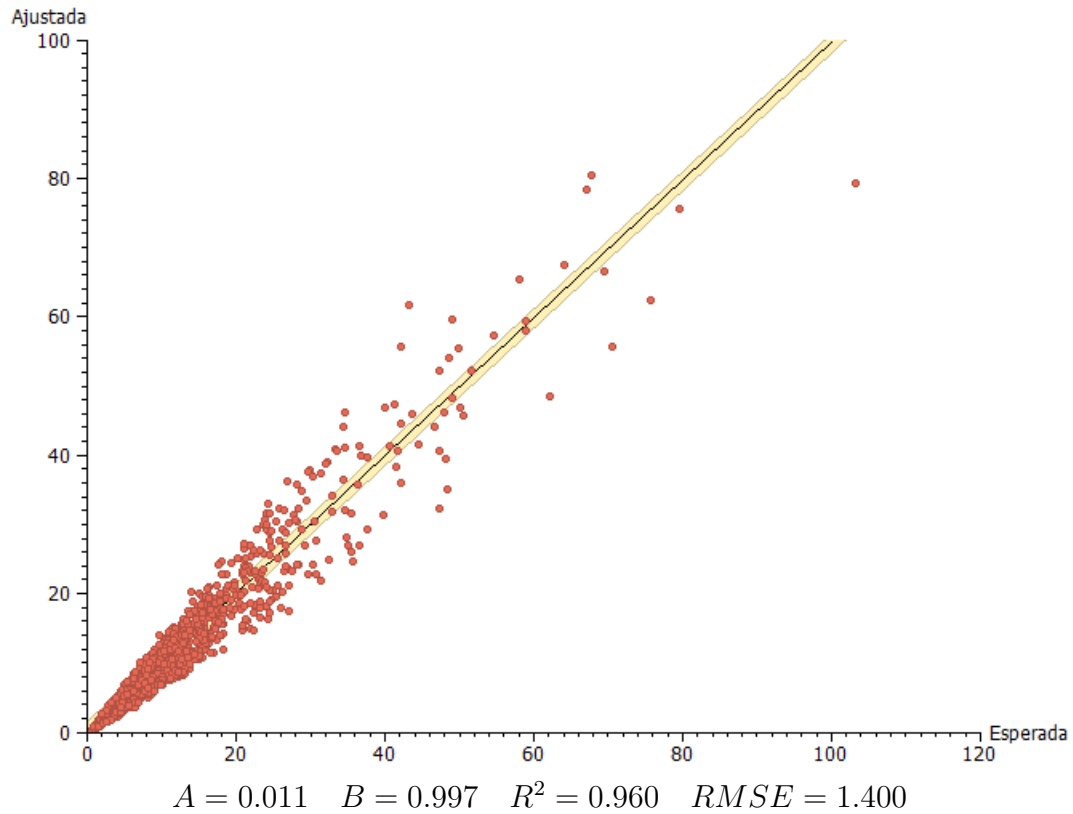


Figura 5.3: Demanda esperada vs demanda balanceada, 5 iteraciones.

Como se puede observar en la figura 5.3, la bondad del ajuste es de un 96%. Esto para el caso en el que se conocen las sumas sobre los orígenes y los destinos, si se considerasen como datos conocidos los volúmenes de arco o de segmento, entonces convendría aplicar un método de descenso, como se verá en la siguiente sección.

### 5.2.2. Resultados con los métodos de descenso para ajuste de matrices

Considérese nuevamente la matriz  $M_2$  obtenida de perturbar  $M_1$ . También considérense los volúmenes de arco como observaciones. En este caso, se cuenta con el volumen en 112 arcos de la red, marcados en color naranja en la figura 5.4.



Figura 5.4: Conteos en 112 arcos de la red.



La dispersión inicial de volúmenes es la siguiente:

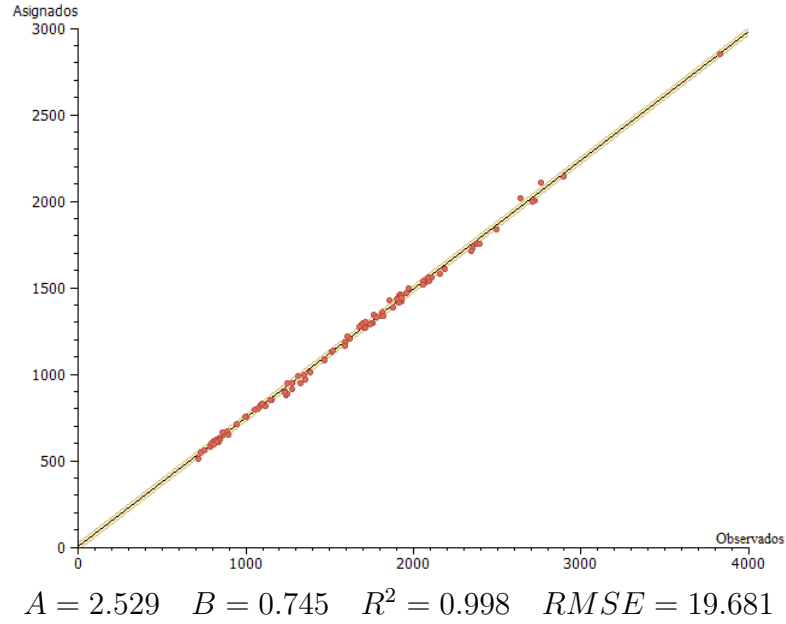


Figura 5.5: Dispersión inicial de los volúmenes de arco.

En este caso, cada punto rojo de la gráfica de la figura 5.5, representa el volumen sobre un arco de la red y tiene como coordenadas  $(V_a, v_a)$ . Los coeficientes  $A$  y  $B$  representan, respectivamente la ordenada al origen y la pendiente de la recta de ajuste. El factor de correlación se representa con  $R^2$  y la raíz del error cuadrático medio con  $RMSE$ .

Para ajustar  $M_2$  se utiliza el modelo de mínimos cuadrados de la ecuación (4.1), considerando  $\alpha = 0.5$  y se utilizan los métodos de máximo descenso y gradiente conjugado multiplicativos para el ajuste. Los resultados se muestran a continuación, tomando una tolerancia de  $10^{-6}$  para las iteraciones:

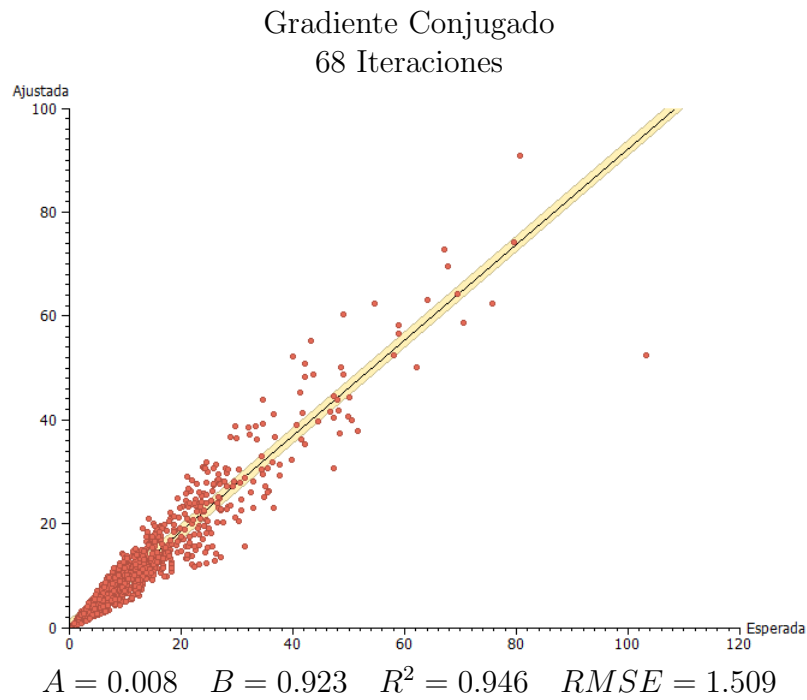
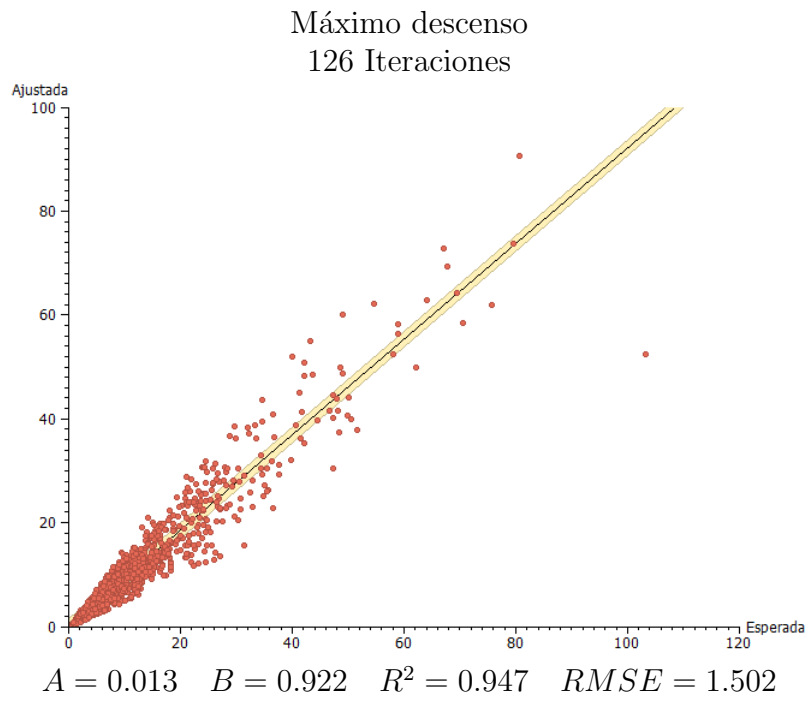


Figura 5.6: Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los arcos y  $\alpha = 0.5$ .

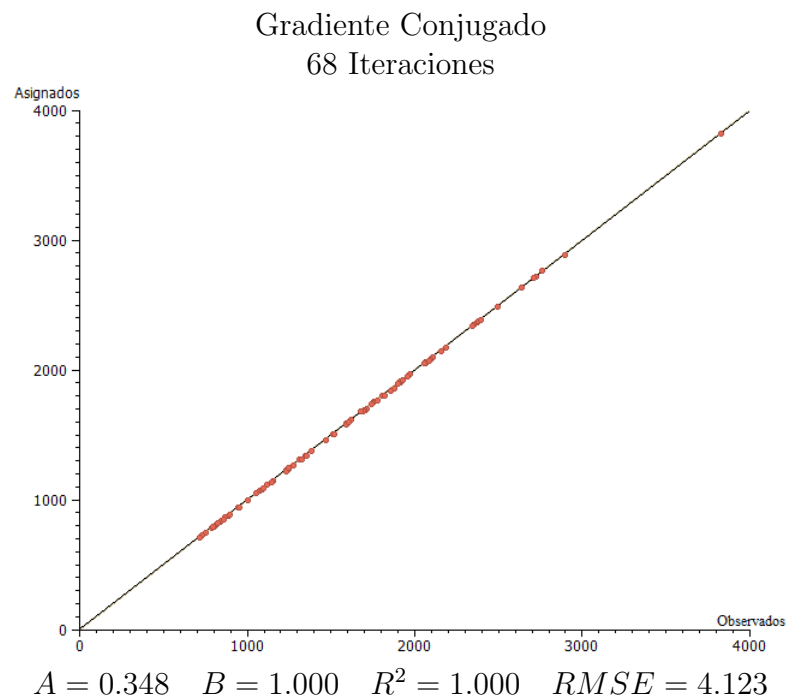
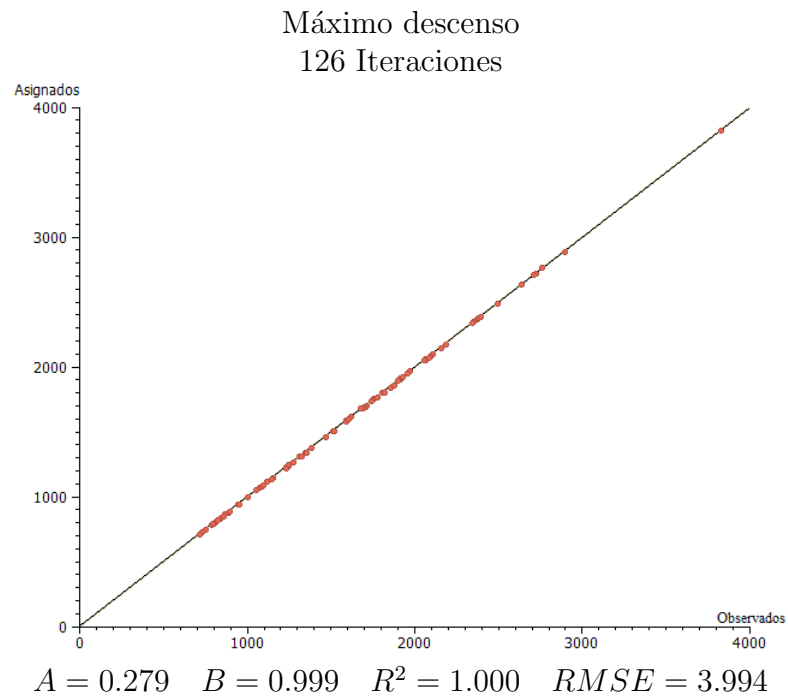


Figura 5.7: Volúmenes observados vs volúmenes asignados, para máximo descenso y gradiente conjugado considerando conteos en los arcos y  $\alpha = 0.5$ .

Como se ve en las figuras 5.6 y 5.7, máximo descenso requiere 58 iteraciones más que gradiente conjugado. La dispersión de volúmenes queda muy bien ajustada para ambos métodos y la dispersión de demanda tiene una bondad de ajuste del 94 %.

En la práctica, obtener los volúmenes de arco resulta un tanto difícil, ya que al tratarse de transporte público, es necesario contar a todos los pasajeros que pasan por cada arco, en cada una de las líneas de transporte. Es por esto que se sugiere utilizar conteos de segmentos, ya que de esta manera, es más fácil tomar alguno de los vehículos de transporte e ir contando el número de pasajeros en cada uno de sus segmentos. Para ejemplificar esto, considérense los conteos en 136 segmentos de tránsito mostrados en la siguiente figura:



Figura 5.8: Conteos en 136 segmentos de la red.

Al aplicar los métodos de descenso se obtiene:

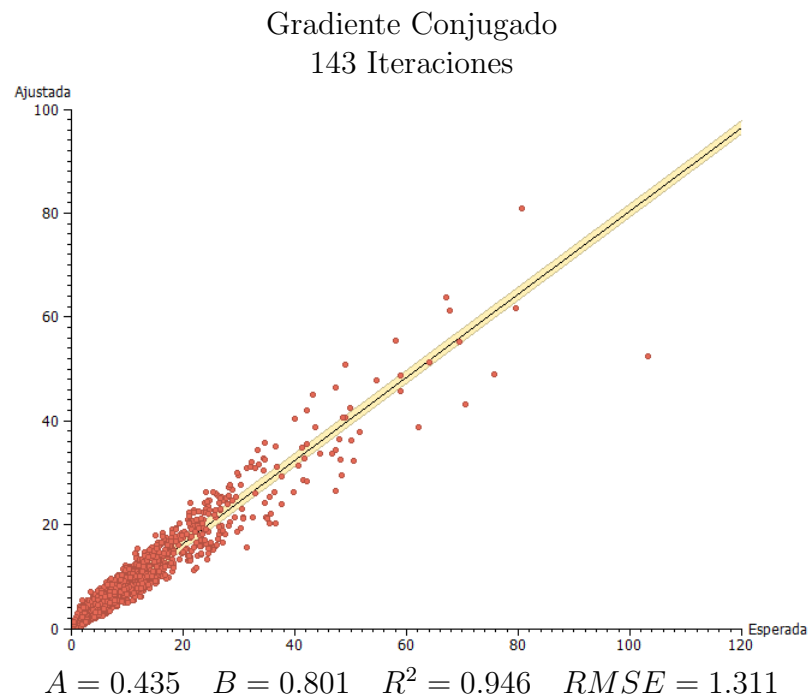
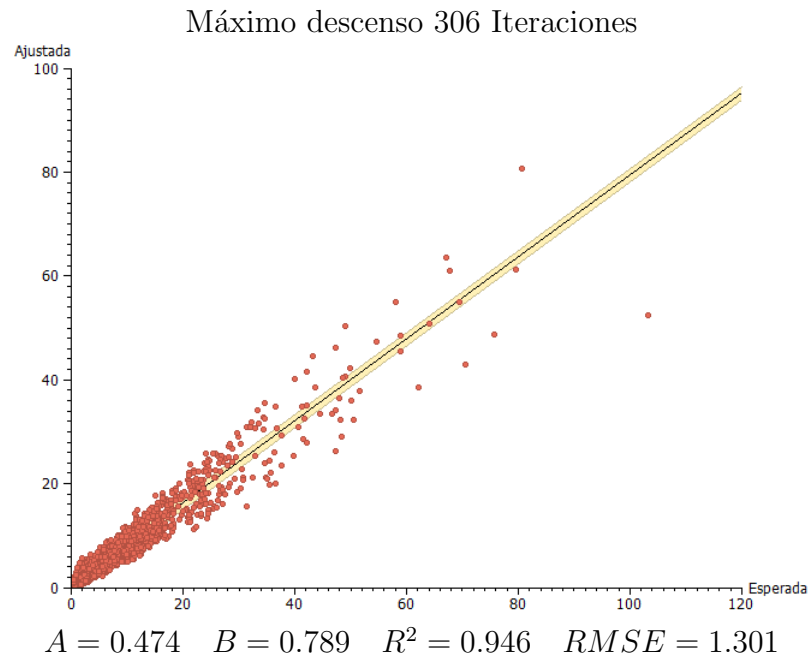


Figura 5.9: Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y  $\alpha = 0.5$ .

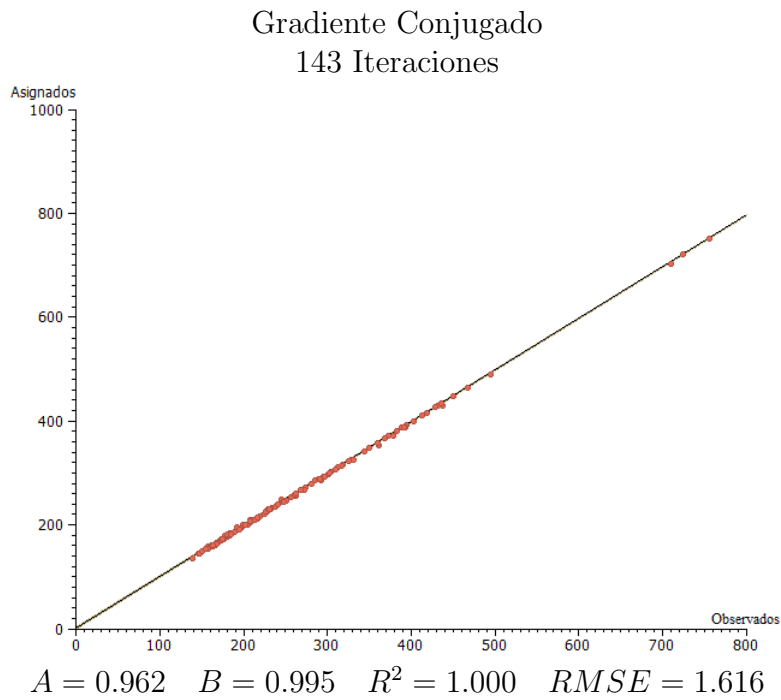
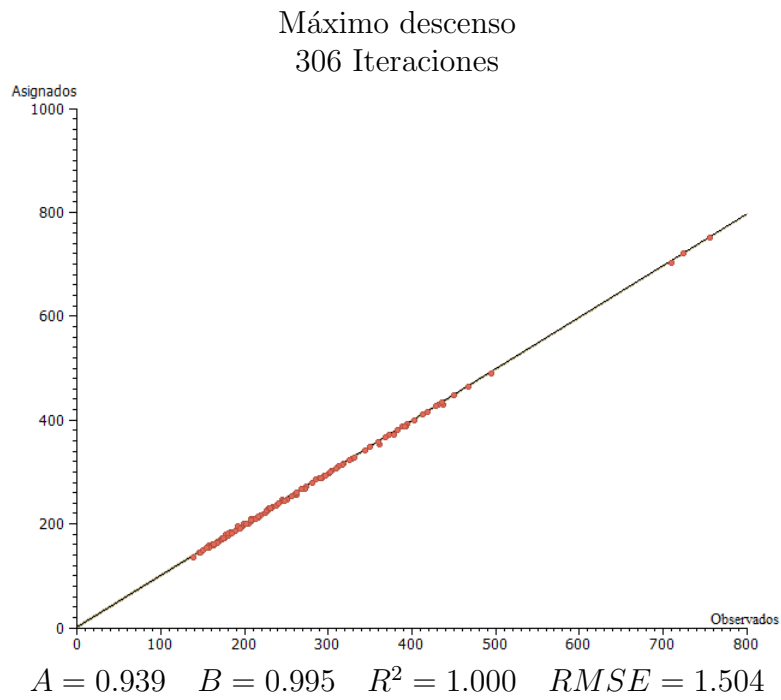


Figura 5.10: Volúmenes asignados vs volúmenes observados, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y  $\alpha = 0.5$ .

Como se puede observar en las figuras 5.9 y 5.10, máximo descenso requirió 163 iteraciones más que gradiente conjugado.

Aunque se requiere un número mayor de iteraciones para hacer un ajuste de demanda con conteos de segmentos que con conteos de arcos, se tendría que considerar la facilidad de obtener cada uno de los conteos. Como se mencionó anteriormente, en la práctica es más fácil la obtención de conteos en segmentos, es por esto que es importante encontrar una manera de reducir el número de iteraciones cuando se tienen dicha clase de datos. A continuación se considerará el modelo de mínimos cuadrados que incluye el factor de penalización  $k$  mostrado en la ecuación (4.3). Los resultados obtenidos para  $k = 100$  son los siguientes:

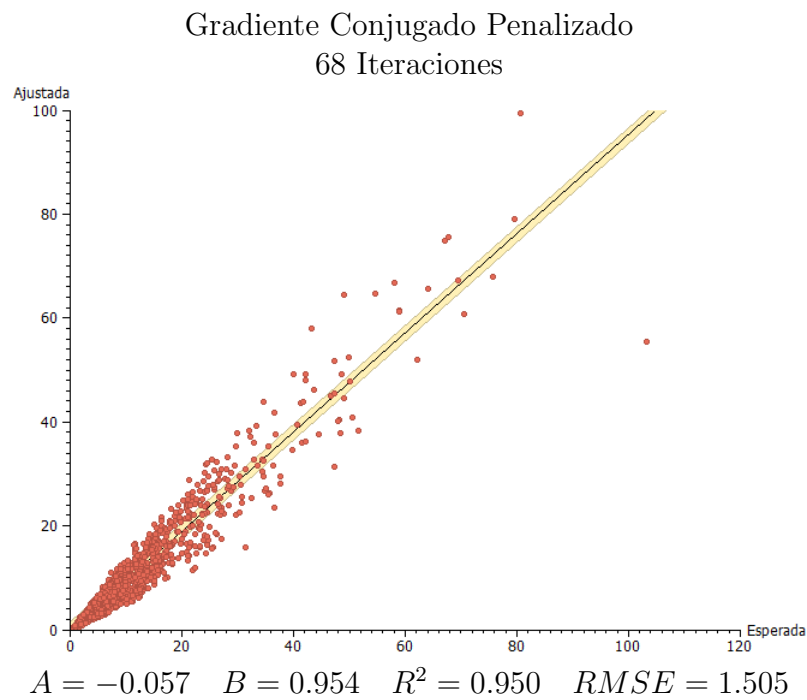
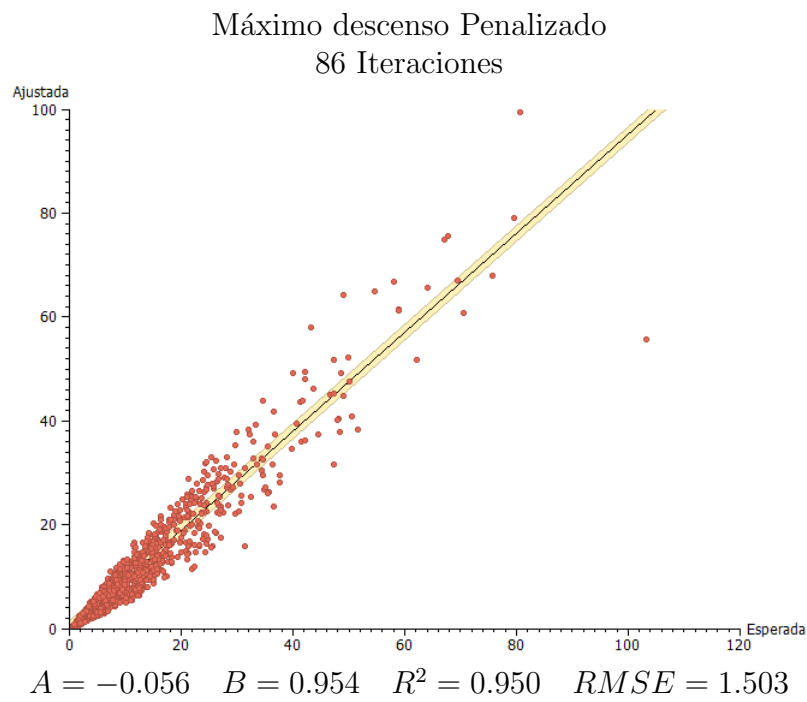


Figura 5.11: Demanda esperada vs demanda ajustada, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y un factor de penalización de  $k = 100$ .



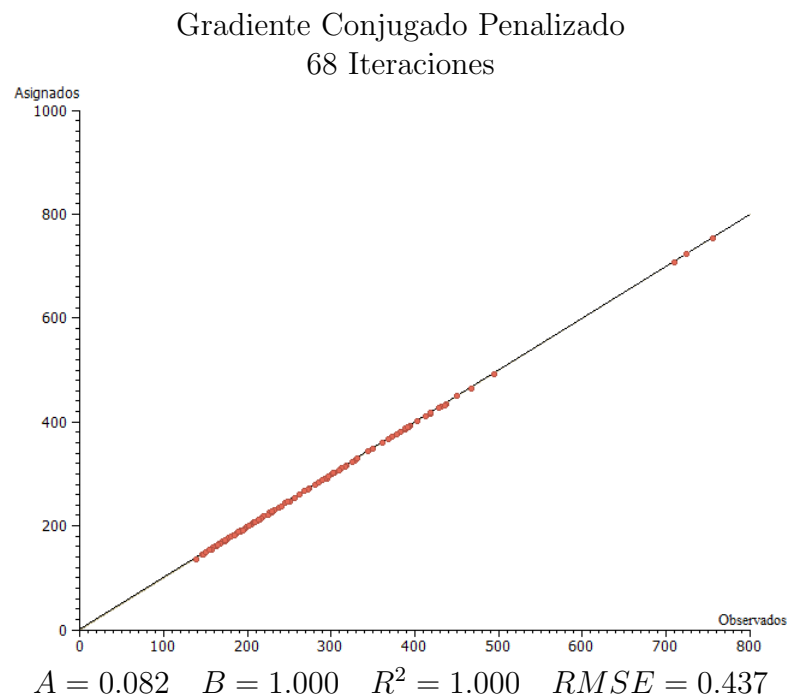
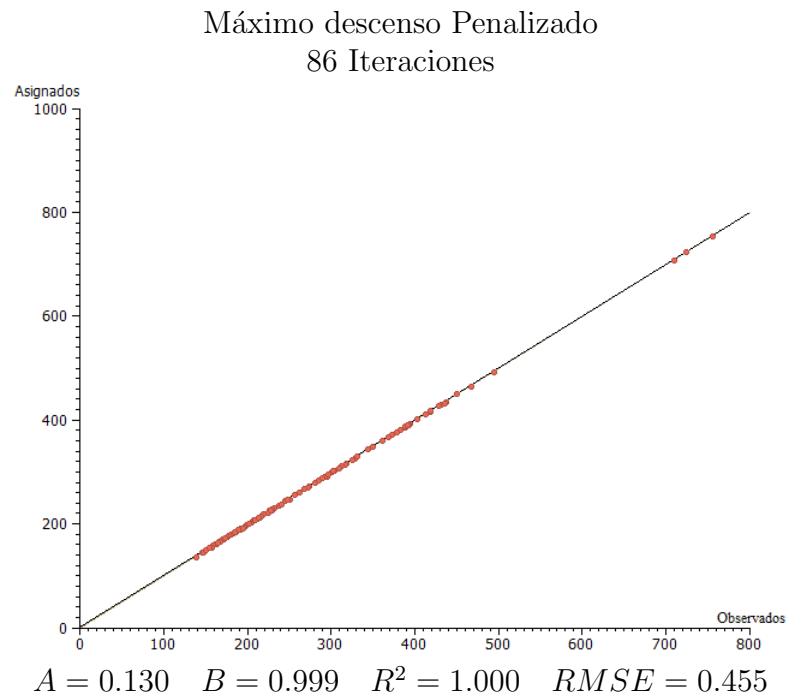


Figura 5.12: Volúmenes observados vs volúmenes asignados, para máximo descenso y gradiente conjugado considerando conteos en los segmentos y un factor de penalización de  $k = 100$ .

En las figuras 5.11 y 5.12 se ve cómo al poner un factor de penalización igual a 100, se reduce el número de iteraciones en ambos métodos; además, se observa que el método de máximo descenso sigue requiriendo de un mayor número de iteraciones de las que requiere el método de gradiente conjugado. Ahora considérese un factor  $k = 1000$ , los resultados obtenidos para gradiente conjugado son los siguientes:

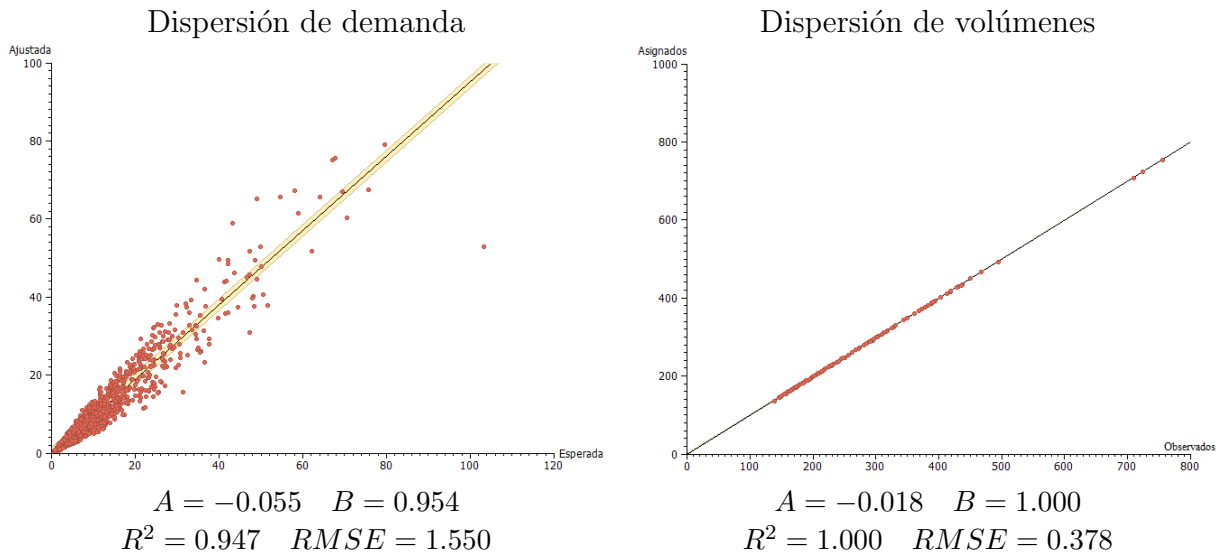


Figura 5.13: Dispersión de demanda y de volúmenes para gradiente conjugado considerando conteos en los segmentos y un factor de penalización de  $k = 1000$  con 54 iteraciones.

En la figura 5.13, se puede ver cómo nuevamente al incrementar el valor del factor de penalización  $k$ , se continúa reduciendo el número de iteraciones. Finalmente, considérese un factor de penalización igual a 10000, los resultados son:

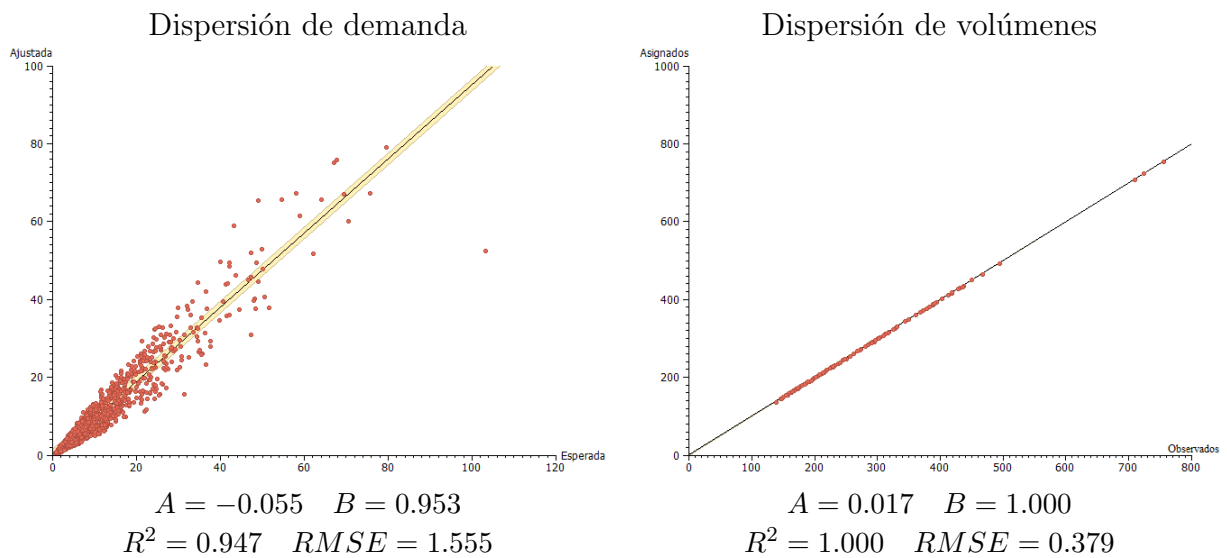


Figura 5.14: Dispersión de demanda y de volúmenes para gradiente conjugado considerando conteos en los segmentos y un factor de penalización de  $k = 10000$  con 53 iteraciones.

En la figura 5.14 se observa que se redujo únicamente una iteración, lo cual nos indica que se está llegando al límite del factor de penalización. En el capítulo siguiente se muestran las conclusiones obtenidas en base a los resultados mostrados en este capítulo.



# Capítulo 6

## Conclusiones y trabajo a futuro

En el presente trabajo se han estudiado varios modelos para estimar matrices de demanda de transporte urbano. Uno de los modelos más simples y usados es el de balance biproporcional, en el que se conoce una matriz de demanda a priori y los totales marginales de la matriz de demanda que se espera obtener. Como se vió en el capítulo anterior, este método converge en pocas iteraciones para redes pequeñas. Podría resultar relativamente fácil obtener el número total de viajes que se originan o que terminan en cierta área, depende de la infraestructura de la red y de los vehículos de transporte, no se debe de perder de vista que los resultados obtenidos serán un indicador de la demanda y no la demanda como tal, es por esto que se ha investigado cada vez más sobre métodos alternativos que proporcionen información más detallada de cada uno de los viajes, como aquellos en los que se consideran conteos en los arcos.

El tema central de este trabajo ha sido el estudio de los modelos de demanda que se basan en el conteo de volúmenes sobre un conjunto predeterminado de arcos de la red. En el modelo simplificado de Spiess, (4.5), se plantea la solución de un problema de mínimos cuadrados, en donde se minimiza la diferencia entre los volúmenes medidos y aquellos que se obtienen por medio de una asignación de tránsito con la demanda desconocida. En modelos un poco más sofisticados se agrega la diferencia entre la matriz origen-destino que se quiere estimar y la conocida a priori, como en (4.1). Debido a que el aspecto central es la minimización de la distancia entre la matriz conocida a priori y la matriz desconocida, se introduce el modelo (4.3), en donde la diferencia de volúmenes medidos y los reales se incorpora en la función objetivo con un parámetro de penalización (4.3), en lugar de la ponderación promediada de ambos términos como en (4.1). Esta es una de las dos principales aportaciones del presente trabajo de tesis.

Para resolver los problemas que surgen de los modelos mencionados anteriormente se introduce un algoritmo de gradiente conjugado multiplicativo, el cual, hasta donde se sabe, es novedoso en el ámbito de la ingeniería del transporte. Esta es la otra aportación importante de esta tesis. Los resultados se comparan con aquellos obtenidos cuando se aplica el método de descenso máximo, obteniendo soluciones muy similares en ambos casos, pero con

la ventaja de que gradiente conjugado realiza menos iteraciones. Además, el efecto de penalizar la diferencia de volúmenes reduce aún más el número de iteraciones. En resumen, ambos, el modelo penalizado y el método de gradiente conjugado multiplicativo, generan un procedimiento más eficiente para estimar la demanda desconocida a partir de la medición de volúmenes sobre un conjunto determinado de arcos ó segmentos de la red de transporte.

Por lo que respecta a los resultados obtenidos en la red de Winnipeg, es importante resaltar que en los experimentos es posible considerar mediciones de volúmenes en los arcos ó en los segmentos. Los resultados numéricos mostrados en este trabajo, indican que el número de iteraciones de los métodos para aproximarse al óptimo, es mayor cuando se utilizan volúmenes de segmentos que cuando se utilizan volúmenes de arcos. Una posible explicación es la siguiente: el número de restricciones (igualdad de volúmenes) en el problema de optimización es mayor cuando se miden volúmenes en los segmentos que cuando se miden en los arcos (recuérdese que por cada arco de la red hay varios segmentos), por lo que el tamaño del problema penalizado (4.25) aumenta al aumentar el número de sumandos. Por otro lado, cuando se miden los volúmenes sobre segmentos se recupera mejor la demanda, como es de esperarse, debido a que la información utilizada es más selectiva, al separar la información sobre los arcos en la información más específica de los volúmenes sobre las diferentes líneas de transporte. Por lo tanto, dependiendo de las necesidades, y del nivel de aproximación que se requiera, es posible elegir arcos ó segmentos, o combinaciones de arcos y segmentos.

Hay varios aspectos que pueden ser estudiados en un trabajo futuro y que no se han abordado en este trabajo. Algunos de ellos son los siguientes:

1. Para acelerar la convergencia del método de gradiente conjugado multiplicativo se puede investigar la posibilidad de introducir un preconditionador ad-hoc para este tipo de problemas. Debido a que un preconditionador estará asociado a un cierto tipo de inversa del problema lineal asociado a las condiciones necesarias para el óptimo, es posible que estudiando el problema dual, asociado al problema de optimización con restricciones, se obtenga cierta idea para obtenerlo. La importancia de un preconditionador será mayor cuando se trate de estimar demanda en redes de gran tamaño, como la del Valle de México, ya que en esos casos el número de iteraciones sin duda aumentará, además de que cada iteración será de mayor costo computacional.
2. Al usar el método de gradiente conjugado y observar mejoras, es natural pensar en otras técnicas de optimización basadas en sub-espacios de Krylov, las cuales en la actualidad se encuentran dentro de los métodos más eficaces del álgebra lineal numérica para resolver problemas de gran escala. Por supuesto que el estudio del método de gradiente conjugado multiplicativo introducido en este trabajo no está agotado; por ejemplo no se ha abordado el problema de estudiar otras variantes más económicas para estimar el parámetro  $\beta$ .
3. Es posible también extender estos modelos incorporando mayor información, como por ejemplo la matriz de varianza-covarianza como en el modelo (3.10). Otra posibilidad

es incorporar otro tipo de restricciones como las tarifas de viaje ó también incluir cotas superiores en la demanda, Romero sugiere en [25] considerar también los totales marginales  $O_p$  y  $D_q$  como restricciones en caso de que se cuente con esos datos. En resumen, se trata de generar nuevo modelos basados en métodos de conteo y formulaciones de proyección (mínimos cuadrados).





# Apéndice A

## Algoritmos

### A.1. Algoritmo de asignación para una red simple

Etapa 1. Cálculo de la estrategia óptima.

1. Inicialización.

$$u_i = \infty, \quad i \in N - D \quad u_D := 0;$$

$$f_i := 0, \quad i \in N;$$

$$S = A; \quad \bar{A} = \emptyset.$$

2. Iteraciones. Obtención del siguiente arco.

Si  $S = \emptyset$ , parar e ir al paso 4.

En caso contrario, encontrar la arista  $a = (i, j) \in S$  que satisfaga:

$$u_j + t_a \leq u_{j'} + t_{a'}, \quad a' = (i', j')$$

Hacer  $S := S - \{a\}$

3. Actualizar etiquetas de nodo.

Si  $u_i > u_j + t_a$ , entonces:

$$u_i := \frac{f_i u_i + f_a (u_j + t_a)}{f_i + f_a}$$

$$f_i := f_i + f_a, \quad \bar{A} = \bar{A} + \{a\};$$

Ir al paso 2.

4. Terminar.

Etapa 2. Asignación de la demanda.

1. Inicialización.

$$v_p = g_{pq}, pq \in PQ;$$

2. Iteraciones. Asignación de viajes. Para cada arco  $a \in A$  hacer en orden decreciente de  $(u_j + t_a)$ :

$$\text{Si } a \in \bar{A}, \text{ entonces } v_a := \frac{f_a}{f_i} v_i, \quad v_j = v_j + v_a, \text{ en caso contrario } v_a := 0.$$

Si  $A := \emptyset$  ir al paso 3.

3. Terminar.

## A.2. Método biproporcional estándar

1. Inicialización.

Asignar  $B_q^0 = 1$  para  $q \in Q$  y  $k = 1$ .

2. Balance de orígenes.

Para cada origen  $p \in P$  calcular

$$a_p^k = \frac{O_p}{\sum_{q \in Q} b_q^{k-1} G_{pq}}$$

3. Balance de destinos.

Para cada destino  $q \in Q$  calcular

$$b_q^k = \frac{D_q}{\sum_{p \in P} a_p^k G_{pq}}$$

4. Criterio de paro.

Si

$$\|a^k - a^{k-1}\| + \|b^k - b^{k-1}\| < \varepsilon_1$$

ir al paso 5, de lo contrario hacer  $k = k + 1$  y regresar al paso 2.

5. Calcular la solución del problema primal.

$$g_{pq} = a_p^k b_q^k G_{pq}$$

parar.

## A.3. Método biproporcional con cotas superiores

Algoritmo 1.

1. Inicialización.

Asignar  $B_q^0 = 1$  para  $q \in Q$  y  $k = 1$ .

2. Balance de orígenes.

Para cada origen  $p \in P$  resolver la ecuación

$$\sum_{q \in Q} \min\{a_p^k b_q^{k-1} G_{pq}, U_{pq}\} = O_p$$

para la variable  $a_p^k$  aplicando el Algoritmo 2.

3. Balance de destinos.

Para cada destino  $q \in Q$  resolver la ecuación

$$\sum_{p \in P} \min\{a_p^k b_q^k G_{pq}, U_{pq}\} = D_q$$

para la variable  $b_q^k$  aplicando el Algoritmo 2.

4. Criterio de paro.

$$\|a^k - a^{k-1}\| + \|b^k - b^{k-1}\| < \varepsilon_1$$

ir al paso 5, de lo contrario hacer  $k = k + 1$  y regresar al paso 2.

5. Calcular la solución del problema primal.

$$\left\| \sum_{q \in Q} \min\{a_p^k b_q^k G_{pq}, U_{pq}\} - O_p \right\| < \varepsilon_2$$

entonces

$$g_{pq} = \min\{a_p^k b_q^k G_{pq}, U_{pq}\}$$

de otra forma el problema primal no es factible, parar.

Algoritmo 2.

1. Inicialización.

Asignar  $F = \sum_i f_i$  y  $U = T$

2. Ordenar los elementos.

Ordenar los elementos  $i$  en orden creciente de los cocientes  $u_i/f_i$ .

3. Leer los elementos.

Usando el orden establecido en el paso 2, hacer para cada  $i$  Si  $u_i F \leq f_i U$  entonces,  $F := F - f_i$  y  $U := U - u_i$ ; de otra forma, ir al paso 4.

Si esta condición no se cumple después de haber leído todos los elementos, entonces el problema no es factible, parar.

4. Calcular la solución óptima.

Asignar la solución  $x = U/F$  y parar.

## A.4. Método triproporcional

1. Inicialización. Para cada  $pq \in PQ$  hacer:

$$g_{pq}^0 = G_{pq}$$

2. Balance de orígenes. Para cada  $pq \in PQ$  calcular:

$$g_{pq}^{k+1} = \frac{O_p}{\sum_{q \in Q} g_{pq}^k} g_{pq}^k$$

$$k = k + 1$$

3. Balance de destinos. Para cada  $pq \in PQ$  calcular:

$$g_{pq}^{k+1} = \frac{D_q}{\sum_{p \in P} g_{pq}^k} g_{pq}^k$$

$$k = k + 1$$

4. Balance de costos. Sea  $I_{pq}$  el intervalo de costos correspondiente al viaje de  $p$  a  $q$ . Entonces: Para cada  $pq \in PQ$  calcular

$$g_{pq}^{k+1} = \frac{R_{I_{pq}}}{\sum_{pq \in PQ} g_{pq}^k \delta_{pq}^{I_{pq}}} g_{pq}^k$$

$$k = k + 1$$

5. Criterio de paro. Si se cumplen todas las restricciones con un cierto grado de tolerancia, entonces parar, de lo contrario ir al paso 2.

## A.5. Método de gradiente conjugado

1. Inicialización. Hacer el contador de iteraciones  $l = 0$  y  $\mathbf{g}^0 = \mathbf{G}$ .

2. Para  $l = 0, \dots, L$ , hacer:

2.1. Asignación para calcular la función objetivo. Hacer una asignación de tránsito con  $\mathbf{g}^l$  para obtener los volúmenes de arco  $v_a^l, \forall a \in A$ .

2.2. Calcular el valor de la función objetivo.

$$Z(\mathbf{g}) = \frac{1}{2} \sum_{a \in \bar{A}} (v_a^l(\mathbf{g}) - V_a)^2 + \frac{k}{2} \sum_{pq \in PQ} (g_{pq}^l - G_{pq})^2$$

Si se alcanzó el número máximo de iteraciones  $L$ , ir al paso 3.

- 2.3. Asignación para calcular el gradiente. Hacer una asignación de tránsito agregando el atributo que contiene los conteos en los segmentos y guardar la información en la matriz del gradiente  $\nabla Z(\mathbf{g}^l)$ . Posteriormente agregar el término correspondiente a la demanda.

$$\nabla Z(\mathbf{g}^l) = \sum_{s \in S_{pq}^l} \pi_s^l \sum_{a \in \bar{A}} \delta_{as}^l (v_a^l - V_a) + k (g_{pq}^l - G_{pq})$$

- 2.4. Calcular la dirección de descenso.

Si  $l = 0$ , tomar como dirección de descenso la del gradiente, de lo contrario:

Calcular  $\beta$ :

$$\beta_{l+1} = -\frac{(\nabla Z(\mathbf{g}^{l+1}))^T (\nabla Z(\mathbf{g}^{l+1}) - \nabla Z(\mathbf{g}^l))}{(\mathbf{d}^l)^T (\nabla Z(\mathbf{g}^{l+1}) - \nabla Z(\mathbf{g}^l))}$$

Calcular la nueva dirección de descenso:

$$\mathbf{d}^{l+1} = -\nabla Z(\mathbf{g}^{l+1}) + \beta \mathbf{d}^l$$

- 2.5. Calcular el tamaño óptimo del paso. Hacer una asignación de tránsito utilizando la dirección de descenso anterior para obtener la derivada de los volúmenes  $v_a$  respecto a  $\lambda$ , posteriormente calcular el tamaño de paso.

$$\lambda^l = \frac{\sum_{pq \in PQ} d_{pq}^l (G_{pq} - g_{pq}^l) + k \sum_{a \in A} v_a(\mathbf{d}^l) (V_a - v_a(\mathbf{g}^l))}{\sum_{pq \in PQ} (d_{pq}^l)^2 + k \sum_{a \in A} v_a(\mathbf{d}^l)^2}$$

- 2.6. Actualizar la matriz de demanda.

$$g_{pq}^{l+1} = g_{pq}^l + \min(\lambda^l, 1) \mathbf{d}^l$$

- 2.7. Actualizar el contador de iteraciones  $l = l + 1$  e ir al paso 2.1.

3. Terminar.



# Bibliografía

- [1] Bacharach M. *Estimating Nonnegative Matrices from Marginal Data*, International Economic Review, Vol. 6, No. 3, pp. 294–310, 1965.
- [2] Beckmann M. J., McGuire C. B. and Winsten C. B. *Studies in the Economics of Transportation*, Yale University Press, New Haven, Conn, 1956.
- [3] Bierlaire M. *Mathematical Models for Transportation Demand Analysis*, Ph.D. Thesis, Facultés Universitaires Notre-Dame de la Paix de Namur, Faculte des Sciences, Département de Mathématique, 1995.
- [4] Casey H. J. *Applications to Traffic Engineering of the Law of Retail Gravitation*, *TraficQuarterly* IX(1): 23–35, 1955.
- [5] Cepeda M., Cominetti R., Florian M. *A Frequency-Based Assignment Model for Congested Transit Networks with Strict Capacity Constraints: Characterization and Computation of Equilibria*, Transportation Research Part B 40, 437–459, 2006.
- [6] Cominetti R., Correa J. *Common-Lines and Passenger Assignment in Congested Transit Networks*, Transportation Science 35 (3), 250–267, 2001.
- [7] Deming W. E., Stephan F. F. *On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known*, Annals of Mathematical Statistics, XI, 4427–444, 1940.
- [8] Draper N. R., Smith H. *Applied Regression Analysis*, John Wiley and Sons, 3a. Ed., 1998.
- [9] Feller W. *Introducción a la Teoría de Probabilidades y sus Aplicaciones*, Limusa, Vol. I, 1975.
- [10] Fernández A. G. *Modelos Matemáticos de Asignación de Tránsito. Aplicación a la Red Metropolitana del Valle de México y sus Efectos en el STC-Metro*, Tesis de Maestría en Ciencias, Departamento de Matemáticas Aplicadas e Industriales, Universidad Autónoma Metropolitana-Iztapalapa, 2013.

- [11] Florian M. *An Improved Linear Approximation Algorithm for the Network Equilibrium (Packet Switching) Problem*. IEEE Proc. Decision and Control, 812–828, 1977.
- [12] Florian M., Guélat J. and Spiess H. *An Efficient Implementation of the PARTAN Variant of the Linear Approximation for the Network Equilibrium Problem*. Networks 17, 319–339, 1987.
- [13] Florian M. *Models and Software for Urban and Regional Transportation Planning: The Contributions of the Center for Reserch on Transportation.*, Interuniversity Research Centre on Enterprise Networks, Logistic and Transportation, No. 11, pp 3–5, 2008.
- [14] Franck M. and Wolfe P. *An Algorithm for Quadratic Programming*. Naval Res. Logist. Quart. 3, 95–110, 1956.
- [15] Fu Q., Liu R. and Hess S. *A Review on Transit Assignment Modelling Approaches to Congested Networks: a New Perspective*, Procedia–Social and Behavioral Sciences, 54[4], pp. 1145–1155, 2012.
- [16] Furness K.P. *Time Function Interaction*, Traffic Engineering and Control, Vol. 7, No. 7, pp 19–36, 1970.
- [17] Hestenes M. R., Stiefel E. *Methods of Conjugate Gradients for Solving Linear Systems*, Journal of Research of the National Bureau of Standards, Vol. 49, No. 6, Artículo de investigación 2379, 1952.
- [18] Hyndmana R. J., Koehlerb A. B. *Another look at measures of forecast accuracy*, International Journal of Forecasting, Vol. 22, No. 4, pp 679–688, 2006.
- [19] <http://www.inegi.org.mx>
- [20] <http://www.inrosoftware.com/en/products/emme/index.php>
- [21] Lamond B. and Stewart N. F. *Bregman's Balancing Method.*, Transportation Research, Vol 15B, pp 239–248, 1981.
- [22] Luenberger D. G. *Linear and Nonlinear Programming*, Second Edition, Addison-Wesley, 1984.
- [23] Nocedal J. and Wright S. J. *Numerical Optimization*, Second ed., Springer, 2006.
- [24] Noriega Y. and Florian M. *Some Enhancements of the Gradient Method for O–D Matrix Adjustment*, Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, University of Montreal, Canadá. No. 4, pp 2–6, 2009.



- [25] Romero D. *Easy Transportation-Like Problems of K-Dimensional Arrays*, Journal of Optimization Theory and Applications, Vol. 66, No. 1, 1990.
- [26] Ross S. M. *Introduction to Probability Models*, ELSEVIER, 9th. Ed., 2007.
- [27] Spiess H. *Contributions a la Théorie et Aux Outils de Planification des Réseaux de Transport Urbain.*, Ph.D. thesis. Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Québec, 1984.
- [28] Spiess H. *Technical Note, Conical Volume-Delay Functions*, Transportation Science, 24[2], pp. 153–158, 1990.
- [29] Spiess H. *A Gradient Approach for the O-D Matrix Adjustment Problem*, EMME/2 Support Center, Haldenstrasse 16, CH-2558 Aegerten, Switzerland, 1990.
- [30] Spiess H. and Florian M. *Optimal Strategies: A New Assignment Model for Transit Networks*, Transportation Research—B, vol 23B, No. 2, pp. 83–102, 1989.
- [31] Stone R., Brown J. A. C. *A Computable Model of Economic Growth, (a Programme for Growth 1)*, London: Chapman and Hall, 1962.
- [32] *Top Ten Algorithms in the 20th Century*: January/February 2000 Issue of Computing in Science & Engineering (a Joint Publication of the American Institute of Physics and the IEEE Computer Society), 2000.
- [33] Van Zuylen H. J. and Willumsen L. G. *The Most Likely Trip Matrix Estimated from Traffic Counts.*, Transpn. Res. B 14, 281–293, 1980.
- [34] Wardrop, J. G. *Some Theoretical of Road Traffic Research*, Proceedings of the Institute of Civil Engineers, Part II, pp. 325–378, 1952.
- [35] Wilson, A. G. *Entropy in Urban and Regional Modelling*, Pion, London, 1970.
- [36] Wilson, A. G. *Urban and Regional Models in Geography and Planning*, J. Wiley and Sons, London, 1974.

**ESTIMACIÓN DE MATRICES  
PARA DEMANDA EN TRANSPORTE**

TESIS QUE PRESENTA:  
**MARÍA VICTORIA CHÁVEZ HERNÁNDEZ**

PARA OBTENER EL GRADO DE  
**MAESTRA EN CIENCIAS  
MATEMÁTICAS APLICADAS E INDUSTRIALES**

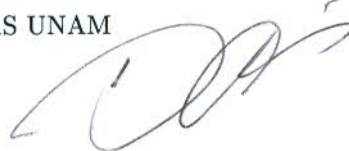
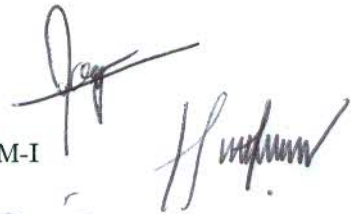
ASESOR: DR. LORENZO HÉCTOR JUÁREZ VALENCIA

JURADO CALIFICADOR:

PRESIDENTE: DR. JOAQUÍN DELGADO FERNÁNDEZ UAM-I

SECRETARIO: DR. LORENZO HÉCTOR JUÁREZ VALENCIA UAM-I

VOCAL: DR. DAVID ROMERO VARGAS UNAM



MÉXICO, D.F. MARZO 2014.