

Máster Interuniversitario en Estadística e Investigación Operativa

Título: Modelos estadísticos de vulnerabilidad para el diseño de redes ópticas dinámicas

Autor: Noemí Germen Galán

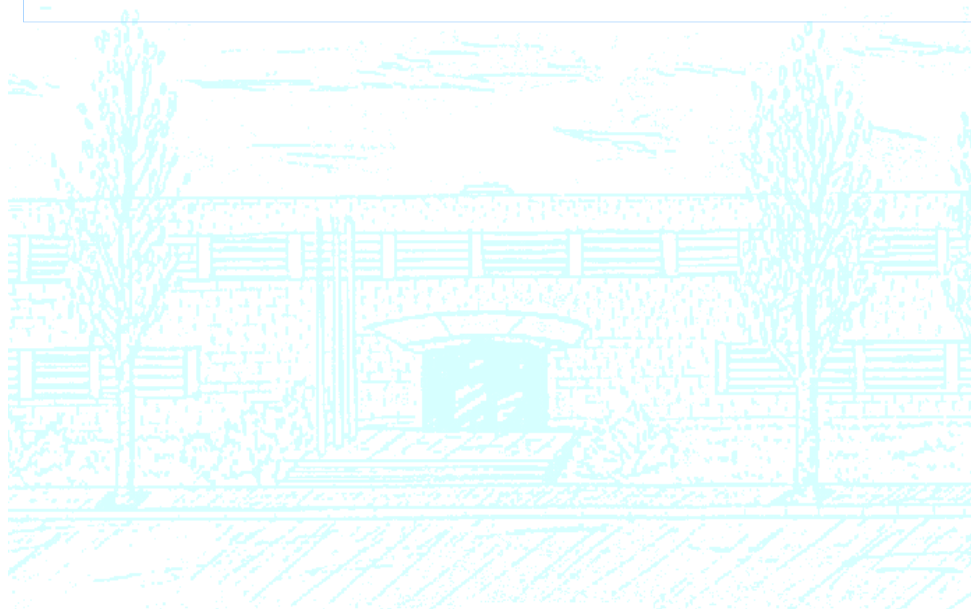
Director: Luis Domingo Velasco Esteban

Co-Director: Marc Ruiz Ramírez

Departamento: Arquitectura de Computadores

Universidad: Universidad Politécnica de Cataluña

Convocatoria: Junio 2014



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística



UNIVERSITAT DE BARCELONA



Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya

Trabajo Fin de Máster

Modelos estadísticos de vulnerabilidad para el diseño de redes ópticas dinámicas

Noemí Germen Galán

Director: Luis Domingo Velasco Esteban

Co- Director: Marc Ruiz Ramírez

UNIVERSIDAD POLITÉCNICA DE CATALUÑA

Departamento de Arquitectura de Computadores

Agradecimientos

Doy las gracias en especial a mis directores Luis Velasco y Marc Ruiz por su paciencia, dedicación y ayuda, tanto en el proceso de obtención de los datos, en la aportación de sugerencias como en introducirme en un mundo de conceptos nuevo para mí como es el de las redes ópticas.

También doy las gracias al Grupo de Comunicaciones Ópticas de la UPC por ceder espacio y equipo para poder hacer el TFM con ellos.

Por último, doy las gracias a todas las personas y amigos por su apoyo y ánimo durante el transcurso del proyecto.

Resumen

En este proyecto se ha abordado la obtención de modelos estadísticos para la predicción del grado de vulnerabilidad de una red. La finalidad de estos modelos es la de servir para el diseño de redes ópticas dinámicas, que se ven sujetas a una evolución constante del tráfico de datos que tienen que soportar. Para conseguir este objetivo, se ha planteado un estudio estadístico desde el diseño del experimento para obtener datos a partir de topologías de operadores reales hasta el análisis y modelización estadísticos a partir de una serie de técnicas comunes. Después de la validación de los modelos y de la evaluación de su potencial de predicción, se han seleccionado los mejores modelos para su integración en problemas de planificación periódica de la red, cuya aplicación se ha introducido con la definición de un caso práctico.

Paraules clau: redes ópticas dinámicas, estadística aplicada, modelos de respuesta binomial

MSC2000: modelo logit, estadística aplicada.

Abstract

In this work, we focus on obtaining statistical models for the prediction of the vulnerability degree on a dynamic optical network. The goal of these models is to be useful for the design of dynamic optical networks subject to changes in the traffic due to the constant evolution. To this aim, we proposed a wide statistical study, covering from the experimental design to obtain meaningful data from real operator topologies to the application of common statistic and regression techniques. After validating the models and evaluating their prediction power, the best ones have been selected for their inclusion in optimization problems dealing with the periodical planning of the network. A use case showing such practical application has been eventually described.

Keywords: dynamic optical networks, applied statistics, binomial models

MSC2000: logit model, applied statistics

Índice

Capítulo 1. Introducción	1
1.1 Motivación y objetivos.....	1
1.2 Organización de la memoria	2
Capítulo 2. Conceptos relacionados	3
2.1 Teoría de Grafos.....	3
2.2 Redes Ópticas.....	5
2.3 Tráfico Dinámico	6
2.4 Recuperación del tráfico ante fallos de enlace.....	7
Capítulo 3. Diseño del experimento	9
3.1 Definición de la variable respuesta.....	9
3.2 Definición de las variables explicativas.....	9
3.3 Metodología de obtención de datos	11
3.3.1 Redes y parámetros de referencia	11
3.3.2 Funcionamiento del simulador.....	11
Capítulo 4. Teoría de modelización	13
4.1 Modelo escogido.....	13
4.2 Modelo logit.....	14
4.2.1 Interpretación del modelo logit	14

4.3	Contrastación y validación del modelo.....	15
4.3.1	Test χ^2 de Pearson	15
4.3.2	Test de devianza	15
4.3.3	Escala o parámetro de dispersión	16
4.4	Normalidad de los Residuos en el modelo logit	16
4.5	Curva de ROC	17
4.6	Predicciones con el modelo	17
Capítulo 5.	Análisis descriptivo	19
5.1	Descriptiva de las variables.....	19
5.2	Análisis de Componentes Principales	21
5.3	Análisis para restaurados y no restaurados por separado	22
5.4	Resumen	25
Capítulo 6.	Estimación de los modelos	27
6.1	Modelos de la red BT	27
6.1.1	Modelo único.....	27
6.1.2	Modelos por categorías.....	32
6.1.3	Modelo único por categorías para la red BT	33
6.1.4	Modelo libre por categorías para BT	35
6.1.5	Elección del mejor modelo por categorías para BT.....	37
6.2	Modelos para la red DT.....	41
6.2.1	Modelo único por categorías de la red DT	43
6.2.2	Modelo libre por categorías para DT.....	45
6.2.3	Elección del mejor modelo por categorías para DT	47
6.3	Modelo de la red TEL	51
6.3.1	Modelo único por categorías de la red TEL.....	53

6.3.2	Modelo libre por categorías para TEL.....	56
6.3.3	Elección del mejor modelo por categorías para TEL	58
6.4	Capacidad predictiva de los modelos	61
6.5	Conclusiones y elección del mejor modelo	62
Capítulo 7.	Caso de aplicación práctica	63
7.1	Diseño periodico en redes dinámicas	63
7.2	Ejemplo de problema de re-planificación dinámica	64
7.2.1	Activación del ciclo de planificación.....	65
7.2.2	Computo de la vulnerabilidad durante la optimización.....	65
Capítulo 8.	Conclusiones.....	67
8.1	Contribuciones y próximos pasos	67
8.2	Valoración personal	68
8.3	Reconocimientos	68
Apéndice A.	Código implementado.....	a
Bibliografía.....		s

Capítulo 1. Introducción

1.1 Motivación y objetivos

La cada vez mayor popularidad y penetración en la sociedad de las aplicaciones basadas en Internet está ampliando de forma dramática la cantidad de tráfico de datos a transportar en las redes de comunicaciones ópticas. Gracias a las más modernas tecnologías de transmisión, la capacidad de estas redes se puede explotar de una manera más profunda y eficiente año tras año. No obstante, el diseño de la mayoría de redes ópticas de comunicaciones operativas hoy en día debe ser replanteado con el fin de adaptarse mejor a una realidad inexistente (e impensable) en los años en los que estas redes fueron ideadas y desplegadas.

De un modo muy rápido, podemos decir que el tráfico está evolucionando en tres direcciones fundamentales: aumentando el volumen, haciéndose más heterogéneo, y tornándose más impredecible. A causa de estas razones (sobre todo la tercera), la operación de las redes está desplazándose desde un esquema estático, donde las conexiones ópticas se establecen de forma manual, a un entorno dinámico donde la red está dotada de los automatismos necesarios para reaccionar ante cualquier evento. Así, una red dinámica es capaz de servir nuevas conexiones y liberar las finalizadas sin necesidad de intervención humana.

En general, las redes de comunicaciones deben de permitir que el servicio pueda mantenerse ante el evento de un fallo (por ejemplo, un corte en un enlace de fibra). Una red dinámica tiene que ser capaz de detectar cuando un fallo ha impactado en la red y lanzar un mecanismo de restauración que permita, en decenas de milisegundos, encontrar nuevas rutas para las conexiones afectadas por el fallo y, así, restablecer el servicio. El éxito de este mecanismo de restauración, que no permite asegurar el 100% de restablecimiento del servicio, depende en gran medida del diseño de la red y de las características del tráfico.

En un entorno como el descrito anteriormente, donde el tráfico evoluciona de forma incierta, cabe pensar lo que hoy puede resultar un buen diseño de la red, se convierta en insuficiente el día de mañana. Así pues, una red sujeta a fallos diseñada para ser robusta tenderá a ir volviéndose más vulnerable a medida que el escenario de tráfico cambie. Por ello, es importante medir el grado de vulnerabilidad y detectar cuando y donde se debe hacer una reconfiguración o ampliación de la red. Esta decisión tiene que ser tomada a partir de datos que se comportan de forma aleatoria, por lo tanto el uso de modelos estadísticos resulta de gran utilidad.

En este proyecto nos centraremos en la obtención de modelos estadísticos para el diseño de redes ópticas de comunicaciones en entornos de tráfico dinámico. Específicamente, nuestro objetivo es el de hallar modelos que permitan estimar, sobre una red en operación sometida a fallos de enlace, si el tráfico afectado por un fallo va a poder ser restaurado completamente o no. Así, una red en operación puede ser reconfigurada o ampliada cuando, a partir de la

predicción con estos modelos, se detecte que la red es lo suficientemente vulnerable como para requerir una actualización. Estos modelos deben permitir detectar qué partes de la red son más vulnerables que otras y, por lo tanto, sobre qué zonas debe centrarse la reconfiguración o ampliación.

Para realizar este estudio, partimos de las topologías de referencia proporcionadas por una serie de operadores de red nacionales, a saber, Telefónica, Deutsche Telekom, y British Telecom. A partir de estas topologías y de una serie de datos y parámetros de referencia, simularemos su comportamiento en un entorno de tráfico dinámico realista sujeto a fallos en los enlaces de fibra óptica. Con los datos recabados de las simulaciones, se obtendrán varios modelos de predicción del grado de vulnerabilidad de un enlace de la red. Tras un exhaustivo proceso de modelización, validación, y predicción, escogeremos aquel o aquellos modelos que permitan predecir tal respuesta con la mayor precisión posible. Finalmente, se ilustrará el uso de estos modelos para el diseño de redes en entornos dinámicos, haciendo hincapié en el escenario de planificación periódica de la red, para el cual se necesita además proveer de un módulo que permita actualizar y reajustar los modelos ante los cambios imprevisibles del tráfico en el futuro.

Debido a la restricción, en términos de tiempo, que supone obtener un volumen suficiente de observaciones para cada red y que nos centramos en analizar las redes antes mencionadas, se busca encontrar un modelo para cada una de ellas. No obstante, el objetivo es intentar encontrar variables que sean significativas para todas las redes, con el fin de proponer variables que pudieran ser utilizadas en un modelo más complejo que pudiera ser adaptable a cualquier red de unas características concretas.

1.2 Organización de la memoria

El resto de la memoria está organizado de la siguiente forma: el capítulo 2 explica todos los conceptos necesarios sobre redes para tener una buena comprensión del proyecto. Se introducen conceptos de teoría de grafos, así como las características de las redes ópticas y el tráfico dinámico. El capítulo 3 define el escenario de trabajo así como la definición de la variable respuesta, las variables explicativas de la topología y las variables explicativas de la carga. El capítulo 4 define aquellos conceptos teóricos necesarios para la modelización de los datos, la utilización del modelo logit, así como los conceptos teóricos utilizados para su validación y posterior predicción. En el capítulo 5 se presenta la descriptiva más relevante de los datos, distinguiendo entre los distintos niveles de la variable respuesta. En el capítulo 6 se presenta el análisis y los resultados principales de la modelización de las tres redes, así como su validación y posterior predicción. En el capítulo 7 se presenta un caso de aplicación de los modelos estadísticos obtenidos para el diseño de redes ópticas dinámicas. Finalmente, en el capítulo 8 se presentan las conclusiones y contribuciones del presente proyecto, además de definir los siguientes pasos de continuación de este trabajo.

Adicionalmente, se añaden varios apéndices aportando aquellos datos más relevantes no incluidos en la memoria, tales como parte del código implementado y resultados extendidos de las figuras y tablas presentados a lo largo del documento.

Capítulo 2. Conceptos relacionados

En el presente capítulo se van a definir una serie de conceptos a partir de los cuales obtendremos una comprensión de este trabajo relacionado con las redes ópticas de comunicaciones. En primer lugar facilitaremos diversas definiciones de teoría de grafos que son útiles para caracterizar las topologías de las redes de comunicaciones. Seguidamente se presentaran la composición y funcionamiento de las redes ópticas flexgrid (que es la tecnología asumida en este proyecto). Finalmente se definen las figuras más relevantes relacionadas al tráfico dinámico y esquemas de recuperación del tráfico ante fallos.

2.1 Teoría de Grafos

A la hora de hablar de topología o estudio de una red, se puede hacer a través de un grafo representado como $G(N,E)$, donde N representa el conjunto de nodos (o vértices) y E es el conjunto de enlaces, siendo $|N|$ y $|E|$ el número de nodos y enlaces respectivamente. Los enlaces pueden ser unidireccionales o bidireccionales. A la hora de medir el tamaño de la red, además de N y E , el valor relativo más usado es el grado nodal medio (δ). Primeramente el grado nodal de un nodo es el número de nodos con el cual está relacionado, por lo tanto, el grado nodal medio es la media de todos los grados nodales, es decir, la media del número de arcos incidentes en él. Se puede calcular como:

$$\delta = \frac{2 \cdot |E|}{|N|} \quad (2.1)$$

Adicionalmente al grado nodal, se pueden encontrar variables que caractericen la posición de un nodo respecto al resto. Por ejemplo, el coeficiente de clúster de un nodo se calcula como se muestra en la ecuación 2.2, donde d representa el grado nodal del nodo e y es el número de arcos entre nodos adyacentes con dicho nodo. A partir de estos valores, se puede calcular el coeficiente de clúster de una red como la media de los coeficientes de clúster de cada nodo.

$$c = \frac{2 \cdot y}{d + (d - 1)} \quad (2.2)$$

Sobre este grafo pueden calcularse caminos o rutas. Un camino o ruta en un grafo es una sucesión de enlaces o nodos contiguos que unen un par de nodos del grafo. De todos los caminos más cortos entre un par de nodos (i,j) , el camino mínimo h_{ij} es aquel cuyo coste es el menor posible. Este coste suele calcularse en número de saltos (número de enlaces usados por el camino) o en distancia recorrida (si los enlaces llevan asociada alguna distancia). De ahora en adelante solo trabajaremos con caminos mínimos en número de saltos. Así, el camino mínimo medio de la red (h) se define como el camino más corto en media entre cualquier par de nodos (i, j) .

A partir de estas distancias mínimas, dos variables fundamentales del grafo pueden ser calculadas: el radio y el diámetro. El radio (*rad*) es la distancia más corta de entre los caminos mínimos más largos que salen de un nodo, por otro lado, el diámetro (*dia*) es el camino mínimo más largo de la red. La definición formal de ambos es la siguiente:

$$rad = \min_{i \in N} \left(\max_{\substack{j \in N \\ i \neq j}} (h_{ij}) \right) \quad (2.3)$$

$$dia = \max_{i \in N} \left(\max_{\substack{j \in N \\ i \neq j}} (h_{ij}) \right) \quad (2.4)$$

Otro concepto relacionado con estas distancias mínimas es el *coeficiente de intermediación* (o *betweenness*) definido como la medida que relaciona los caminos mínimos de un grafo y el grado de centralidad de un elemento, sea nodo o arco. En otras palabras, para un cierto nodo (o arco) se calcula el número de caminos mínimos que pasan a través de ese nodo (o arco). A partir de los valores de *intermediación* de todos los elementos del grafo se puede calcular adicionalmente la media y la desviación estándar. La segunda proporciona información acerca de cómo se distribuyen los caminos mínimos por la red, un valor bajo indica el hecho de estar todos los nodos (o arcos) utilizados de forma más o menos uniforme, por consiguiente, la carga de una red de transporte con esta tipología, estará distribuida de forma natural, es decir, de forma más balanceada pudiendo así resultar efectiva para un mejor uso de la capacidad.

Una de las características principales de un grafo es la conectividad, que puede ser medida de varias formas. Podemos decir que un grafo es *k*-conexo si es necesario eliminar un mínimo de *k* enlaces para desconectar el grafo. Por lo general, las redes de comunicaciones son al menos 2-conexas. Otra forma de medir la conectividad de una red es a través de la información que proporciona el análisis espectral del grafo. Para ello, es necesario construir la matriz *Q* formada por la resta de la matriz Laplaciana Λ y la matriz de adyacencias *M*. La matriz Laplaciana Λ es una matriz diagonal de tamaño $|N| \times |N|$ donde el elemento *i* de la diagonal contiene el grado nodal del nodo *i*. Por otra parte, la matriz de adyacencias *M* tiene dimensión $|N| \times |N|$ y cada celda m_{ij} vale 1 si existe un enlace entre los nodos *i* y *j*, y 0 en caso contrario. De los valores propios de esta matriz, podemos obtener varias variables, de las cuales las más importantes son el segundo y el quinto valor propio, que se denominan como conectividad algebraica y radio espectral, respectivamente. En la Fig. 2.1 podemos ver el valor de la conectividad algebraica para varias topologías. Un valor mayor de conectividad implica un número mayor de enlaces a eliminar para desconectar la red (líneas punteadas).

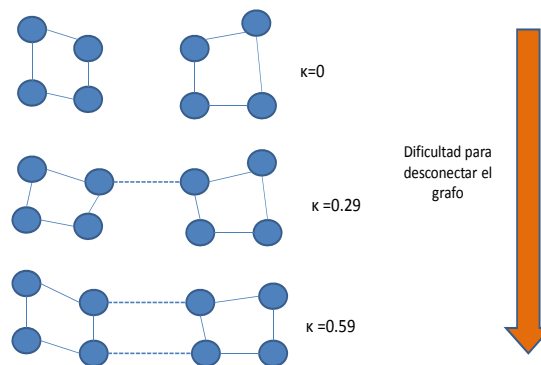


Fig. 2-1 Ejemplo de conectividad algebraica

Pueden encontrarse más detalles acerca de estas variables y otras de similar índole en [Vm11].

2.2 Redes Ópticas

Las redes ópticas están constituidas por una topología como la descrita en el apartado 2.1, junto con una serie de equipos y dispositivos ópticos. Un enlace se corresponde con una fibra óptica y un nodo representa todo el equipamiento necesario para poder establecer y transformar conexiones ópticas a través de las fibras de la red. Una fibra óptica, tiene una capacidad determinada útil a la cual denominaremos espectro o ancho espectral (S) medida en Giga-Hertzios (GHz) y viene caracterizada por el tipo de equipamiento instalado en los nodos.

Una conexión óptica es un camino (también llamado circuito) el cual se establece en la red y permite transmitir servir el tráfico de una demanda. En una misma fibra y momento se puede transportar diversas conexiones de forma simultánea porque cada una de ellas ocupa una porción del espectro óptico diferente. De otra forma, si dividimos el espectro en varios fragmentos (o longitudes de onda), cada uno solo podrá soportar una conexión en un momento dado. Las conexiones ópticas tienen el fin de servir las peticiones de conexión de clientes que llegan a la red. El objetivo es encontrar una ruta con espectro suficiente para poder atender el ancho de banda demandado por el cliente. Una petición de conexión consta mayoritariamente de la siguiente información: los nodos de origen y destino, el destino donde se va a realizar la conexión y la cantidad de ancho de banda a transportar, expresado normalmente en Megabits por segundo (Mbps) o Gigabits por segundo (Gbps).

En este momento, podemos definir dos tipos de redes ópticas: rígidas y elásticas. En una *red óptica rígida* cada conexión óptica ocupa la misma cantidad de espectro durante todo su tiempo de operación, siendo el ancho ocupado igual para todas las conexiones. En contraste, una *red óptica elástica* permite alojar cada conexión con un ancho de banda distinto que, además, puede variar a lo largo del tiempo. En este proyecto, nos centraremos en redes ópticas flexibles.

Entre las tecnologías estudiadas para el desarrollo de redes ópticas elásticas, una de las cuales tiene un mejor resultado entre rendimiento y complejidad tecnológica, es la tecnología *flexgrid* [Li11]. En este tipo de redes, el espectro óptico está dividido en porciones con la misma anchura, denominadas segmentos. Dicha anchura puede ser distinta según sean los equipos instalados en la red, siendo Δ el parámetro de anchura del segmento en GHz (normalmente serán valores pequeños como 6.25GHz). Cada segmento ocupa una posición en el espectro caracterizado por su frecuencia central, lo cual permite ordenar los segmentos de forma lógica y ordenando las frecuencias centrales de menor a mayor. Cada conexión tiene un determinado número de segmentos en función del ancho de banda requerido para la conexión. Esta porción de espectro formada por segmentos contiguos y destinados al transporte de una conexión se denomina *canal* y puede caracterizarse mediante su frecuencia central y el número de segmentos. En la Fig. 2.2 podemos ver una representación de un enlace de una red óptica flexgrid.

El espectro asignado a cada conexión óptica debe cumplir dos condiciones: la *contigüidad* y la *continuidad de espectro*. La contigüidad es la restricción de que todos los segmentos de un canal sean contiguos, es decir, sus frecuencias centrales sean consecutivas. Por otro lado, la continuidad de espectro requiere que el canal asignado a una conexión sea el mismo en todos los enlaces usados por esta. Así pues, encontrar una conexión que cumpla estas condiciones no es una tarea fácil y requiere de la resolución de un problema de optimización que

comúnmente se denomina como problema de enrutamiento y asignación de espectro (en inglés, *routing and spectrum assignment, RSA*). En la Fig 2-3 observamos como una conexión entre los nodos *b* y *d* de dos segmentos no puede ser servida por la ruta más corta debido a esta necesidad de continuidad. Así, se pueden observar canales contiguos de 2 segmentos en la ruta más corta pero no coinciden en el espectro. Para ello es necesario servir la conexión por una ruta más larga con canales continuos disponibles.

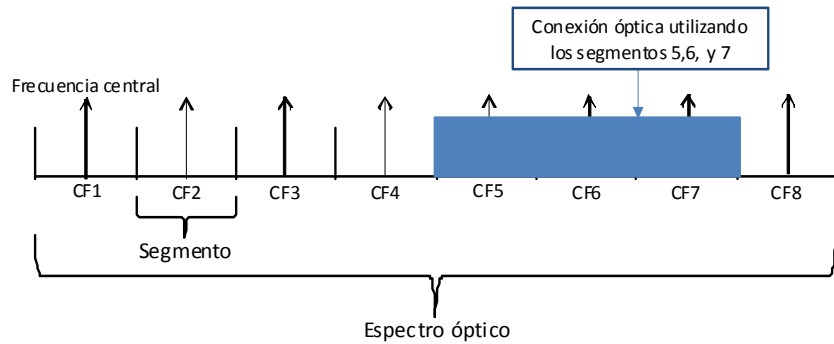


Fig. 2-2 Representación de un enlace de fibra óptica

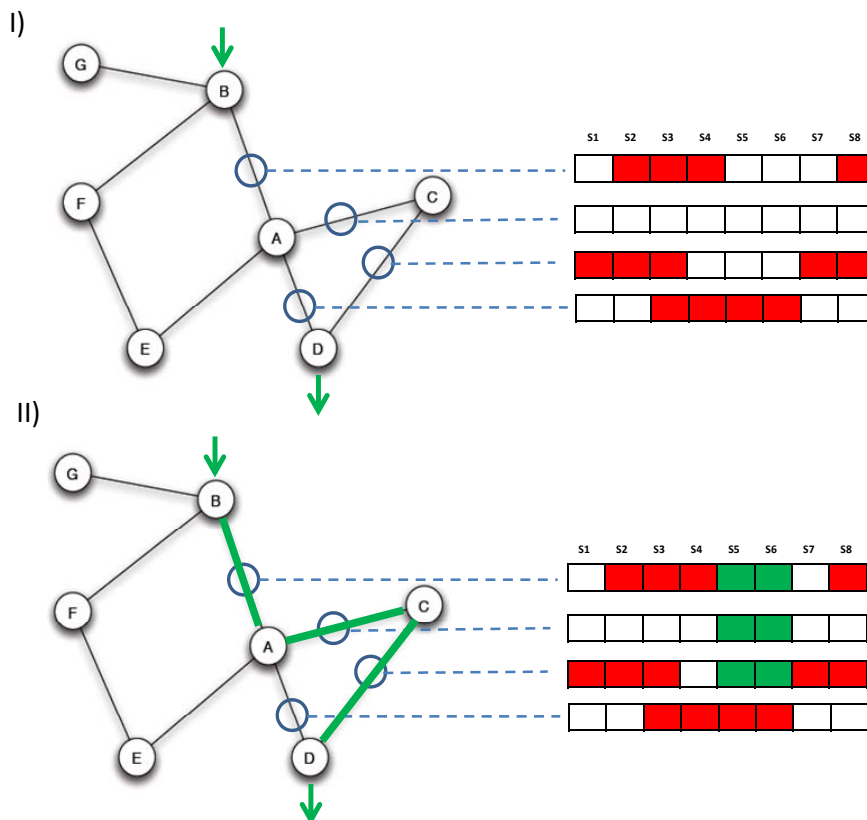


Fig. 2-3 Establecimiento de una conexión de la red óptica

2.3 Tráfico Dinámico

Las redes ópticas elásticas se despliegan con el objetivo de proporcionar servicio de conectividad a ciertos clientes mediante conexiones que permanecen establecida durante un tiempo hasta que se eliminan, liberando así recursos en la red que podrán ser utilizados por

otras conexiones. Este funcionamiento dinámico se puede modelar a partir de varios fundamentos de teoría de tele-tráfico [ITU05].

Podemos medir la cantidad media de conexiones en la red en un determinado instante como intensidad de tráfico (I). De forma genérica, consideraremos que las peticiones de conexión llegan a un nodo de la red siguiendo una distribución de Poisson con un cierto tiempo medio entre llegadas (iat). Por otra parte, el tiempo que esa conexión permanece establecida (si ha podido ser servida) sigue una distribución exponencial con un cierto tiempo medio de servicio (ht). Entonces, la intensidad se calcula como $I = ht / iat$ y la unidad de medida estándar es el Erlang. Como ya hemos visto, en redes elásticas, las conexiones ocuparán un tamaño de espectro diferente según sea la petición de ancho de banda. Eso nos lleva a concluir que la intensidad no es suficiente para caracterizar el tráfico transportado en la red. Por eso, definimos el perfil de tráfico (TP) como la proporción media de peticiones de cada tipo de cliente que llega a la red. Este perfil, que tendrá como dimensión el número de clases distintas de cliente (caracterizadas por su demanda en Gb/s), es un vector de distribución de probabilidades cuya suma debe dar 1.

En entornos de tráfico dinámico, el grado de servicio de la red se mide con la probabilidad de que una petición de conexión no pueda ser aceptada en el momento de su llegada. Esta probabilidad se denomina probabilidad de bloqueo (PB) y es una de las figuras más usadas para evaluar la capacidad efectiva de la red. Así, podemos considerar una probabilidad de bloqueo del 1% como el valor más alto admisible para dar un servicio mínimamente decente a nuestros clientes. Aunque la red tenga capacidad sobrante, no vamos a permitir más intensidad ya que empeoraría el grado de servicio alcanzando un nivel no deseable. Otra vez aquí, nos encontramos con que no es lo mismo rechazar una demanda de 400 Gbps que una de 10 Gbps. Así, en redes ópticas elásticas se suele hablar de la probabilidad de bloqueo ponderada, haciendo uso del TP para dar un peso distinto a cada conexión rechazada.

2.4 Recuperación del tráfico ante fallos de enlace

En este proyecto estudiaremos la red ante el evento de fallos (cortes) en un enlace de fibra. Nótese que este evento no es menor, ya que una red nacional de tamaño medio está sujeta a varias decenas de este tipo de fallo en el transcurso de un año. Asumiremos que en el tiempo transcurrido entre el impacto de un fallo y la reparación de éste (tradicionalmente 12h-24h para una red nacional con un servicio de mantenimiento normal), ningún otro fallo impacta la red. Ante este escenario, los operadores de red pueden apostar por dos esquemas de recuperación de las conexiones ópticas afectadas por el corte, que son la protección y la restauración.

El esquema de protección consiste esencialmente en, a la hora de establecer la conexión óptica para servir la petición, establecer una conexión secundaria de protección que no comparta ningún enlace con la conexión primaria. Así, ante el evento de un fallo en cualquiera de los enlaces de la conexión primaria, el tráfico se conmuta rápidamente (pocos milisegundos) a la conexión secundaria sin que se vea afectado significativamente el servicio. Este esquema tiene múltiples variantes, entre las cuales destacan la protección dedicada y la protección compartida [Ve09]. En cualquiera de ellos, se puede asegurar que el 100% del tráfico afectado por un fallo podrá ser recuperado. Finalmente, cabe destacar que la máxima eficiencia de este esquema se obtiene cuando el tráfico es estático (sin cambios entre periodos de ampliación de la red) o poco dinámico (las conexiones se establecen por semanas o meses).

Bajo un esquema de restauración, las conexiones se sirven con una única conexión primaria. Cuando un fallo afecta a ésta, se ejecuta automáticamente un mecanismo que encuentra, con los recursos de capacidad disponible, una nueva ruta para la conexión que no utilice en enlace afectado. Bajo este mecanismo, no se puede asegurar el 100% de la recuperación del tráfico (restorabilidad), ya que ésta depende en gran medida de los recursos disponibles en el momento del fallo. Además, durante el tiempo necesario para obtener la nueva ruta y establecer la conexión de restauración (en el orden de cientos de milisegundos), el tráfico afectado por el corte resta bloqueado. No obstante, la mayor eficiencia en el uso de recursos con respecto a la protección hace de la restauración el método más apropiado para entornos de tráfico muy dinámico (cuando el tiempo de servicio de las conexiones es del orden de minutos).

En este proyecto, asumiremos un escenario de tráfico dinámico sujeto a fallos únicos de enlace donde se recupera el tráfico afectado mediante un mecanismo de restauración.

Capítulo 3. Diseño del experimento

El presente capítulo consta de 2 partes. En la primera se presenta la definición de la variable respuesta, así como de las variables explicativas. En la segunda parte, se presentan los escenarios de referencia usados para la obtención de datos y se describe brevemente el funcionamiento del simulador utilizado para la obtención de datos de la variable respuesta.

3.1 Definición de la variable respuesta

Tal y como se ha especificado en la introducción, los modelos estadísticos resultantes de este proyecto deben permitir estimar si la restauración del tráfico afectado por un fallo en un enlace se podrá restaurar completamente o no. Para ello, definimos el concepto de vulnerabilidad como el resultado de dicha restauración. Así, la variable respuesta de nuestro estudio, llamada Y , la cual puede tomar dos únicos niveles, a saber:

- $Y=1$, cuando todas las conexiones afectadas por un fallo se pueden restaurar.
- $Y=0$, cuando al menos una conexión no se puede restaurar.

En otras palabras, cuando $Y=0$ diremos que la red es vulnerable a ese fallo. Los modelos para predecir Y serán funciones de variables que tienen relación con la topología y con el tráfico transportado en la red. A continuación se describen las variables explicativas utilizadas.

3.2 Definición de las variables explicativas

Como se verá más adelante en el proceso de obtención de datos, las observaciones se generan por cada fallo de enlace que se procesa. Así, una observación está relacionada con un corte en un enlace específico de la red. Esa observación se puede caracterizar por dos tipos de variables explicativas: las que tienen que ver con la topología de la red resultante tras el impacto del fallo y las que miden el número, volumen, y tipo de las conexiones que usaban en enlace antes del fallo (y que han debido de ser restauradas por el mecanismo de restauración).

Antes de entrar en los detalles de las variables, podemos ver como cualquier variable que describe la topología (por ejemplo conectividad algebraica, diámetro, radio espectral, etc) puede ser calculada cada vez que hay un fallo con la red residual tras eliminar el enlace fallado. Así, si consideramos, por ejemplo, que la variable *algConnect* contiene el valor de la conectividad algebraica de la red cuando todos los enlaces están operativos, definimos *algConnectF* como la conectividad algebraica resultante tras eliminar el enlace fallado. Además, definimos *algConnectVar* como el cociente entre *algConnectF* y *algConnect*. Utilizaremos esta notación para varias topologías de la red.

La lista de variables explicativas de la topología para cada observación (con el nombre con el que serán identificadas en los ficheros de datos) són:

link_node_A: Identificador de primer nodo del enlace fallado
link_node_B: Identificador de segundo nodo del enlace fallado
radiusF: Radio de la red tras el fallo
radiusVar: Variación en el radio de la red
diameterF: Diámetro de la red tras el fallo
diameterVar: Variación en el diámetro de la red
algConnectF: Conectividad algebraica de la red tras el fallo
algConnectVar: Variación en la conectividad algebraica de la red
spectralRadiusF: Radio espectral de la red tras el fallo
spectralRadiusVar: Variación en el radio espectral de la red.
avgLbetF: Coeficiente de intermediación medio tras el fallo
avgLbetVar: Variación en el coeficiente de intermediación medio tras el fallo

A parte de esta lista de variables descriptivas de la topología después del fallo, guardamos información sobre el tráfico llevado en el enlace antes del fallo. Para ello, es necesario previamente describir dos conceptos. En primer lugar, guardamos el ancho de banda ocupado por cada conexión (definido como *bw_path*) el cual se calcula como el ancho de banda de la conexión multiplicado por la longitud en número de saltos de la ruta. La interpretación de esta variable es que, a mayor sea su valor, más largas y pesadas serán las conexiones.

Analizando el origen y destino de las conexiones que atraviesan el enlace fallado, podemos separar tres grupos: *grup1* que contiene todas aquellas conexiones que tienen el origen y el destino en los dos pares de nodos adyacentes al enlace fallado; *grup2* que contiene todas aquellas conexiones que tienen o bien el origen o bien el destino en alguno de los nodos adyacentes del enlace fallado; y *grup3* que contiene el número de conexiones que ni empiezan ni acaban en alguno de los nodos adyacentes del enlace fallado.

Con las descripciones anteriores realizadas, podemos describir la lista de variables descriptivas de la carga consideradas:

min_bw_path: Valor mínimo de la variable *bw_path* de todas las conexiones en el enlace
mean_bw_path: Valor medio de la variable *bw_path* de todas las conexiones en el enlace
max_bw_path: Valor máximo de la variable *bw_path* de todas las conexiones en el enlace
min_path_length_before: Longitud de la ruta más corta (en número de saltos) de una conexión en el enlace.
mean_path_length_before: Longitud media de la ruta (en número de saltos) de una conexión en el enlace.
max_path_length_before: Longitud de la ruta más larga (en número de saltos) de una conexión en el enlace.
Grup1: Número de conexiones que pertenecen al grupo 1
X.Grup1: Porcentaje de conexiones (respecto del total) que pertenecen al grupo 1
Grup2: Número de conexiones que pertenecen al grupo 2
X.Grup2: Porcentaje de conexiones (respecto del total) que pertenecen al grupo 2
Grup3: Número de conexiones que pertenecen al grupo 3
X.Grup3: Porcentaje de conexiones (respecto del total) que pertenecen al grupo 3

Con el fin de obtener datos para varias redes de referencia, hemos hecho uso de un simulador, cuya configuración y funcionamiento se describen en la siguiente sección.

3.3 Metodología de obtención de datos

3.3.1 Redes y parámetros de referencia

Los datos usados para la modelización se han obtenido a partir de las redes de referencia de tres operadores nacionales: Telefónica (TEL), British Telecom (BT), y Deutsche Telekom (DT). La Fig. 3-1 muestra la topología de las tres redes así como el número de nodos, enlaces, y el grado nodal medio.

En el simulador hemos configurado cada una de estas redes con enlaces con un ancho espectral de 2 THz y un tamaño de segmento de 6.25 GHz. La conexiones generadas pueden pertenecer a 3 velocidades distintas: 40, 100, y 400 Gb/s, con una proporción de 67%, 27%, y 6% respectivamente. La red es operada a una intensidad de tráfico que devuelve en promedio una probabilidad de bloqueo cercana al 1%. Para dicha intensidad, del entorno de un 10% de los fallos que impactan la red no pueden ser restaurados completamente, lo que asumimos como el umbral máximo de vulnerabilidad permitida. Los fallos se generan con un tiempo medio entre fallos consecutivos lo suficientemente grande como para asegurar que ha habido un recambio completo de las conexiones establecidas en la red.

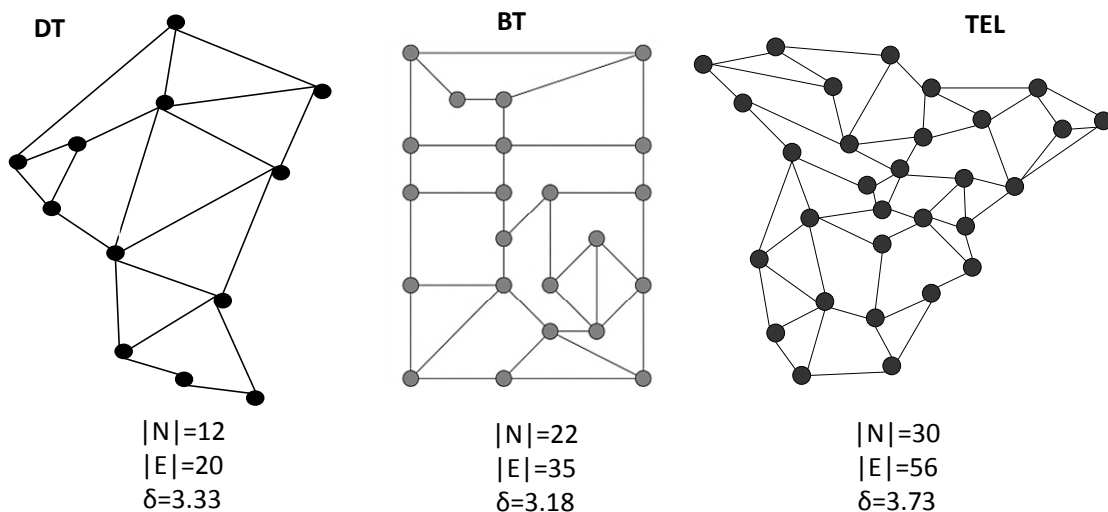


Fig. 3-1 Topologías usadas

3.3.2 Funcionamiento del simulador

Para obtener los datos de la variable respuesta hemos utilizado un simulador de tráfico dinámico en redes ópticas flexibles implementado en OMNeT++ [OMNeT]. Este entorno permite desarrollar cualquier simulador de eventos discretos con aplicación a las redes de comunicaciones. Los eventos más importantes que se resuelven en el simulador son:

Inicialización: dado que la generación de conexiones entrantes se realiza independientemente por cada nodo, se genera aleatoriamente el tiempo en que se recibirá la primera petición de conexión con origen en cada uno de estos nodos. Además, se genera el tiempo del primer fallo de cada uno de los enlaces siguiendo una distribución de probabilidad exponencial con una tasa que depende de la longitud en km del enlace de fibra.

Llegada de petición de conexión: al recibir este evento, en primer lugar se computa en qué momento se recibirá la siguiente petición con origen en ese nodo. En segundo lugar, se elige el nodo de destino de la petición siguiendo una distribución uniforme entre el resto de nodos de

la red. Luego, se ejecuta el algoritmo de enrutamiento y asignación de espectro definido en [Ca12] con el fin de encontrar una conexión óptica que sirva dicha petición. Nótese que este algoritmo no siempre encuentra una conexión factible a causa de la falta de recursos de capacidad disponible; en ese caso la petición de conexión es rechazada. En el caso de ser aceptada, se calcula el tiempo que esta conexión estará establecida a partir de una distribución exponencial con una tasa de servicio media que se configura mediante parámetros externos. Al calcular este valor, se genera un nuevo evento, el de fin de establecimiento de dicha conexión. Finalmente, estadísticas tales como si la petición se ha aceptado o no, que ruta y que espectro se ha asignado en caso de aceptación, así como otros resultados se guardan en el fichero de datos de salida correspondiente al aprovisionamiento de conexiones.

Final de establecimiento de conexión: cuando llega un evento como este, se liberan todos los recursos de capacidad utilizados por dicha conexión y esta se elimina de la lista de conexiones actualmente establecidas en la red.

Fallo en un enlace: igual que en el caso de recibir una nueva petición de conexión, el primer paso consiste en generar el siguiente fallo en dicho enlace. Después, el enlace se etiqueta como fallado, las conexiones que usan ese enlace se marcan como conexiones a restaurar, y se liberan todos los recursos ocupados por dichas conexiones. Entonces, el algoritmo de restauración dinámica detallado en [Ca14] se ejecuta con el fin de encontrar una ruta alternativa para cada una de las conexiones afectadas la cual evite el enlace que acaba de fallar. El éxito de esta restauración puede no ser completa, es decir, alguna de las conexiones puede ser rechazada por falta de recursos de capacidad en la red. En ese caso, se marca el fallo como no restaurado. Por el contrario, si el tráfico afectado por el corte ha sido satisfactoriamente re-enrutado, entonces el fallo se marca como restaurado. Este resultado, además de los detalles de la restauración, se guardan en el fichero de salida de resultados de la restauración.

Fin de la simulación: cuando el número de peticiones de conexión generadas llega a un cierto umbral (fijado en los parámetros de ejecución del simulador), la simulación se finaliza, retornando los ficheros de resultados de aprovisionamiento y restauración. Es este fichero de resultados de restauración el que contiene las observaciones de la variable respuesta que se utilizarán para la obtención de los modelos.

La configuración del simulador se establece de modo que aseguramos que un número significativo de fallos (más de 100) impacta cada uno de los enlaces de la red. Así podemos obtener varias observaciones que afectan a un mismo enlace. A este conjunto de datos se le aplicarán diversas técnicas estadísticas con el fin de obtener y validar modelos. En el siguiente capítulo se describen los conceptos teóricos fundamentales de dichas técnicas estadísticas.

Capítulo 4. Teoría de modelización

En este capítulo se van a explicar y definir los conceptos y técnicas estadísticas utilizadas para realizar el análisis del presente trabajo. Se presentará el tipo de modelo más adecuado para analizar los datos asumiendo que la variable respuesta es binomial, además de explicar su interpretación. Se introducirá el concepto de odds ratio con el cual se interpretan los coeficientes de los modelos. Para la validación del modelo se utilizará el test de Pearson y el test de devianza, además de calcular la sobredispersión de los datos a partir del parámetro de dispersión. Para escoger el mejor modelo, se utilizará el análisis de la devianza, ANOVA, y la curva de ROC. Para finalizar, se mostrará cómo se calculan las predicciones a través de las tablas de contingencias.

4.1 Modelo escogido

A la hora de ejecutar un modelo donde la variable endógena Y es una variable dicotómica, el modelo lineal de elección binaria tiene diferentes métodos para desarrollar el modelo. Los más utilizados son el modelo lineal de probabilidad truncado, el modelo logit y el modelo probit. Cuando la variable endógena se supone como una función de distribución uniforme, estamos ante el modelo lineal de probabilidad lineal truncado, cuando se distribuye como una función de distribución normal 0 y 1 estamos ante un modelo probit y, en el último caso, si la función de distribución de la variable endógena es una binomial estamos ante un modelo logit. La diferencia entre estos modelos respecto al modelo de regresión lineal comúnmente usado, es el hecho de tener la variable endógena como una función de distribución uniforme, normal o logística en vez de tener un comportamiento cuantitativo.

El modelo de probabilidad lineal truncado se caracteriza por realizar una regresión sobre la variable respuesta binaria utilizando la metodología de Mínimos Cuadrados Ordinarios (MCO). En este caso, se encuentran limitaciones tales como tener el valor de la estimación del modelo fuera del rango 0 – 1, La perturbación aleatoria no puede seguir una distribución normal, presenta problemas en la heterocedasticidad y el coeficiente de determinación R^2 está subestimado.

El modelo probit, como se ha mencionado anteriormente, tiene función de distribución de una normal para la variable endógena y el modelo logit la distribución binomial (también llamada logística). Tal y como se comenta en [Me03], *“la similitud existente entre las curvas de la normal tipificada y de la logística, los resultados de ambos modelos no difieren entre sí, siendo las diferencias operativas, debidas a la complejidad que presenta el cálculo de la función de distribución normal frente a la logística”*. Por este motivo se suele utilizar la función logit binomial y es la utilizada en el siguiente proyecto.

4.2 Modelo logit

El modelo logit se puede interpretar en términos probabilísticos para medir la probabilidad de ocurrencia ($Y_i=1$) cuando la variable endógena es una función de distribución binomial. Esto hace que el modelo logit relacione la variable endógena Y_i con las variables explicativas X_{ki} a través de una función de distribución. Esta función de distribución utilizada tal y como se ve en la ecuación 4.1 puede interpretarse en términos probabilísticos para medir la probabilidad de ocurrencia del acontecimiento de estudio ($Y_i=1$), pero tiene un inconveniente y es el hecho de no poder interpretar los coeficientes de las variables del modelo al igual que se hace en los modelos lineales convencionales.

$$Y_i = \frac{e^{\left(\sum_{r=0}^R b_r x_{qr}\right)}}{1 + e^{\left(\sum_{r=0}^R b_r x_{qr}\right)}} + \varepsilon_i \quad (4.1)$$

4.2.1 Interpretación del modelo logit

Cuando se muestra el resultado del modelo logit, los valores de los coeficientes corresponden al valor del test de Wald. El estadístico de Wald proporciona saber si es verdadero el valor del parámetro de la muestra o no utilizando la chi-cuadrado para realizar el contraste, dando significativo el p-valor cuando se rechaza la hipótesis nula de no ser la variable influyente en el modelo [Ha01]. Por ello para poder interpretarlos, no se pueden utilizar directamente dichos valores sino que se ha de hacer de otro modo.

El no poder interpretarlos de forma convencional no significa que no se pueda utilizar la información aportada. El hecho de ver si los coeficientes de las variables son significativos nos informan si esta influye en el modelo y nos aporta información y por otro lado, se observa el signo de los coeficientes. Estos nos indicarán la dirección de aumento de la probabilidad cuando aumenta la variable explicativa correspondiente, siendo un signo negativo del coeficiente de la variable, disminuciones en la probabilidad de tener ($Y=1$) y un signo positivo aumentos en la probabilidad de tener ($Y=1$). Una vez tenemos esta información, para poder interpretar los coeficientes se calcula el cociente entre la probabilidad de ocurrencia frente a la probabilidad de no suceder, tal y como se indica en la ecuación 4.2. Este suceso se denomina odds ratio y es la ventaja odds del suceso y se interpreta en términos del cambio proporcional (sea en aumento o en disminución) producidos en el odds ratio de la variable concreta por cada unidad de cambio en la variable endógena.

$$Prob(Y_i = 1) = M_i = \frac{M_i}{1 - M_i} = e^{\alpha + \beta_k x_{ki}} \quad (4.2)$$

El valor odds ratio resultante puede tener tres casos, el primero es cuando el coeficiente de la variable del modelo resulta ser igual a 0, en este caso su odds ratio es igual a 1, lo cual no produce ningún efecto de dicha variable sobre la variable respuesta del modelo. El segundo se ocurre cuando el valor del coeficiente de la variable a calcular el odds ratio es superior a 0 y positivo, indicando un valor de odds ratio superior a la unidad, lo cual implica que para cada unidad de aumento del coeficiente de la variable, esta producirá un efecto significativo sobre la variable respuesta y más cercano a la probabilidad de ocurrir ($Y=1$). Por último, cuando es menor a 0 y por tanto negativo, un incremento de una unidad del coeficiente de la variable, producirá un efecto negativo sobre la variable respuesta haciendo que la probabilidad de tener un fallo no restaurado sea mayor.

Además, al igual que en los modelos lineales comunes donde la variable endógena es

cuantitativa, se analiza el modelo en términos de AIC definido por Akaike [Hi74] para valorar la calidad del mismo, por ello un modelo con buena calidad será aquel cuyo valor del AIC sea menor. En apartado siguiente se continúa con la contrastación y validación del modelo.

4.3 Contrastación y validación del modelo

Una vez escogemos aquel cuyo valor de AIC es menor, se prosigue el análisis para su validación. A diferencia del análisis de regresión tradicional donde podemos interpretar los coeficientes de los modelos y aplicar la normalidad de los residuos, se ha tenido que utilizar otras técnicas estadísticas para poder contrastar y validar el modelo. En el presente proyecto se utiliza el test de Pearson, el test de devianza y la escala o parámetro de dispersión descritos a continuación.

4.3.1 Test χ^2 de Pearson

El test χ^2 de Pearson [Sa13] es una medida del error para medir la bondad de ajuste del modelo multinomial, en el presente proyecto el modelo binomial, que cuantifica la diferencia entre el valor observado y el valor estimado. Su estadístico es donde se recogen los residuos estandarizados o de Pearson del modelo logit, y es la diferencia entre el valor observado de la variable respuesta y el estimado dividido por la estimación de la desviación típica. Se define como:

$$\chi^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \frac{(Y_i - \hat{M}_i)^2}{\hat{M}_i(1 - \hat{M}_i)} \quad (4.3)$$

Podemos ver el hecho de ser el estadístico similar a la suma de cuadrados de los residuos en los modelos de regresión utilizados convencionalmente pero con la diferencia de no utilizar una distribución normal. Un buen ajuste del modelo será mejor cuando el valor del estadístico esté cerca de 0. Para ello se realiza el test de Pearson para contrastar la siguiente hipótesis nula:

$$H_0 = Y_i - \hat{M}_i = e_i = 0 \quad (4.4)$$

La cual se distribuye como una chi-cuadrado con (n-k) grados de libertad, por lo cual su valor se compara con el valor teórico de las tablas chi-cuadrado a la hora de contrastar la hipótesis nula. Si el valor calculado es superior al valor teórico se rechaza la hipótesis nula, es decir, su p-valor es significativo, el error será significativamente distinto de cero y por tanto, se trataría de un mal ajuste.

4.3.2 Test de devianza

El test de devianza corresponde al estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste de los modelos multinomiales, en nuestro caso el modelo logit binomial [Sa13]. El test está basado en la similitud de los datos agrupados. Se trata de la comparación por cociente de las frecuencias observadas y esperadas de cada clase que hay en los datos. Su estadístico es el siguiente:

$$G = 2 \left[\sum_{q=1}^Q \sum_{j=1}^k \frac{Y_j}{q} \ln \left(\frac{Y_j/q}{\hat{m}_j/q} \right) \right] \quad (4.5)$$

Es decir $G = -2 \ln \Lambda$. Donde y_j/q corresponden a los valores observados y \hat{m}_j/q a los esperados. Tiene como distribución asintótica la chi-cuadrado al igual que el test de Pearson para hacer el contraste, teniendo como grados de libertad $Q - (n+1) * (k-1)$, es decir, la diferencia entre la dimensión del espacio paramétrico y la dimensión de este espacio bajo hipótesis nula. La hipótesis nula del test viene a ser:

$$H_0 = G = -2 \ln \Lambda = 0 \quad (4.6)$$

Al igual del estadístico de Pearson, si el valor calculado es superior al valor teórico se rechaza la hipótesis nula de ser significativamente distinto de cero y por tanto, se trataría de un mal ajuste del modelo a los datos.

4.3.3 Escala o parámetro de dispersión

A la hora de saber si el modelo creado tiene sobredispersión de los datos, otra técnica utilizada es calcular la escala o parámetro de dispersión. Para ello se utilizan los valores resultantes de calcular el estadístico de Pearson y el valor residual de devianza que viene dada por defecto en el modelo logit.

Cuando calculamos es estadístico de Pearson, recordemos que se trata de la suma de residuos al cuadrado, pues bien, si utilizamos el valor resultante al estadístico de Pearson y lo dividimos por el valor residual de devianza, obtenemos la escala o parámetro de dispersión tal y como se observa en 4.7.

$$escala = \frac{\sum_{i=1}^n e_i^2}{Valor\ residuo\ devianza} \quad (4.7)$$

Si su valor es elevado nos indica la existencia de una elevada dispersión en los datos, en caso contrario estamos ante un buen ajuste del modelo a los datos y por tanto las variables que lo representan nos dan buena información sobre los datos a analizar.

4.4 Normalidad de los Residuos en el modelo logit

En los modelos logit con distribución binomial, a la hora de realizar los residuos, ya hemos visto dos test que nos proporcionan el buen ajuste del modelo además de calcular su escala, ambos test no tienen una distribución normal y por tanto son difíciles de analizar. En la regresión clásica utilizada habitualmente, se lleva a cabo el análisis de los residuos a través de la distribución normal. Por tanto, Al tener una variable endógena binomial con valores entre 0 y 1, los datos se mueven en el rango comprendido entre estos valores y la discrepancia entre valores observados y predichos por el modelo es generalmente grande. Por último analizar los residuos de forma tradicional no es a primera vista intuitivo a la hora de interpretarlos [Ca10]. Por ello no se utilizará esta técnica y se hará una alternativa más idónea como es la curva de ROC, dicha técnica se detallada en el siguiente apartado.

4.5 Curva de ROC

La curva de ROC (en inglés *Receiver Operating Characteristic*) permite evaluar la capacidad del modelo para discriminar. Gráficamente representa los efectos de la selección del umbral s , entre 0 y 1, con respecto a la sensibilidad y especificidad. Siendo s el valor resultante del área bajo la curva, sensibilidad la proporción de verdaderos valores 1 que han sido correctamente estimados y especificidad la proporción de verdaderos valores 0 que han sido correctamente predichos. El área bajo la curva ROC es la probabilidad de que una observación elegida al azar, tenga mayor probabilidad de ser ($Y=1$) frente a una observación ($Y=0$). Por lo tanto, lo idóneo es tener el mayor valor posible del área bajo la curva, es decir, su capacidad de discriminación. Tiene como rangos posibles entre 0.5 y 1, como se muestra más abajo, correspondiendo al primero el peor caso y al último el ideal. Un buen modelo ha de tener un área bajo la curva con valor mínimo a 0.7.

0.9 – 1	= ajuste excelente
0.8 – 0.9	= ajuste bueno
0.7 – 0.8	= ajuste medio
0.6 – 0.7	= ajuste malo
0.5 – 0.6	= ajuste muy malo

Gráficamente representa la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes puntos de corte. En la Fig. 4.1 se muestran varios ejemplos de curvas de ROC en el cual vemos mejor valor del área bajo la curva a más pronunciada y cercana está a la unidad.

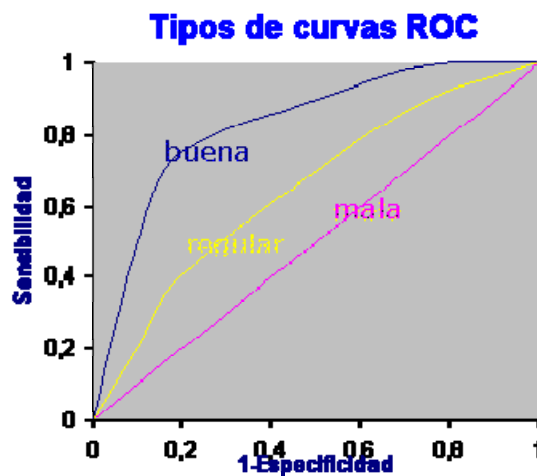


Fig. 4-1 Ejemplo curva de ROC

4.6 Predicciones con el modelo

Para realizar predicciones con los modelos logit, se puede hacer uso de tablas de contingencias. Anteriormente hemos hablado en la curva de ROC de sensibilidad y especificidad. Bien, se trata de valores de las tablas de contingencias que son utilizadas para ver la capacidad predictiva del modelo. No obstante, las realizadas en el presente proyecto tienen la finalidad de mejorar la capacidad predictiva de restaurar un fallo en la red y por tanto el orden de las variables en la tabla es distinto.

Tabla 4-1 Ejemplo de tabla de contingencias

Respuesta	Predicción	
	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	Verdaderos no restaurados ESPECIFICIDAD	Falsos no restaurados
$Y = 1$	Falsos restaurados	Verdaderos Restaurados SENSIBILIDAD

En la Tabla 4.1 vemos un ejemplo de cómo serán las tablas predictivas utilizadas. En ellas podemos ver la variable endógena como la respuesta Y con los dos posibles casos de ocurrencia y también tenemos la predicción como las Y^{\wedge} . Por tanto a la hora de ver la sensibilidad y especificidad del modelo vemos los verdaderos restaurados son la especificidad y los verdaderos no restaurados la sensibilidad.

A la hora de ver las proporciones de cada uno de los casos, los valores predictivos del modelo, en el caso de los no restaurados corresponde al cociente entre valor de los verdaderos no restaurados respecto a la suma de los verdaderos no restaurados y los falsos no restaurados. Se haría análogamente para los falsos no restaurados, así como para los valores predictivos del modelo de los verdaderos restaurados.

Capítulo 5. Análisis descriptivo

En el siguiente capítulo se va a llevar a cabo la exploración previa de los datos obtenidos para las redes BT, DT y TEL. Se han calculado los análisis descriptivos más relevantes de las diferentes redes, tanto para las diversas variables explicativas de las bases de datos como para aquellas variables destacables y con diferencias entre fallos en la red que han podido ser restaurados y los no restaurados. Para ello se han llevado a cabo diferentes técnicas estadísticas como resúmenes descriptivos, histogramas, cómputo de correlaciones, además de añadir el análisis de componentes principales (ACP). La finalidad es ver si el conjunto de variables explicativas es lo suficientemente diverso como para resultar de utilidad su inclusión en el posterior capítulo de modelización.

5.1 Descriptiva de las variables

En este apartado vamos a realizar el análisis descriptivo de los datos. Primeramente en la tabla 5-1 se visualiza un resumen de las diferentes redes con el total de observaciones, el número de fallos no restaurados y el número de fallos restaurados, es decir, las dos opciones de la variable endógena. La red TEL tiene un número significativamente mayor de datos frente a las otras dos redes. La proporción de observaciones dónde el fallo no ha podido ser restaurado ($Y = 0$) en TEL es del 18% frente al 10% aproximado en las otras dos, esto puede ser indicador inicial de diferencia entre las redes.

Tabla 5-1 Resumen de las redes

Red	Nº observaciones	Restauradas ($Y = 1$)	No restauradas ($Y = 0$)
BT modelo	1082	969	113
DT modelo	1976	1767	209
TEL modelo	5379	4380	979

A continuación, en la tabla 5-2 vemos la descriptiva de las variables explicativas, los valores máximos, mínimos, la media y mediana de cada una de ellas, además del primer y tercer cuartil. No hay valores destacables en las descriptivas, el grado nodal medio de la red, avgDegree, obtiene un total de tres valores, el mismo valor en cada red, el radiusVar se mueve entre 1 y 1.5, algConnectVar y spectralRadiusVar están entre 0.5 y 1, además X.Grup1, X.Grup2 y X.Grup3, obviamente, tienen mínimo 0 y máximo 1 al tratarse del porcentaje de las variables originales Grup1, Grup2 y Grup3 respectivamente. spectralRadiusF y diameterVar tienen los mínimos y máximos similares comprendidos entre 0.90 y 2.5. El resto de variables tienen valores mínimos y máximos diferentes.

Tabla 5-2 Resumen descriptivo de las variables explicativas

avgDegree	avgLbetF	algConnectF	spectralRadiusF	radiusVar
Min. :3.180	Min. :0.06000	Min. :0.1300	Min. :0.94	Min. :1.00
1st Qu.:3.330	1st Qu.:0.06000	1st Qu.:0.2400	1st Qu.:0.97	1st Qu.:1.00
Median :3.730	Median :0.06000	Median :0.2500	Median :0.99	Median :1.00
Mean :3.565	Mean :0.07707	Mean :0.2888	Mean :1.33	Mean :1.07
3rd Qu.:3.730	3rd Qu.:0.09000	3rd Qu.:0.2600	3rd Qu.:1.46	3rd Qu.:1.00
Max. :3.730	Max. :0.13000	Max. :0.4900	Max. :2.42	Max. :1.50
diameterVar	algConnectVar	spectralRadiusVar	link_node_A	link_node_B
Min. :1.500	Min. :0.5400	Min. :0.6900	Min. : 0.000	Min. : 1.00
1st Qu.:1.750	1st Qu.:0.9200	1st Qu.:0.9500	1st Qu.: 3.000	1st Qu.: 7.00
Median :1.750	Median :0.9700	Median :0.9800	Median : 8.000	Median :12.00
Mean :1.852	Mean :0.9374	Mean :0.9615	Mean : 9.704	Mean :13.75
3rd Qu.:2.000	3rd Qu.:0.9900	3rd Qu.:1.0000	3rd Qu.:16.000	3rd Qu.:20.00
Max. :2.500	Max. :1.0000	Max. :1.0000	Max. :27.000	Max. :29.00
Grup1	Grup2	Grup3	X.Grup1	X.Grup2
Min. : 0.000	Min. : 0.00	Min. : 0.00	Min. :0.00000	Min. :0.0000
1st Qu.: 0.000	1st Qu.: 9.00	1st Qu.: 1.00	1st Qu.:0.00000	1st Qu.:0.4400
Median : 1.000	Median :14.00	Median : 7.00	Median :0.05000	Median :0.5800
Mean : 1.591	Mean :14.42	Mean :10.17	Mean :0.08809	Mean :0.6025
3rd Qu.: 2.000	3rd Qu.:19.00	3rd Qu.:16.00	3rd Qu.:0.11000	3rd Qu.:0.7800
Max. :12.000	Max. :52.00	Max. :66.00	Max. :1.00000	Max. :1.0000
X.Grup3	mean_path_length_before	min_path_length_before	max_path_length_before	
Min. :0.0000	Min. :1.000e+00	Min. :1.000	Min. : 1.000	
1st Qu.:0.0600	1st Qu.:3.119e+03	1st Qu.:1.000	1st Qu.: 5.000	
Median :0.3300	Median :2.725e+09	Median :1.000	Median : 6.000	
Mean :0.3098	Mean :1.824e+10	Mean :1.341	Mean : 6.169	
3rd Qu.:0.5000	3rd Qu.:4.048e+10	3rd Qu.:2.000	3rd Qu.: 7.000	
Max. :1.0000	Max. :6.429e+10	Max. :9.000	Max. :15.000	
mean_bw_path	min_bw_path	max_bw_path	restored.	
Min. :4.000e+01	Min. : 40.00	Min. : 40	Min. :0.0000	
1st Qu.:3.540e+02	1st Qu.: 40.00	1st Qu.: 800	1st Qu.:1.0000	
Median :2.541e+09	Median : 40.00	Median :1600	Median :1.0000	
Mean :1.622e+12	Mean : 63.18	Mean :1850	Mean :0.8454	
3rd Qu.:3.263e+12	3rd Qu.: 80.00	3rd Qu.:2800	3rd Qu.:1.0000	
Max. :1.693e+13	Max. :3600.00	Max. :6000	Max. :1.0000	

Seguimos con los histogramas a fin de ver visualmente el comportamiento de las variables explicativas. La Fig. 5-1 muestra los gráficos donde vemos, tal y como se comentó anteriormente, el grado nodal medio siempre el mismo valor numérico para cada red y, por consiguiente, muestra sus tres valores. La conectividad algebraica (algConnectVar) tiene tendencia a ser mayor cuando aumenta su valor, sus valores están comprendidos entre 0.5 y 1, los nodos A y B son diferentes, nos sugieren el hecho de no estar relacionados entre sí en los fallos. El valor de la longitud de la ruta varía, entre el valor de la ruta más corta (min_path_length_before) y el valor de la ruta más larga (max_path_length_before) hay una gran diferencia, en el primer caso hay mayor concentración de observaciones entre 0 y 2 y en la ruta más larga, las observaciones están más dispersas y con valores comprendidos entre 2 y 8.

Los Grup1, Grup2 y Grup3 tienen comportamientos diferentes, sugiriendo heterogeneidad en las conexiones establecidas. Además, para la variable bw_path explicada en el capítulo 3, vemos sus tres gráficos del mínimo, media y máximo respectivamente, con diferente comportamiento. Hasta ahora podemos ver diversidad en los datos.

Continuamos con las correlaciones dos a dos de las variables a fin de saber si hay correlación entre ellas. En la Tabla 5-3 vemos el ranking de correlaciones donde se incluyen aquellas con valor superior a 0.70. Las correlaciones superiores a 0.85 corresponden a la relación del spectralRadiusF con avgLbetF y algConnectF, estas correlaciones tienen sentido al tratarse de variables de la topología y por tanto están relacionadas entre sí. Una correlación de 0.84 entre link_node_A y link_node_B, no es una correlación superior a 0.90 y por tanto no tenemos una excesiva correlación entre los dos links. El resto de correlaciones tienen valores menores o igual a 0.80 informando de no tener una correlación alta. En definitiva, tenemos una buena dispersión de los datos sin correlaciones elevadas entre las variables.

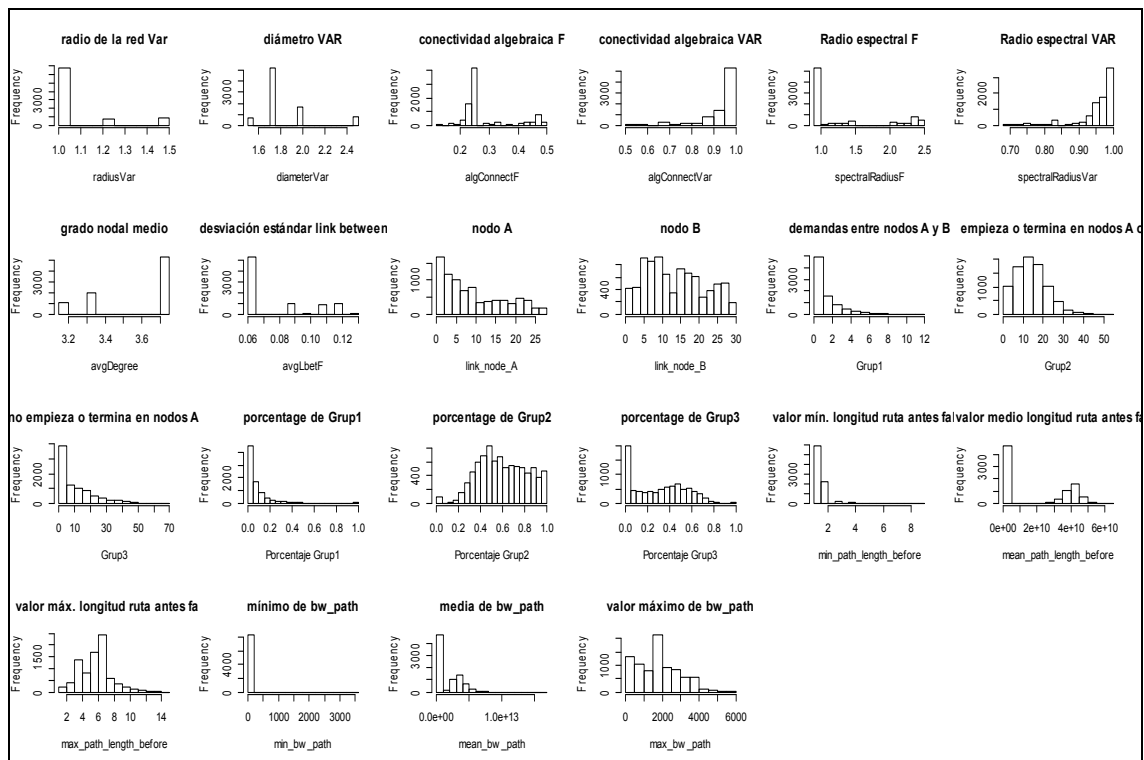


Fig. 5-1 Histogramas variables explicativas

Tabla 5-3 Ranking de correlaciones

avgLbetF	spectralRadiusF	0.96
algConnectF	spectralRadiusF	0.87
link_node_A	link_node_B	0.84
Grup3	X.Grup3	0.8
max_bw_path	max_path_length_before	0.79
mean_bw_path	mean_path_length_before	0.78
radiusVar	DiameterVar	0.76
diameterVar	spectralRadiusF	0.73

Los datos nos sugieren una diversidad entre ellos, para saber si las variables escogidas tienen un comportamiento diverso, se ha llevado a cabo otra técnica estadística como es el análisis de componentes principales (ACP) explicado en el siguiente apartado.

5.2 Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) tiene como finalidad reducir la dimensionalidad de un conjunto de datos, es decir, reducir el número de variables perdiendo la menor información posible. Para ello crea una transformación lineal a través de un sistema de coordenadas del conjunto original de datos donde la varianza con mayor valor es la

denominada primera componente principal, el segundo valor de la varianza en los datos con mayor valor es la segunda componente principal, y así sucesivamente. Cada nueva componente principal es una combinación lineal de las variables originales, siendo estas independientes entre sí. Normalmente se escogen dos o tres componentes principales [Pe02]. Para tener la transformación lineal de los datos, primeramente se hace la matriz de covarianza o matriz de coeficientes de correlación. Esta transformación cambia las antiguas coordenadas a las nuevas coordenadas, y es la denominada transformación lineal, a partir de la cual se puede reducir la dimensionalidad de datos.

En la Fig. 5.2, se muestra el ACP de los datos, que muestra una clara diversidad entre las variables explicativas. Grup2 con respecto a su porcentaje, la variable X.Grup2, están en direcciones distintas, intuitivamente, al ser el porcentaje creado respecto a la anterior se podría pensar en una alta similitud entre ellas, pero tanto en el ranking de correlaciones como en el ACP vemos lo contrario. Lo mismo sucede entre Grup1 y Grup3 respecto a sus variables de porcentaje creadas a partir de ellas. En general, se observa una gran dispersión de las observaciones en sus componentes principales, lo cual nos indica que estamos contemplando una amplia variedad de casos en nuestro conjunto de datos.

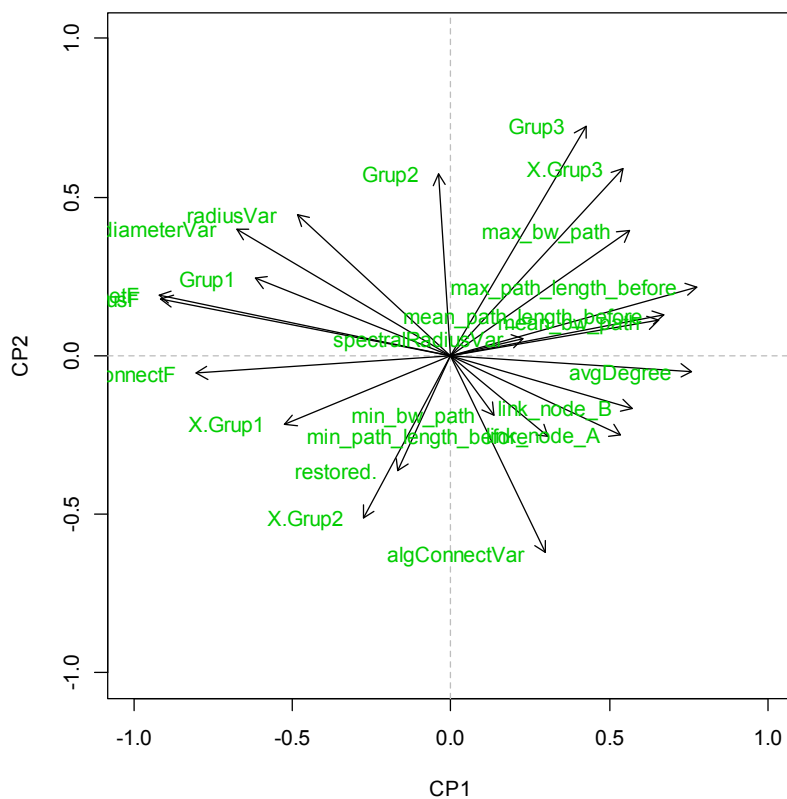


Fig. 5-2 Análisis de Componentes Principales

5.3 Análisis para restaurados y no restaurados por separado

La finalidad del presente proyecto es mejorar la probabilidad de restauración cuando se ha producido un fallo en la red, por tanto, se ha realizado una breve descriptiva con algunas variables para ver si hay diferencias en la misma variable cuando la variable respuesta presenta valores Y=1 o Y=0. En la tabla 5-4 podemos ver la descriptiva para 6 variables, primeramente para todas las observaciones y luego para los casos donde la variable endógena tiene valor igual a 1, finalizando con las que tienen valor igual a 0.

Tabla 5-4 Resumen descriptivo de las variables separadas en Y = 1 y Y = 0

Todos los datos			
algConnectVar	link_node_B	X.Grup2	X.Grup3
Min. :0.5400	Min. : 1.00	Min. :0.0000	Min. :0.0000
1st Qu.:0.9200	1st Qu.: 7.00	1st Qu.:0.4400	1st Qu.:0.0600
Median :0.9700	Median :12.00	Median :0.5800	Median :0.3300
Mean :0.9374	Mean :13.75	Mean :0.6025	Mean :0.3098
3rd Qu.:0.9900	3rd Qu.:20.00	3rd Qu.:0.7800	3rd Qu.:0.5000
Max. :1.0000	Max. :29.00	Max. :1.0000	Max. :1.0000
mean_bw_path	max_bw_path	restored.	
Min. :4.000e+01	Min. : 40	Min. :0.0000	
1st Qu.:2.271e+02	1st Qu.: 800	1st Qu.:1.0000	
Median :3.041e+02	Median :1600	Median :1.0000	
Mean :3.279e+02	Mean :1850	Mean :0.8454	
3rd Qu.:3.921e+02	3rd Qu.:2800	3rd Qu.:1.0000	
Max. :3.600e+03	Max. :6000	Max. :1.0000	
Para Y = 1			
algConnectVar	link_node_B	X.Grup2	X.Grup3
Min. :0.5400	Min. : 1.00	Min. :0.0000	Min. :0.0000
1st Qu.:0.9300	1st Qu.: 7.00	1st Qu.:0.4500	1st Qu.:0.0400
Median :0.9700	Median :13.00	Median :0.6000	Median :0.3000
Mean :0.9455	Mean :13.77	Mean :0.6144	Mean :0.2911
3rd Qu.:1.0000	3rd Qu.:20.00	3rd Qu.:0.7900	3rd Qu.:0.4900
Max. :1.0000	Max. :29.00	Max. :1.0000	Max. :1.0000
mean_bw_path	max_bw_path		
Min. :4.000e+01	Min. : 40		
1st Qu.:1.216e+02	1st Qu.: 700		
Median :3.041e+02	Median :1600		
Mean :3.116e+02	Mean :1710		
3rd Qu.:3.731e+02	3rd Qu.:2400		
Max. :3.600e+03	Max. :6000		
Para Y = 0			
algConnectVar	link_node_B	X.Grup2	X.Grup3
Min. :0.5400	Min. : 1.00	Min. :0.0000	Min. :0.0000
1st Qu.:0.8600	1st Qu.: 8.00	1st Qu.:0.3700	1st Qu.:0.2300
Median :0.9500	Median :11.00	Median :0.5000	Median :0.4600
Mean :0.8935	Mean :13.69	Mean :0.5376	Mean :0.4117
3rd Qu.:0.9900	3rd Qu.:20.00	3rd Qu.:0.6800	3rd Qu.:0.6000
Max. :1.0000	Max. :29.00	Max. :1.0000	Max. :1.0000
mean_bw_path	max_bw_path		
Min. :1.389e+02	Min. : 300		
1st Qu.:3.149e+02	1st Qu.:2000		
Median :3.832e+02	Median :2400		
Mean :4.174e+02	Mean :2620		
3rd Qu.:4.719e+02	3rd Qu.:3200		
Max. :2.800e+03	Max. :6000		

Para algConnectVar vemos la misma proporción entre el mínimo y el máximo, pero su comportamiento es diferente al tener un primer cuartil en 0.93 para Y=1 y 0.86 en su contrario (Y=0). El link_node_B y X.Grup2 les pasa lo mismo, los valores del primer cuartil son menores en el caso de Y=0 con respecto a Y=1. En el resto de variables, los primeros cuartiles de las observaciones con la variable endógena igual a 1 son superiores, es decir, pasa a la inversa. Cabe destacar el hecho de que la variable mean_bw_path tiene un valor considerablemente mayor en la media cuando Y=0 respecto de Y=1.

Para poder visualizar mejor estas diferencias explicadas hasta ahora, en la Fig. 5-3 se ven los comportamientos desiguales en las variables cuando las observaciones de la variable respuesta tienen valores igual a 1 frente a las contrarias. En todas las variables hay diferencias, sobre todo para el porcentaje de Grup3. Para mean_bw_path en los fallos en la red que no han podido ser restaurados frente a los contrarios, vemos una asimetría más pronunciada en el segundo caso respecto al primero. Finalmente, max_bw_path tiene comportamientos diferentes en ambos casos, para Y=1 tiene una forma más asimétrica y en Y=0 el volumen de observaciones están más desplazadas y comprendidas entre los valores 1000 y 4000.

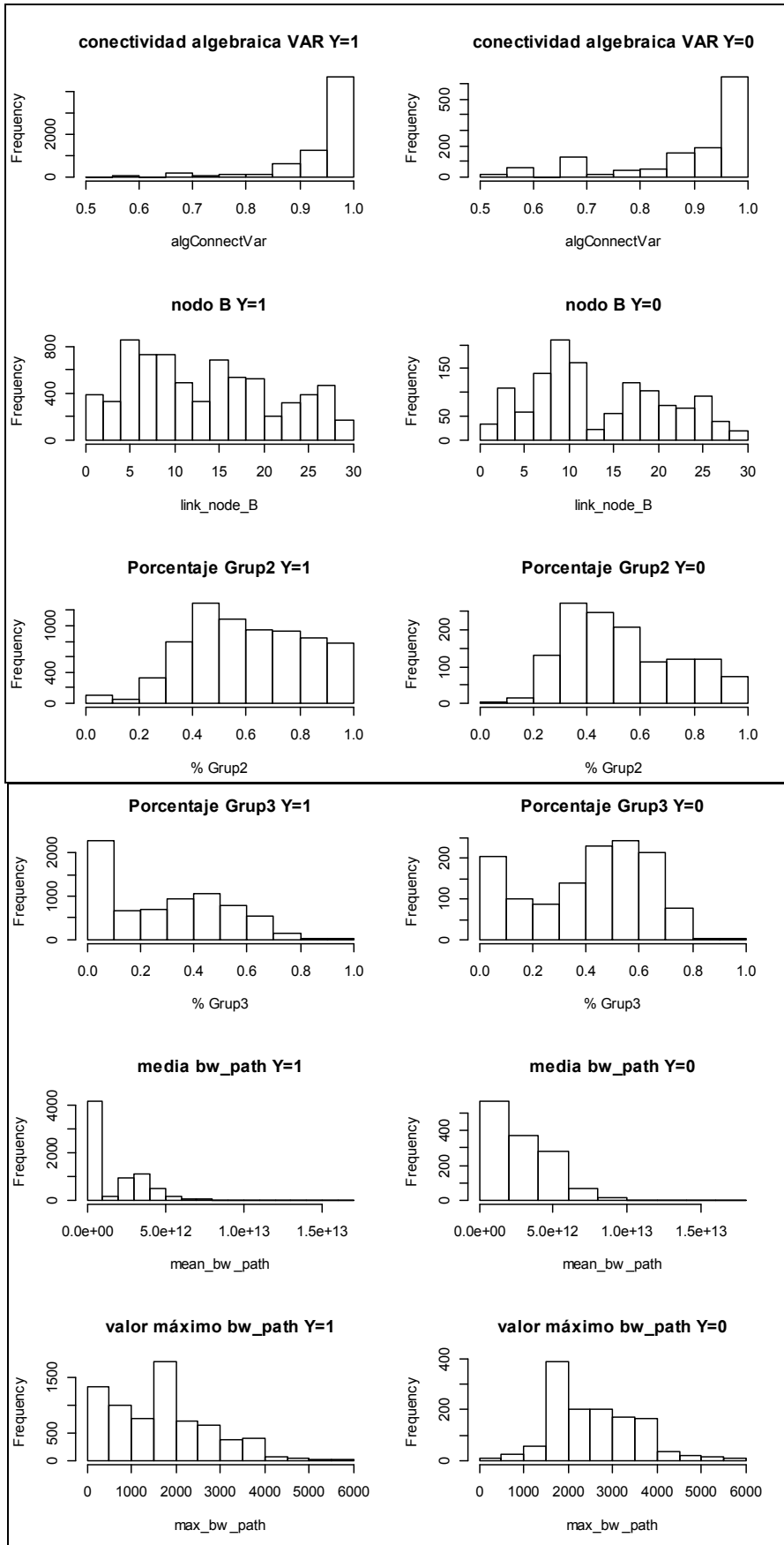


Fig. 5-3 Histogramas de una selección de variables explicativas

5.4 Resumen

En este capítulo hemos visto, por una parte, que el conjunto de variables explicativas seleccionadas es diverso y, por otra, que existen diferencias significativas entre los valores de dichas variables explicativas entre las observaciones que pertenecen a los distintos niveles de la variable respuesta. Estas dos características nos incrementan las opciones de que algún subconjunto de estas variables forme un buen predictor de la variable respuesta, lo cual es esencial para la obtención de un modelo válido.

Tras este análisis exploratorio que ha servido para validar el conjunto de datos obtenidos, en el siguiente capítulo atacaremos la modelización de la variable respuesta.

Capítulo 6. Estimación de los modelos

En este capítulo se presentaran los modelos para las redes BT, DT y TEL, presentadas en el orden en cuyos datos fueron obtenidos. Partiendo de la red BT se presenta el mejor modelo con un número reducido de variables. Además de esta aproximación, se planteará la segmentación de la variable respuesta en categorías con el fin de encontrar modelo apropiado para cada categoría, así como un único modelo con el mismo conjunto de variables pero distintos coeficientes por categorías. De las conclusiones sacadas tras el análisis de los estadísticos de decisión, se seleccionarán aquellos modelos con un mejor ajuste y se extenderán a las redes DT, y TEL. Al final del proceso se seleccionará aquel modelo que presente un mejor balance entre calidad y simplicidad, promoviendo aquellas opciones que permitan comparar distintas redes entre sí mediante modelos similares. Finalmente, se evaluará el poder predictivo de los modelos seleccionados en comparación con el modelo nulo a través del análisis de las tablas de contingencia.

6.1 Modelos de la red BT

6.1.1 Modelo único

A la hora de buscar los modelos para la red BT, se ha decidido buscar un modelo con 4 o 5 variables como máximo. Se calcula la búsqueda del mejor modelo para representar los datos con todas las observaciones, para ello se calcula el modelo nulo, seguidamente se realizan los modelos con una, dos, tres, cuatro y cinco variables escogiendo cada una de ellas sobre el modelo con menor AIC, es decir, para el modelo con una variable se hizo las combinaciones posibles y se seleccionó el modelo con AIC más bajo, así para el de dos variables y también para el de tres, cuatro y cinco variables. Una vez tenemos cada modelo se interpretan los signos de los coeficientes del modelo, el test chi-cuadrado de Pearson, el test de la devianza y, se calcula el parámetro de dispersión. Una vez lo tenemos, se hace la comparación a través del anova entre los modelos con 4 y 5 variables a fin de saber cuál es el mejor y se finaliza con la curva de ROC del modelo escogido para ver el ajuste del modelo.

Empezamos el análisis con el modelo nulo. Recordemos que al tener la variable respuesta como una binomial, el modelo calculado es el modelo logit. El resultado se muestra a continuación:

```

glm(formula = restored. ~ 0, family = binomial(link = "logit"),
    data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.177   1.177   1.177   1.177   1.177

No Coefficients
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1500  on 1082  degrees of freedom
Residual deviance: 1500  on 1082  degrees of freedom
AIC: 1500
Number of Fisher Scoring iterations: 0

```

El AIC es de 1500. El test de Pearson de 0.49 siendo no significativo, por lo tanto indica el no rechazo de la hipótesis nula de tener un buen ajuste, el test de devianza da significativo con valor $4.44e-16$, por lo tanto se rechaza la hipótesis nula de ser un modelo ajustado a los datos y la escala o parámetro de dispersión es igual a 1. Al no tener variables el modelo nulo, se prosigue el análisis con el modelo de una variable. Para calcularlo, se realizaron todos los modelos con una variable posibles a fin de obtener aquel cuyo valor del AIC era menor. El modelo resultante es el mostrado seguidamente.

```

glm(formula = restored. ~ 0 + algConnectF, family = binomial,
    data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2773   0.3943   0.3943   0.4342   0.6918

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
algConnectF  10.0611    0.4672   21.54  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1499.97  on 1082  degrees of freedom
Residual deviance:  660.02  on 1081  degrees of freedom
AIC: 662.02
Number of Fisher Scoring iterations: 5

```

La variable algConnectF tiene un AIC de 662.02, el test de Pearson es de 0.99, y el test de devianza igual a 1, ambos no significativos, en el capítulo 5 comenté si el p-valor tanto en el test de Pearson como en el de devianza era significativo había un mal ajuste del modelo a los datos, por consiguiente, estos resultados nos indican un buen ajuste del modelo a los datos. El parámetro de dispersión es de 0.87 y el coeficiente de la variable es significativo y positivo. A partir de este modelo, se ejecutan todos los posibles modelos con dos variables, siendo el de AIC más pequeño el siguiente:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path, family =
binomial,
    data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2094   0.0859   0.2001   0.4147   1.7229

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
algConnectF  25.488741   1.875295   13.59  <2e-16 ***
max_bw_path  -0.001545   0.000152  -10.16  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1499.97  on 1082  degrees of freedom
Residual deviance:  493.59  on 1080  degrees of freedom
AIC: 497.59
Number of Fisher Scoring iterations: 6

```


La variable introducida es max_bw_path con un coeficiente negativo y significativo, el valor del AIC de 497.59 menor al del modelo anterior, el test de Pearson de 0.99 y el test de devianza igual a 1, al ser no significativos informan del buen ajuste del modelo a los datos. Además el parámetro de dispersión es 0.90, menor al modelo anterior.

Se sigue el análisis, realizando la búsqueda de otra variable más al modelo, recordemos la finalidad de encontrar un modelo con 4 o 5 variables. El modelo con tres variables y con menor valor de AIC incluye como tercera variable mean_bw_path con coeficiente negativo y además significativo, el AIC es de 494.46, el test de Pearson de 0.99 y el test de devianza 1, tal y como se ha comentado en los modelos anteriores, indican los dos test un buen ajuste del modelo a los datos. El parámetro de dispersión es de 0.87, un valor bajo por lo tanto no hay una elevada sobredispersión en los datos. Los coeficientes del modelo, siguen con los mismos signos las variables del modelo anterior y la añadida tiene signo negativo. El modelo se muestra seguidamente:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path,
    family = binomial, data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2037  0.0834  0.1912  0.4117  1.6723

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  26.9409423  2.0219794  13.324  <2e-16 ***
max_bw_path  -0.0014220  0.0001604  -8.867  <2e-16 ***
mean_bw_path -0.0014974  0.0006166  -2.429  0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1499.97  on 1082  degrees of freedom
Residual deviance:  488.46  on 1079  degrees of freedom
AIC: 494.46

Number of Fisher Scoring iterations: 7

```

Al considerar tres variables aun pocas, se realizó la búsqueda de una cuarta variable. Más abajo vemos el modelo con 4 variables. Los resultados para el test de la devianza es 1 y para el test de Pearson 0.99. El AIC es de 492.25 y la escala igual a 0.92, la sobredispersión entre los datos no es elevada. Tal y como vimos en los modelos anteriores, resultados superiores al nivel de significación, el modelo se ajusta a los datos.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
    X.Grup3, family = binomial, data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2585  0.0815  0.1863  0.4091  1.6950

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  27.6790044  2.0762101  13.332  < 2e-16 ***
max_bw_path  -0.0013180  0.0001678  -7.853  4.05e-15 ***
mean_bw_path -0.0015598  0.0006483  -2.406  0.0161 *
X.Grup3      -0.9942129  0.4886065  -2.035  0.0419 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1499.97  on 1082  degrees of freedom
Residual deviance:  484.25  on 1078  degrees of freedom
AIC: 492.25

Number of Fisher Scoring iterations: 7

```

Se ha de tener en cuenta los signos de los coeficientes por ser los valores del estadístico de Wald y no poderse interpretar como en los modelos lineales utilizados normalmente. Los signos de los coeficientes de las tres primeras variables se mantienen igual al modelo visto anteriormente, la cuarta variable corresponde a X.Grup3, tiene un coeficiente negativo y es significativa. Finalmente, el modelo con 5 variables siguiendo su elección como en los modelos anteriores, tiene el AIC igual a 489.67 tal y como se ve abajo.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3 + min_bw_path, family = binomial, data = x300_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3850  0.0793  0.1849  0.3983  1.6674

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  28.2119155  2.1418059  13.172 < 2e-16 ***
max_bw_path  -0.0011811  0.0001788  -6.606 3.95e-11 ***
mean_bw_path -0.0028598  0.0008818  -3.243  0.00118 **
X.Grup3      -1.2597775  0.5090587  -2.475  0.01333 *
min_bw_path   0.0029432  0.0018971   1.551  0.12080
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1499.97  on 1082  degrees of freedom
Residual deviance:  479.67  on 1077  degrees of freedom
AIC: 489.67

Number of Fisher Scoring iterations: 7

```

Los coeficientes de las cuatro variables vistas en el modelo anterior mantienen los mismos signos, la última, min_bw_path, es positiva y significativa. El modelo da importancia a las tres variables creadas a partir de bw_path lo cual nos informa que tanto el ancho como el peso nos da información. El test de la devianza es de 1 y el estadístico de Pearson 0.25 informan de bondad de ajuste del modelo a los datos, aunque el test de Pearson acerca su valor a la significación notablemente con respecto al modelo anterior, el parámetro de sobredispersión es de 1.03, un poco mayor al modelo anterior.

Una vez se tienen todos los modelos con los resultados de los valores del AIC, los signos de los coeficientes, el test de Pearson y el test de la devianza para cada uno de ellos, la tabla 6-1 proporciona un resumen de los resultados de los modelos estudiados en el apartado anterior. En ella se puede ver la similitud en los test entre los modelos con una, dos y tres variables, los otros son distintos. El AIC es menor y el parámetro de dispersión o escala, salvo para el modelo nulo y con una variable, a partir del modelo con tres variables aumenta.

Tabla 6-1 Tabla resumen de los modelos para la red BT

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión/escala
Modelo nulo	1500	0.49	4.44e-16	1
Modelo con 1 variables	662.02	0.99	1	0.90
Modelo con 2 variables	497.59	0.99	1	0.87
Modelo con 3 variables	494.1	0.99	1	0.87
Modelo con 4 variables	492.25	0.97	1	0.92
Modelo con 5 variables	489.67	0.25	1	1.03

Esta información, no es suficiente para elegir el mejor modelo ; por ello, se realizó el análisis de la devianza, es decir, la anova para contrastar los modelos de 4 y 5 variables a fin de saber el más idóneo a los datos. El resultado nos indica un p-valor significativo.

Analysis of Deviance Table					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3 + min_bw_path					
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	1078	484.25			
2	1077	479.67	1	4.5865	0.03222 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Por consiguiente, la hipótesis nula de ser el modelo con 5 variables mejor respecto al modelo con 4 variables con un p-valor significativo, se rechaza. Ello y juntamente con unos resultados de AIC, el test de devianza similares entre los dos modelos, pero con el test de Pearson más cercano a la significación y la escala mayor al modelo con 4 variables, nos confirma el modelo único es el de 4 variables. Seguidamente, vamos a interpretar los coeficientes del modelo en términos de odds ratio. Para interpretar los coeficientes, recordar si un valor es positivo o negativo tiene una repercusión diferente al modelo.

Un valor positivo de 1.05e+12 en la variable algConnectF, al ser superior a 1, nos indica para la conectividad algebraicaF de poder ser restaurada, si incrementa una unidad la variable, esta tiene un incremento de (1-1.05e+12)% para un mismo valor de la variable algConnectF y tendrá un efecto significativo sobre la variable respuesta. Es decir, al ser un valor bastante mayor a 1, cualquier incremento en la conectividad algebraicaF tendrá un efecto significativo sobre la variable respuesta mejorando la restauración del fallo producido. Las otras tres variables tienen un coeficiente negativo, el odds ratio es de 0.99 para max_bw_path y mean_bw_path y 0.37 para X.Grup3. Los odds inferiores a 1 nos informan para cada unidad de incremento en cada una de las tres variables, ésta disminuye en 0.01% y 0.63 respectivamente, es decir, es decir, esta nos aleja de obtener una probabilidad cercana a la unidad en la variable respuesta (Y=1), es decir, nos dificulta la restauración del fallo. Los cálculos de los odds ratio se muestran seguidamente.

Odds Ratio de los coeficientes:	
algConnectF	$\exp(27.6790044) = 1.05e+12$
max_bw_path	$\exp(-0.0013180) = 0.99$
mean_bw_path	$\exp(-0.0015598) = 0.99$
X.Grup3	$\exp(-0.9942129) = 0.37$

La escala, en unidades, de las cuatro variables del modelo no son las mismas, en el capítulo 3 se comentó el hecho de hacer la variable bw_path y a partir de ella, dos de las tres variables creadas mean_bw_path y max_bw_path se incluyen en el modelo único, su magnitud en unidades, para comparar en la misma escala, tendría que multiplicarse por 1000 el valor del coeficiente y luego aplicar la exponencial, así se obtendría el odds ratio en la misma magnitud de algConnectF. El resultado sería max_bw_path $\exp(-0.0013180*1000) = 0.27$ y para mean_bw_path $\exp(-0.0015598*1000) = 0.21$. Igualmente son valores menores a la unidad y por tanto para cada unidad de incremento en cada una de las variables, ésta disminuye en 0.73% y 0.79 respectivamente sobre la variable respuesta. Del mismo modo, X.Grup3 se trata del porcentaje de la variable Grup3 y por tanto se podría multiplicar por 100 para tener su valor en unidades. El resultado es X.Grup3 $\exp(-0.9942129*100) = 6.64e-44$, con valor prácticamente 0, entonces una unidad de incremento en X.Grup3 no influye en la variable respuesta al ser el odds ratio igual a 1.

Dados los resultados de ser mejor modelo el de cuatro variables, se ha calculado la curva de ROC para el modelo original en la Fig. 6.1. En ella se observa la curva de ROC con valor a 0.88 informando de ser un buen ajuste al modelo. Un valor en la curva de ROC entre 0.8 y 0.9 es un buen ajuste del modelo tal y como vimos en el capítulo 4. Por lo tanto, el modelo seleccionado es un buen candidato para representar los datos de la red BT y será utilizado en el análisis de las otras redes a fin de saber si distintas redes se pueden modelar con las mismas variables.

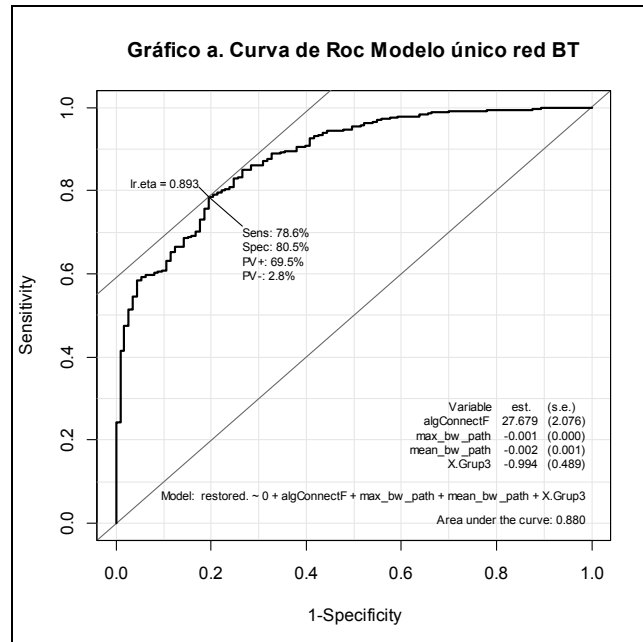


Fig.6-1 Curva de ROC para el modelo único de la red BT

6.1.2 Modelos por categorías

A la hora de encontrar el modelo de una red y así ver sus características, se demostró diferencia entre las observaciones cuando la variable respuesta es igual a 1 o 0. Por ello, se ha creado una variable que permita separar las observaciones por categorías. Después de probar varias tentativas, se ha optado por crear la variable $\log.bw_path_demand$ como función de algunas de las variables explicativas:

$$\log.bw_path_demand = \frac{\log(mean_bw_path \cdot Total)}{algConnectVar} \quad (6.1)$$

Como se puede ver, utiliza el logaritmo de la variable $mean_bw_path$ multiplicado por el total y lo divide por $algConnectVar$. Se ha decidido hacer 6 categorías con una proporción similar entre las categorías 2 y 5. La primera categoría corresponde a la mayoría de datos donde la variable respuesta tiene valores igual a 1 y para la última pasa lo contrario, las observaciones serán prácticamente todas igual a 0. Los puntos de corte de cada categoría para la red BT se muestran en la Tabla 6-2.

Tabla 6-2 Puntos de corte de las categorías

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Red BT	0 – 3.49	3.50 – 3.81	3.82 – 4.13	4.14 – 4.62	4.63 – 7.99	> 8.00

En la Fig. 6-2 podemos ver el número de fallos no restaurados para la red BT según las categorías creadas. En este caso vemos una evolución a medida que aumenta la categoría cuando Y=0. No tiene sentido modelizar las categorías primera y última al no darnos apenas información cuando ocurren los dos sucesos por este motivo se calculará el análisis para las categorías 2, 3, 4 y 5.

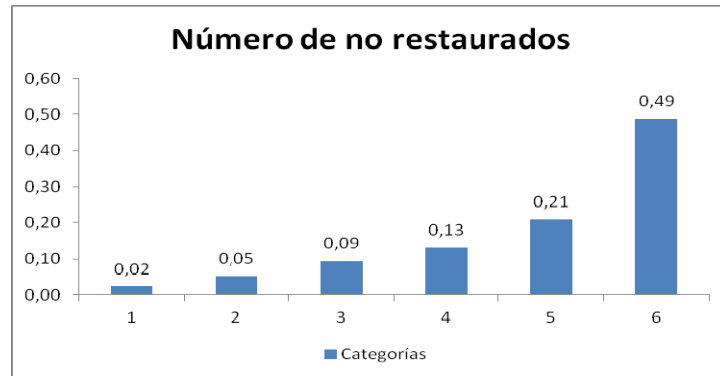


Fig. 6-2 Número de no restaurados para las distintas categorías

6.1.3 Modelo único por categorías para la red BT

En el siguiente apartado, se aplica el modelo único del apartado 6.1.1 a cada una de las categorías, de la segunda a la quinta de la red BT. Para la categoría 2, el modelo es el siguiente:

```
glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x300_1_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.96241  0.08915  0.13775  0.21523  0.68405

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  28.033462   7.242800   3.871 0.000109 ***
max_bw_path  -0.001542   0.000701  -2.199 0.027845 *
mean_bw_path  0.001257   0.002921   0.430 0.666887
X.Grup3      -3.839722   2.861382  -1.342 0.179625
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 270.327  on 195  degrees of freedom
Residual deviance:  38.741  on 191  degrees of freedom
AIC: 46.741

Number of Fisher Scoring iterations: 7
```

El AIC es de 46.741, tiene un total de 195 observaciones, al estar en una categoría, el tamaño de las observaciones es menor y por eso el AIC también, el test de Pearson es de 0.99 y del de devianza 1 indicando un buen ajuste del modelo. El parámetro de dispersión es de 0.75 no teniendo una sobredispersión elevada en los datos. Los coeficientes de las variables para algConnectF y mean_bw_path es positivo y para las otras dos negativo, además las variables significativas son algConnectF y max_bw_path y las otras no.

Para la tercera categoría, aplicamos el modelo único para los datos de dicha categoría. Presenta un AIC de 59.65, el test de Pearson 0.99, el test de devianza igual a 1, el número de observaciones es de 191 y la escala es de 0.67, menor a la categoría anterior. Los coeficientes de los signos en este caso, algConnectF y X.Grup3 tienen signo positivo. Menos la variable

algConnectF todas son no significativas en los coeficientes. El modelo se muestra seguidamente:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x300_l_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6460  0.1270  0.1932  0.2864  0.9197

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF 25.3967465  7.2869271  3.485 0.000492 ***
max_bw_path -0.0010539  0.0006412 -1.644 0.100256
mean_bw_path -0.0017334  0.0014701 -1.179 0.238355
X.Grup3      1.2004227  2.2230160  0.540 0.589199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 264.78  on 191  degrees of freedom
Residual deviance:  51.65  on 187  degrees of freedom
AIC: 59.65

Number of Fisher Scoring iterations: 7

```

Para la siguiente categoría, la cuarta, el modelo único seleccionado para todos los datos de la red BT del apartado 6.1.1 aplicado a los datos de la categoría, se puede observar más abajo.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x300_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4556  0.2766  0.4036  0.5373  1.1668

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF 17.0259366  4.7832682  3.559 0.000372 ***
max_bw_path -0.0008154  0.0003334 -2.446 0.014449 *
mean_bw_path -0.0020008  0.0019311 -1.036 0.300153
X.Grup3      1.8446569  1.4121045  1.306 0.191445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 270.33  on 195  degrees of freedom
Residual deviance: 126.95  on 191  degrees of freedom
AIC: 134.95

Number of Fisher Scoring iterations: 5

```

Los signos de los coeficientes de las variables varían de las otras categorías, tienen signo positivo en algConnectF y X.Grup3 y negativo en el resto, también observamos algConnectF y max_bw_path son significativos. El número de observaciones es de 195, el AIC es de 134.95, la escala o parámetro de dispersión es de 0.95, el test de Pearson 0.67 y el test de devianza 0.99. Todo ello indica un buen ajuste del modelo a los datos. Para finalizar el análisis de las categorías de la red aplicando el modelo único, se aplica dicho modelo para la categoría 5, siguiendo los mismos pasos. A continuación se muestra el resultado, donde el AIC es de 184.82, el número de observaciones 196, los coeficientes de las variables mantienen el mismo signo de las categorías anteriores, y la variable X.Grup3 es la única no significativa del modelo. El test de Pearson es de 0.37 y el test de devianza 0.78. Ambos resultados no significativos, confirman el buen ajuste del modelo a los datos, pero con valores menos cercanos a la unidad y en el test de Pearson, más cercano a la significación. Finalmente, el parámetro de dispersión es de 1.03 no siendo muy elevado, pero si mayor a las categorías anteriores.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x300_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4307 -0.6323  0.3950  0.7021  2.3545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  32.4942947   5.1954006   6.254 3.99e-10 ***
max_bw_path  -0.0009751   0.0003561  -2.738  0.00618 **
mean_bw_path -0.0087539   0.0028255  -3.098  0.00195 **
X.Grup3       0.6086553   1.0958089   0.555  0.57859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 271.71 on 196 degrees of freedom
Residual deviance: 176.82 on 192 degrees of freedom
AIC: 184.82

Number of Fisher Scoring iterations: 5

```

6.1.4 Modelo libre por categorías para BT

Una vez tenemos los modelos por categorías para el modelo único, se ha creado con las variables más idóneas para cada categoría, que será el llamado modelo libre por categorías, donde se escoge cada variable, hasta un total de 4, con el mismo procedimiento del apartado 6.1.1, es decir, proporcionando un menor valor del AIC del modelo. A continuación, se muestra el modelo libre con cuatro variables para la segunda categoría con un AIC igual a 38.749. El modelo se muestra abajo.

```

glm(formula = restored. ~ 0 + x.Grup2 + lbetLink + mean_path_length_before +
  min_path_length_before, family = binomial, data = x300_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.93478  0.01158  0.04689  0.20532  1.16623

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2       5.16813   2.22388   2.324  0.0201 *
lbetLink     41.10146  27.99896   1.468  0.1421
mean_path_length_before -0.18261   0.09404  -1.942  0.0522 .
min_path_length_before  3.92942   3.04895   1.289  0.1975
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 270.327 on 195 degrees of freedom
Residual deviance: 30.749 on 191 degrees of freedom
AIC: 38.749

Number of Fisher Scoring iterations: 9

```

Se observan los signos de los coeficientes de las variables positivos para todas menos para mean_path_length_before siendo este negativo. Esta variable conjuntamente con X.Grup2 son significativas al 10%, el resto no lo son. El test de Pearson y el test de devianza son igual a 1 indicando, como se ha comentado en los anteriores modelo, el buen ajuste del modelo. Su escala o parámetro de dispersión es de 0.54, valor menor al modelo presentado con el modelo único para su misma categoría.

Seguimos con el modelo libre de 4 variables y menor AIC de la categoría 3. El resultado se muestra a continuación.

```

glm(formula = restored. ~ 0 + Grup2 + Grup3 + mean_path_length_before +
    min_path_length_before, family = binomial, data = x300_1_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.38624   0.02090   0.05398   0.21220   0.92835

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
Grup2             0.35841    0.11475   3.123  0.00179 **
Grup3             0.34754    0.16034   2.168  0.03019 *
mean_path_length_before -0.09919    0.04723  -2.100  0.03570 *
min_path_length_before  0.92312    0.70366   1.312  0.18956
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 264.782 on 191 degrees of freedom
Residual deviance: 39.803 on 187 degrees of freedom
AIC: 47.803

Number of Fisher Scoring iterations: 8

```

Se observa los signos de los coeficientes de todas las variables, excepto para mean_path_length_before, positivo. Salvo la variable min_path_length_before, el resto de coeficientes es significativo al 5%. El AIC es de 47.803 y la escala de 0.32, también menor al modelo único de su categoría. El test de Pearson y el de devianza son igual a 1. En resumen, el modelo libre obtiene valores similares en los dos test al modelo único es su categoría, la diferencia es el AIC, el parámetro de dispersión al ser en este caso menores y la significatividad de las variables.

El modelo libre para la categoría 4, a diferencia del modelo único, contiene todos los coeficientes de las variables positivas menos para la variable max_bw_path. El AIC es de 127.72 y el test de Pearson de 0.88. El parámetro de dispersión 0.88 y el test de devianza es 0.99. En conjunto, nos indica ser un buen modelo, el test de Pearson es más cercano a la unidad del modelo único propuesto en su categoría y el AIC menor. Los coeficientes de las variables son significativas al 5% para max_bw_path y mean_path_length_before, el resto no son significativas. El modelo se presenta a continuación:

```

glm(formula = restored. ~ 0 + lbetLink + mean_path_length_before +
    max_bw_path + spectralRadiusF, family = binomial, data = x300_1_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4453   0.2125   0.3701   0.5217   1.4417

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
lbetLink         8.2743569    9.6636184   0.856  0.391866
mean_path_length_before  0.0347048    0.0154471   2.247  0.024660 *
max_bw_path      -0.0010941    0.0003288  -3.328  0.000875 ***
spectralRadiusF    2.0214576    1.1582581   1.745  0.080940 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 270.33 on 195 degrees of freedom
Residual deviance: 119.72 on 191 degrees of freedom
AIC: 127.72

Number of Fisher Scoring iterations: 5

```

El modelo libre, con las cuatro variables escogidas tal y como se ha ido indicando en el apartado 6.1.1 para la última categoría analizada, la quinta, tiene el siguiente resultado:


```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + max_bw_path +
  avgLbetF, family = binomial, data = x300_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4225 -0.6155  0.3800  0.7009  2.4235

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  28.6506831  6.0833047   4.710 2.48e-06 ***
mean_bw_path -0.0091753  0.0028428  -3.228 0.00125 **
max_bw_path  -0.0010118  0.0003628  -2.789 0.00529 **
avgLbetF     14.1016999 10.6583928   1.323 0.18582
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 271.71 on 196 degrees of freedom
Residual deviance: 175.34 on 192 degrees of freedom
AIC: 183.34

Number of Fisher Scoring iterations: 5

```

Las variables, salvo avgLbetF son las mismas al modelo único, esta es no significativa en el modelo, los coeficientes de algConnectF y avgLbetF son positivos en contra de los otros dos con signo contrario. El test de Pearson es de 0.35 y el test de devianza 0.80, ambos similares al modelo único para la red BT e informando de estar ajustado el modelo a los datos. El parámetro de dispersión es de 1.04 informando de una no muy elevada sobredispersión.

6.1.5 Elección del mejor modelo por categorías para BT

Una vez tenemos los modelos por categorías de los apartados 6.1.3 y 6.1.4, se elige el mejor modelo por categoría, el modelo único por categoría. Al tener varios modelos por categorías, las tablas 6-3 y 6-4 tienen el resumen de los modelos y la información más destacada de ellos. En ellas se pueden ver diferencias en los resultados del modelo único con el modelo libre para cada una de las categorías, la diferencia en la categoría 2 y 3 está en un mejor valor del parámetro de dispersión y menor AIC, algo de esperar al tratarse del modelo con las variables más idóneas de su categoría. En la categoría 5, tiene un test de Pearson no significativo pero

Tabla 6-3 resumen modelo único por categorías de la red BT

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo único BT de la categoría 2	46.741	0.99	1	0.75
Modelo único BT de la categoría 3	59.65	0.99	1	0.67
Modelo único BT de la categoría 4	134.95	0.67	0.99	0.95
Modelo único BT de la categoría 5	184.82	0.37	0.78	1.03

Tabla 6-4 resumen modelo libre por categorías de la red BT

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo libre BT de la categoría 2	38.749	1	1	0.54
Modelo libre BT de la categoría 3	47.803	1	1	0.32
Modelo libre BT de la categoría 4	127.72	0.88	0.99	0.88
Modelo libre BT de la categoría 5	183.34	0.35	0.80	1.04

considerablemente menor del resto de categorías además de ser también un poco más cercano a la unidad en el modelo libre con respecto al único en el test de devianza.

La significación de algConnectF en todas las categorías del modelo único, las dos últimas categorías del modelo libre han incluido max_bw_path con un coeficiente significativo y en la última categoría también incluir mean_bw_path y algConnectF significativamente evidencian la similitud de la categoría con las variables el modelo único.

Para decidir cuál de los dos modelos presentados es mejor en cada categoría, se ha llevado a cabo el análisis de la devianza anova, para saber el más idóneo. Más abajo se muestran los resultados del anova, donde el p-valor significativo en todas las categorías implica no poder rechazar la hipótesis nula de ser el modelo único presentado en 6.1.1 para la red BT, el mejor a representar los datos en cada una de las categorías. Es decir, las variables del modelo son las mismas para cada una de las categorías. Aun teniendo menor valor de AIC para cada categoría con el modelo libre, el modelo único por categorías es el presentado en el apartado 6.1.3.

Analysis of Deviance Table Categoría 2					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + X.Grup2 + lbetLink + mean_path_length_before + min_path_length_before					
	Resid.	Df	Resid. Dev	Df	Deviance Pr(>Chi)
1	191		38.741		
2	191		30.749	0	7.9929 0
Analysis of Deviance Table Categoría 3					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + Grup2 + Grup3 + mean_path_length_before + min_path_length_before					
	Resid.	Df	Resid. Dev	Df	Deviance Pr(>Chi)
1	187		51.650		
2	187		39.803	0	11.846 0
Analysis of Deviance Table Categoría 4					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + lbetLink + mean_path_length_before + max_bw_path + spectralRadiusF					
	Resid.	Df	Resid. Dev	Df	Deviance Pr(>Chi)
1	191		126.95		
2	191		119.72	0	7.2289 0
Analysis of Deviance Table Categoría 5					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + algConnectF + mean_bw_path + max_bw_path + avgLbetF					
	Resid.	Df	Resid. Dev	Df	Deviance Pr(>Chi)
1	192		176.82		
2	192		175.34	0	1.482 0

A continuación vamos a ver el resumen del modelo único por categoría con los coeficientes de las variables en la tabla 6-5. Siempre tiene signo positivo el coeficiente de algConnectF, menos para la segunda categoría X.Grup3 tiene signo positivo. Las más variables son max_bw_path y mean_bw_path, la primera tiene coeficiente negativo en todas las categorías excepto para la cuarta, y la segunda tiene coeficiente positivo en la segunda categoría y en el resto negativo.

Para interpretar los coeficientes de las variables de cada categoría se aplica el odds ratio a cada coeficiente en cada uno de los modelos único por categorías al igual que se hizo en el apartado 6.1.2. El resultado figura en la Tabla 6-6. Para algConnectF, el odds ratio es superior a

Tabla 6-5 Tabla resumen modelo único por categorías

Modelo	Coefficientes variables	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo único BT de la categoría 2	algConnectF 28.03 max_bw_path -0.001 mean_bw_path 0.001 X.Grup 3 -3.84	46.741	0.99	1	0.75
Modelo único BT de la categoría 3	algConnectF 25.40 max_bw_path -0.001 mean_bw_path -0.001 X.Grup 3 1.20	59.65	0.99	1	0.67
Modelo único BT de la categoría 4	algConnectF 8.27 max_bw_path 0.03 mean_bw_path -0.001 X.Grup 3 2.02	134.95	0.67	0.99	0.95
Modelo único BT de la categoría 5	algConnectF 32.50 max_bw_path -0.0009 mean_bw_path -0.008 X.Grup 3 0.61	184.82	0.37	0.78	1.03

la unidad, por tanto un incremento de una unidad en la variable aumenta la probabilidad de ser la variable respuesta $Y=1$ en un $(1-1.49e+12)\%$ para la categoría 2, 107% en la tercera, 3903.95% en la cuarta y $(1-1.30e+14)\%$ en la quinta. Los valores de los odds ratio para algConnectF son altos por tanto, para cada unidad de aumento de la variable en su categoría, me facilita la restauración del fallo. En el caso de max_bw_path se mantiene igual en todas las categorías excepto para la cuarta donde incrementa en un 0.03% por cada unidad que incrementa la variable y esto provoca un empeoramiento para encontrar una red ancha y larga y mejorar la restauración ($Y=1$), en el resto de casos, al ser un odds ratio menor a la unidad, para cada unidad de incremento en cada una de las categorías, la variable tiene un efecto sobre la variable respuesta a la inversa, esta disminuye en 0.01%, lo cual es mejor porque nos comunica el hecho de restaurar redes que no sean tan largas y pesadas por cada unidad de incremento en la variable. Para mean_bw_path nos pasa lo mismo, cuando el odds ratio es inferior a la unidad, como hemos comentado para max_bw_path, excepto un leve incremento del 0.002% en la segunda categoría cuando aumenta una unidad la variable y tendrá el efecto contrario sobre la variable respuesta.

Finalmente, X.Grup3 tiene el valor del odds ratio positivo en la tercera, cuarta y quinta categoría donde para cada unidad que incrementa la variable, esta nos indica una mejora de 2.32%, 6.54% y .84% respectivamente, en la restauración de los fallos al tener mayor número de enlaces que no empiezan o terminan en los nodos A y B, por tanto es más fácil encontrar una ruta para ellos al no ser nodos de inicio o fin de la red. Para la segunda categoría nos informa de lo contrario, para cada unidad de incremento de la variable, esta disminuirá la probabilidad de restauración en un 0.98%.

Tabla 6-6 Coeficientes de las variables de los modelos únicos por categoría

Categoría 2	algConnectF	$\exp(28.03) = 1.49e+12$
	max_bw_path	$\exp(-0.001) = 0.99$
	mean_bw_path	$\exp(0.001) = 1.001$
	X.Grup3	$\exp(-3.84) = 0.02$
Categoría 3	algConnectF	$\exp(25.40) = 107418687183$
	max_bw_path	$\exp(-0.001) = 0.99$
	mean_bw_path	$\exp(-0.001) = 0.99$
	X.Grup3	$\exp(1.20) = 3.32$
Categoría 4	algConnectF	$\exp(8.27) = 3904.95$
	max_bw_path	$\exp(0.03) = 1.03$
	mean_bw_path	$\exp(-0.001) = 0.99$
	X.Grup3	$\exp(2.02) = 7.54$
Categoría 5	algConnectF	$\exp(32.50) = 1.30e+14$
	max_bw_path	$\exp(-0.0009) = 0.99$
	mean_bw_path	$\exp(-0.008) = 0.99$
	X.Grup3	$\exp(0.61) = 1.84$

Para acabar el apartado, se ha realizado la curva de ROC para todas las categorías mostrado en la Fig. 6.4. En las categorías 2, 3 y 5, se observa la curva de ROC con valores entre 0.8 y 0.9 indicando un buen ajuste del modelo, la cuarta categoría con valor del área bajo la curva igual a 0.718, donde pasa a ser un ajuste medio. En general, tenemos un buen ajuste del modelo único por categorías.

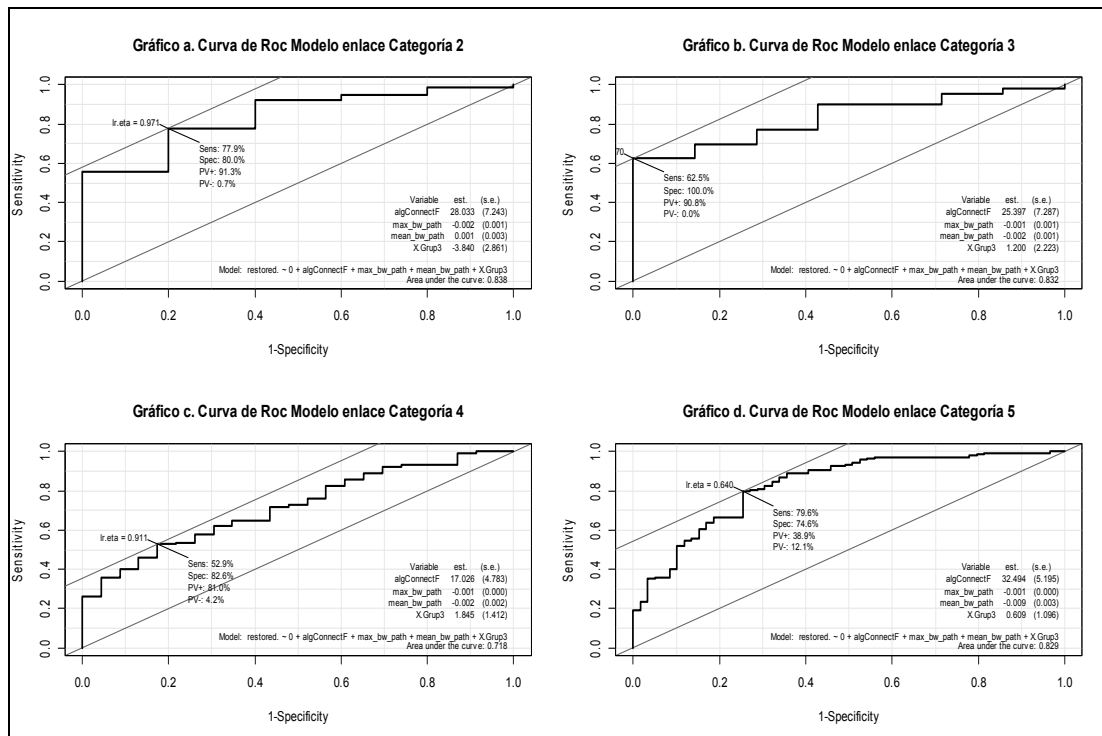


Fig. 6-3 Curva de ROC para las categorías de la red BT

En resumen, para los datos de la red BT tenemos un único modelo, todo y tener menor valor del AIC y un poco más de sobredispersión en los datos, pero de forma muy leve. Tanto para todos los datos como para las diferentes categorías, el modelo único presentado en el apartado 6.1.1 nos ofrece un buen ajuste. A continuación, se van a analizar las dos redes restantes, buscando contrastar si realmente el modelo único tanto para todos los datos, como el modelo único por categorías, también nos puede servir para analizar otras redes.

6.2 Modelos para la red DT

A la hora de buscar los modelos para la red DT, primero se ha aplicado el modelo único para todos los datos resultante de la red BT para los datos de la nueva red, seguidamente se ha buscado un modelo con 4 variables y se le ha llamado modelo libre. Dicho modelo se calcula a través de la búsqueda de aquellas variables, hasta un total de cuatro, escogiendo cada una de ellas a través del modelo con menor AIC, es decir, se siguió los mismos pasos para añadir variable a variable pero como en la red anterior, pero ahora solo presentamos dos modelos, el modelo único aplicado a la red DT y el modelo libre. Una vez tenemos cada modelo se interpretan los signos de los modelos, el test chi-cuadrado de Pearson, el test de la devianza y, se calcula la escala. Se lleva a cabo la comparación y elección del mejor modelo (modelo único) a través del anova y se finaliza con la interpretación de los coeficientes del modelo único a través de los odds ratio y la curva de ROC del modelo escogido.

El modelo único de la red BT aplicado a la red DT tiene un AIC igual a 951.24, los coeficientes de las variables mantienen el signo, es decir, algConnectF y X.Grup3 son positivos y las otras dos tienen el coeficiente negativo. Todas las variables tienen el coeficiente significativo. Ahora bien, el test de Pearson es igual a 0 y el de devianza a 1 habiendo contradicción entre ellos, mientras uno nos indica un buen ajuste de los datos el otro no. La escala es de 3.59, valor no muy elevado y sugiriendo una dispersión de los datos no muy elevada. El modelo se ve más abajo.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
     X.Grup3, family = binomial, data = x260_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0745  0.1202  0.2235  0.4179  4.1484

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  13.0711675  0.7436409  17.577 < 2e-16 ***
max_bw_path  -0.0010757  0.0001776  -6.059 1.37e-09 ***
mean_bw_path -0.0078586  0.0011783  -6.670 2.57e-11 ***
X.Grup3       1.6642021  0.4745114   3.507 0.000453 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2739.32  on 1976  degrees of freedom
Residual deviance:  943.24  on 1972  degrees of freedom
AIC: 951.24

Number of Fisher Scoring iterations: 6

```

El otro modelo presentado para todos los datos de la red DT es el modelo libre que vemos a continuación. Vemos dos variables del modelo único anterior incluidas en el modelo libre (algConnectF y mean_bw_path). Todos los coeficientes de las variables son significativos, mientras que el valor del AIC es de 926.27. Al igual que en el modelo anterior, el test de Pearson es igual a 0 y el de devianza 1. El cambio viene dado a través de la escala, siendo esta menor con un valor igual a 2.61.

```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
max_path_length_before, family = binomial, data = x260_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0014   0.1212   0.2104   0.3555   4.0474

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF    15.309824   0.943552  16.226 < 2e-16 ***
mean_bw_path   -0.008011   0.001106  -7.241 4.44e-13 ***
link_node_A    -0.129630   0.026406  -4.909 9.15e-07 ***
max_path_length_before -0.343649   0.090439  -3.800 0.000145 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2739.32 on 1976 degrees of freedom
Residual deviance: 918.27 on 1972 degrees of freedom
AIC: 926.27

Number of Fisher Scoring iterations: 6

```

Para poder escoger el mejor modelo para representar la red DT, el modelo único, se realiza el anova el cual nos indica una significación al contrastar los dos modelos. Un p-valor significativo nos informa del no rechazo del modelo 1, es decir, el modelo único para la red BT utilizado inicialmente en el análisis. Por tanto, para la red DT también tenemos el mismo modelo único.

```

Analysis of Deviance Table

Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
max_path_length_before
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1972      943.24
2      1972      918.27 0    24.97 0

```

Para finalizar vamos a interpretar los coeficientes del modelo en términos de odds ratio, Para algConnectF tenemos un odds ratio superior a la unidad, nos indica para cada unidad incrementada en la variable, tiene un incremento de (474491)% sobre la variable respuesta y esta se acercará más a conseguir restaurar el fallo.

```

Odds Ratio de los coeficientes:
algConnectF    exp(13.07) = 474492
max_bw_path    exp(-0.001)= 0.99
mean_bw_path   exp(-0.007)= 0.99
X.Grup3        exp(1.66) = 5.26

```

Para max_bw_path y mean_bw_path el odds ratio es de 0.99, por tanto cada unidad de incremento en cada una de las variables, ésta disminuye en 0.01% respectivamente, es decir, nos aleja de obtener una probabilidad cercana a la unidad en la variable respuesta (Y=1) lo cual es positivo al decirnos que para poder restaurar no debe de aumentar el ancho y el largo de las conexiones ya que esto dificultará la restauración. Para X.Grup3, el valor del odds ratio es positivo entonces, cada unidad que incrementa la variable, esta tiene un efecto incremental del 4.26% sobre la variable respuesta, mejorando la restauración de los fallos al tener mayor número de conexiones que no empiezan o terminan en los nodos A y B y siendo más fácil encontrar una ruta para ellos.

Para finalizar, se ha realizado la curva de ROC mostrada en la Fig. 6.4. Se observa el área bajo la curva de ROC con valor igual a 0.876 indicando un buen ajuste del modelo, al ser un valor superior al 0.8. En resumen, el modelo único presentado es un buen modelo.

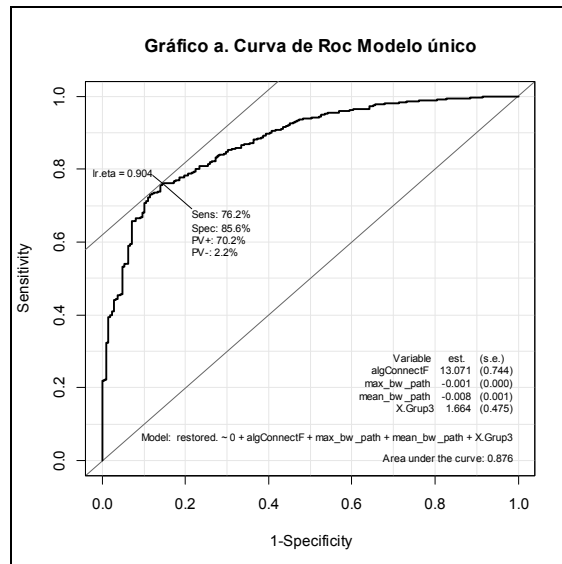


Fig. 6-4 Curva de ROC para el modelo único de la red DT

6.2.1 Modelo único por categorías de la red DT

Para analizar las categorías de la red DT, se ha calculado, tal y como se hizo en el apartado 6.1.2, una escala de 6 categorías, donde la primera y la última no nos proporcionan información relevante para modelizar pero las otras cuatro sí. En la tabla 6-7 vemos los puntos de corte que se han creado a partir de la variable creada para hacer las categorías. Los intervalos no tiene el mismo número de observaciones para cada categoría, esta se tuvo que hacer manualmente al ser los datos ajustados en la escala y por cada decimal se hacían agrupaciones con una desigualdad considerable.

Tabla 6-7 Punto de corte para la red DT

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Red DT	0 – 6.11	6.12 – 7.95	7.96 – 8.83	8.84 – 10.31	10.32 – 13.89	> 13.90

A continuación, empezamos el análisis con la misma dinámica la red BT cuando se hizo las categorías, primeramente se ha analizado para cada grupo dos modelos, el primer modelo corresponde al modelo único por categorías de la red BT. Para la categoría 2 para la red DT, el modelo mostrado seguidamente corresponde al modelo único por categoría de la red BT pero aplicado a los datos actuales. La categoría 2 contiene 461 observaciones, el AIC es de 58.56, los coeficientes de las variables tienen signo negativo para todas las variables excepto para algConnectF, la variable max_bw_path es la única significativa del modelo, indicando a las otras no ser relevantes. El test de Pearson es de 1 y el test de devianza 1. Ambos resultados no significativos, nos confirman un buen ajuste del modelo a los datos, el parámetro de dispersión es de 0.61, valor bajo indicando no haber demasiada sobredispersión.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x260_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1676   0.0761   0.0973   0.1225   1.0648

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  1.283e+01  2.138e+00   6.000 1.98e-09 ***
max_bw_path -3.107e-03  1.410e-03  -2.204  0.0275 *
mean_bw_path  6.648e-04  5.458e-03   0.122  0.9031
X.Grup3      2.212e+02  4.178e+04   0.005  0.9958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.08  on 461  degrees of freedom
Residual deviance:  50.56  on 457  degrees of freedom
AIC: 58.56

Number of Fisher Scoring iterations: 22

```

Seguimos con la siguiente categoría, observando a continuación que contiene 445 observaciones, el AIC del modelo es de 147.41, los coeficientes de las variables tienen signo negativo para las variables max_bw_path y mean_bw_path y las otras dos positivo. La variable mean_bw_path y algConnectF tienen los coeficientes significativos. El test de Pearson es de 0.12 y el test de devianza 1. Al ser el test de Pearson y el de devianza no significativos, nos confirman un buen ajuste del modelo para representar los datos, aunque el test de Pearson presenta valores próximos a la significación. Por último, el parámetro de dispersión es de 1.08.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x260_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1292   0.1331   0.1849   0.2975   1.1187

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF 13.6540888  1.7324059   7.882 3.23e-15 ***
max_bw_path -0.0008591  0.0005282  -1.627  0.103809
mean_bw_path -0.0094675  0.0025235  -3.752  0.000176 ***
X.Grup3      2.4731949  1.9294512   1.282  0.199908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 616.90  on 445  degrees of freedom
Residual deviance: 139.41  on 441  degrees of freedom
AIC: 147.41

Number of Fisher Scoring iterations: 7

```

El análisis para la categoría 4 se muestra a continuación, más abajo tenemos el resultado de aplicar el modelo único por categoría a los datos. El número de observaciones de la categoría es de 455. En el modelo, vemos un AIC de 196.89, los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF, la variable max_bw_path y algConnectF tienen los coeficientes significativos. El test de Pearson es de 0.97 y el de devianza 1. Al ser ambos test no significativos, informan de un buen ajuste del modelo a representar los datos. Además, el parámetro de dispersión es de 0.88, un valor bajo. El modelo tiene buenos indicios de ser apropiado aun teniendo dos variables no significativas al igual de las otras categorías anteriores.


```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x260_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9100   0.2112   0.2916   0.3733   1.8317

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF 12.1970489  1.7003704   7.173 7.33e-13 ***
max_bw_path -0.0013882  0.0004458  -3.114 0.00185 **
mean_bw_path -0.0022770  0.0018073  -1.260 0.20771
X.Grup3      -0.0399826  1.1753539  -0.034 0.97286
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 630.76  on 455  degrees of freedom
Residual deviance: 188.89  on 451  degrees of freedom
AIC: 196.89

Number of Fisher Scoring iterations: 6

```

Para finalizar el análisis de las categorías con el modelo único, a continuación vemos el resultado de aplicar el modelo único a los datos de la categoría 5. El número de observaciones de la categoría es de 451. En el modelo, los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF y X.Grup3, la variable max_bw_path es la única no significativa del modelo. El valor del AIC es de 391.33, El test de Pearson es de 0.54 y el test de devianza 0.99. Al ser ambos test no significativos, nos confirman un buen ajuste del modelo. Adicionalmente, el parámetro de dispersión es de 0.99, un valor bajo. A continuación se sigue con el siguiente apartado con el modelo libre por categorías.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x260_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6079   0.3146   0.4809   0.6328   1.5425

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF 10.3809787  1.7658245   5.879 4.13e-09 ***
max_bw_path -0.0003024  0.0003314  -0.913 0.361430
mean_bw_path -0.0104354  0.0027562  -3.786 0.000153 ***
X.Grup3      2.5461130  0.7542277   3.376 0.000736 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.22  on 451  degrees of freedom
Residual deviance: 383.33  on 447  degrees of freedom
AIC: 391.33

Number of Fisher Scoring iterations: 5

```

6.2.2 Modelo libre por categorías para DT

El modelo libre por categorías es el modelo libre presentado para la red DT en el apartado anterior, pero ahora se va a aplicar a las distintas categorías. A continuación vemos el resultado de aplicar el modelo libre a la segunda categoría. El valor del AIC es de 48.19, los coeficientes de las variables son negativos en todas excepto para algConnectF, todos los coeficientes de las variables son significativas del modelo al 5%. El test de Pearson y el test de devianza es 1. Ambos resultados no significativos e igual a la unidad, nos confirman un buen ajuste del modelo. El parámetro de dispersión es de 0.24, un valor bajo confirmado una sobredispersión de los datos leve. Otro dato a tener en cuenta es el hecho de tener la variable

mean_bw_path y algConnectF en el modelo libre, lo cual indica la importancia de estas en otra red y no solo para la red BT analizada anteriormente.

```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
  max_path_length_before, family = binomial, data = x260_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.81692  0.00831  0.01609  0.11633  0.85846

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
algConnectF    40.008390  11.879001   3.368 0.000757 ***
mean_bw_path   -0.008581   0.003758  -2.283 0.022410 *
link_node_A    -0.583242   0.252258  -2.312 0.020773 *
max_path_length_before -2.601455  1.042669  -2.495 0.012596 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.082  on 461  degrees of freedom
Residual deviance:  40.188  on 457  degrees of freedom
AIC: 48.188

Number of Fisher Scoring iterations: 10

```

El modelo libre para la categoría 3, se muestra a continuación. Los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF, la variable max_path_length_before es la única no significativa del modelo. El AIC es de 142.05, el test de Pearson de 0.7 y el test de devianza 1. El parámetro de dispersión es de 0.96, valor menor al modelo único para la misma categoría. Para la cuarta categoría el modelo libre se puede ver más abajo.

```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
  max_path_length_before, family = binomial, data = x260_l_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1350  0.1130  0.1785  0.2840  1.3478

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
algConnectF    17.347557  2.924340   5.932 2.99e-09 ***
mean_bw_path   -0.008914   0.002155  -4.137 3.52e-05 ***
link_node_A    -0.241700   0.082849  -2.917 0.00353 **
max_path_length_before -0.430808  0.299314  -1.439 0.15006
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 616.90  on 445  degrees of freedom
Residual deviance: 134.05  on 441  degrees of freedom
AIC: 142.05

Number of Fisher Scoring iterations: 7

```

Los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF, la variable max_path_length_before es la única no significativa del modelo. El AIC es de 142.05, el test de Pearson de 0.7 y el test de devianza 1. El parámetro de dispersión es de 0.96, valor menor al modelo único para la misma categoría.

Para la cuarta categoría el modelo libre se puede ver más abajo. El valor del AIC es de 202.42, los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF, la variable mean_bw_path está incluida en el modelo libre, link_node_A es la única no son las significativas del modelo al 10%, al 5% no es significativo el coeficiente de la variable mean_bw_path. El test de Pearson es de 0.97 y el de devianza 1. Al ser ambos test no significativos, reafirman un buen ajuste del modelo a los datos. El parámetro de dispersión es

de 0.88, un valor bajo e igual al modelo único presentado para la misma categoría.

```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
max_path_length_before, family = binomial, data = x260_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6876   0.2670   0.3027   0.3419   2.5301

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF    13.862385   2.409314   5.754 8.73e-09 ***
mean_bw_path    -0.003865   0.002290  -1.688  0.0914 .
link_node_A     -0.006805   0.099109  -0.069  0.9453 .
max_path_length_before -0.560265   0.263912  -2.123  0.0338 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 630.76  on 455  degrees of freedom
Residual deviance: 194.42  on 451  degrees of freedom
AIC: 202.42

Number of Fisher Scoring iterations: 6

```

Para finalizar el análisis de las categorías, el modelo libre para la quinta categoría tiene el valor del AIC es de 389.09, El test de Pearson es de 0.4 y el test de devianza 0.99, confirmando un buen ajuste del modelo representar los datos. El parámetro de dispersión es de 1.01, un valor bajo y un poco menor al modelo único de la categoría. Los coeficientes de las variables tienen signo negativo en todas excepto para algConnectF, todas las variables son significativas del modelo al 10%, el coeficiente de max_path_length_before no es significativo al 5%. El modelo presentado parece tener mejores resultados, al modelo único propuesto para la categoría.

```

glm(formula = restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
max_path_length_before, family = binomial, data = x260_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5580   0.2605   0.4919   0.6294   1.7312

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF    15.902401   2.281781   6.969 3.19e-12 ***
mean_bw_path    -0.008510   0.002505  -3.398  0.00068 ***
link_node_A     -0.118831   0.039889  -2.979  0.00289 **
max_path_length_before -0.335128   0.188601  -1.777  0.07558 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.22  on 451  degrees of freedom
Residual deviance: 381.09  on 447  degrees of freedom
AIC: 389.09

Number of Fisher Scoring iterations: 5

```

6.2.3 Elección del mejor modelo por categorías para DT

Al igual que se hizo en el apartado 6.1.5 se han creado dos tablas resumen de los modelos, la Tabla 6-8 muestra los modelos único por categoría y la 6-9 los modelos libre por categoría para la red DT. Se observa las diferencias de los resultados del modelo único con el modelo libre para cada una de las categorías. A las categorías 2, 3 y 5, el modelo libre tiene menor AIC, las categorías 2 y 3 tienen el parámetro de dispersión menor en el modelo libre respecto al único. El test de Pearson para el modelo libre en la quinta categoría tiene un valor menor a la unidad respecto al modelo libre pero en el resto de categorías es igual o más aproximado a 1. La no significación de algunos de los coeficientes de las variables de los modelos tanto del único

como del libre para cada categoría, pueden indicar la necesidad de incluir una variable aun sin ser significativa en su modelo.

Tabla 6-8 Tabla resumen del modelo único por categorías para la red DT

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo único DT de la categoría 2	58.56	1	1	0.61
Modelo único DT de la categoría 3	147.41	0.12	1	1.08
Modelo único DT de la categoría 4	196.89	0.97	1	0.88
Modelo único DT de la categoría 5	391.33	0.54	0.99	0.99

Tabla 6-9 Tabla resumen del modelo libre por categorías para la red DT

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo libre DT de la categoría 2	48.19	1	1	0.24
Modelo libre DT de la categoría 3	142.05	0.7	1	0.96
Modelo libre DT de la categoría 4	202.42	0.97	1	0.88
Modelo libre DT de la categoría 5	389.09	0.4	0.99	1.01

Hasta ahora se han visto los modelos pero existe la necesidad de comparar los modelos entre sí en cada categoría a fin de saber cuál es el mejor. Para hacerlo, se lleva a cabo el análisis de la devianza, la anova. Más abajo se muestran los resultados donde se observa el p-valor significativo para todas las categorías, por lo tanto, en todos los casos no se puede rechazar la hipótesis nula de ser el modelo único por categoría, el mejor a representar los datos en cada una de las categorías de la red DT. Por consiguiente, todo y tener un modelo con menor valor de AIC para cada grupo e incluso una escala o parámetro de dispersión menor en alguna de las categorías, el modelo único por categorías es el mejor.

Analysis of Deviance Table Categoría 2	
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3	
Model 2: restored. ~ 0 + algConnectF + mean_bw_path + link_node_A + max_path_length_before	
Resid. Df	Resid. Dev Df Deviance Pr(>Chi)
1	457 50.560
2	457 40.188 0 10.372 0
Analysis of Deviance Table Categoría 3	
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3	
Model 2: restored. ~ 0 + algConnectF + mean_bw_path + link_node_A + max_path_length_before	
Resid. Df	Resid. Dev Df Deviance Pr(>Chi)
1	441 139.41
2	441 134.05 0 5.3646 0
Analysis of Deviance Table Categoría 4	
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3	
Model 2: restored. ~ 0 + algconnectF + mean_bw_path + link_node_A + max_path_length_before	
Resid. Df	Resid. Dev Df Deviance Pr(>Chi)
1	451 188.89
2	451 194.42 0 -5.5337 0

Analysis of Deviance Table Categoría 5

```

Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + algConnectF + mean_bw_path + link_node_A +
max_path_length_before
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      447      383.33
2      447      381.09  0  2.2357  0
    
```

Esta conclusión nos indica la importancia del modelo único inicial encontrado en la red DT para representar los datos de otras redes. Seguidamente vamos a interpretar los coeficientes de las variables para cada categoría.

Tabla 6-10 Tabla con los coeficientes del modelo único por categoría

Modelo	Coefficientes variables	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo único DT de la categoría 2	algConnectF 1.283e+01 max_bw_path -3.107e-03 mean_bw_path 6.648e-04 X.Grup 3 2.212e+02	58.56	1	1	0.61
Modelo único DT de la categoría 3	algConnectF 13.65 max_bw_path -0.0008 mean_bw_path -0.009 X.Grup 3 2.47	147.41	0.12	1	1.08
Modelo único DT de la categoría 4	algConnectF 12.20 max_bw_path -0.001 mean_bw_path -0.002 X.Grup 3 -0.04	196.89	0.97	1	0.88
Modelo único DT de la categoría 5	algConnectF 10.38 max_bw_path -0.0003 mean_bw_path -0.01 X.Grup 3 2.55	391.33	0.54	0.99	0.99

Para algConnectF en todas las categorías el coeficiente es positivo, max_bw_path en contra es negativo, para mean_bw_path excepto en la categoría 2 tiene signo negativo y por último X.Grup3 tiene el coeficiente positivo en todas las categorías menos para la cuarta. A la hora de calcular los odds ratio tal y como se observa en la tabla 6-11, un odds ratio inferior a la unidad comunica para cada unidad de incremento del coeficiente de la variable con el odds ratio menor a 1, implica una disminución, en el tanto por ciento restante al valor de la unidad. Para algConnectF, en todas las categorías es superior a la unidad, para cada unidad de aumento de dicha variable en cualquiera de las categorías, esta tendrá un efecto significativo en la variable respuesta en un 373247.6%, 847459.9%, 198788.2 y 32207.96% respectivamente. En el caso de max_bw_path se mantiene igual en todos los casos, es decir, el odds ratio es menor a la unidad y por eso para cada unidad de incremento en cada una de las categorías, esta disminuye en 0.01%, algo bueno porque nos dice que max_bw_path para aumentar la probabilidad de restaurar el fallo no ha de aumentar, es decir, no ha de ser mayor el ancho y largo de la red para tener mejor porcentaje de restauración. Para mean_bw_path, en todos los

casos o es igual o menor a la unidad, para la segunda categoría al tener valor igual a 1 nos indica que no tiene efecto por cada unidad de incremento de la variable sobre la variable respuesta, las otras categorías pasa igual a max_bw_path, una unidad de incremento disminuye en un 0.01%. Finalmente, X.Grup3 tiene un odds ratio superior a la unidad en la categorías dos, tres y cinco, aumentando en un (1-1.16e+96)%, 10.82% y 11.81% respectivamente y para la categoría 4 lo hace en sentido contrario, cada unidad de aumento, decrece en un 0.01% respecto a la variable respuesta.

Tabla 6-11 Coeficientes de las variables de los modelos únicos por categoría de la red DT

Categoría 2	algConnectF	$\exp(1.283e+01) = 373248.6$
	max_bw_path	$\exp(-3.107e-03) = 0.99$
	mean_bw_path	$\exp(6.648e-04) = 1.00$
	X.Grup3	$\exp(2.212e+02) = 1.16e+96$
Categoría 3	algConnectF	$\exp(13.65) = 847460.9$
	max_bw_path	$\exp(-0.0008) = 0.99$
	mean_bw_path	$\exp(-0.009) = 0.99$
	X.Grup3	$\exp(2.47) = 11.82$
Categoría 4	algConnectF	$\exp(12.20) = 198789.2$
	max_bw_path	$\exp(-0.001) = 0.99$
	mean_bw_path	$\exp(-0.002) = 0.99$
	X.Grup3	$\exp(-0.04) = 0.96$
Categoría 5	algConnectF	$\exp(10.38) = 32208.96$
	max_bw_path	$\exp(-0.0003) = 0.99$
	mean_bw_path	$\exp(-0.01) = 0.99$
	X.Grup3	$\exp(2.55) = 12.81$

Finalmente mostramos la curva de ROC para cada categoría de la red DT el modelo único por categoría. En la Fig. 6-5 podemos ver las categorías 2 y 3 con el área bajo la curva superiores 0.8, al estar entre 0.8 y 0.9 sus valores nos indican un buen ajuste del modelo. Las categorías 4 y 5 tienen un valor de 0.78 y 0.75 respectivamente, ambas entre 0.7 y 0.8 con un ajuste medio. Todo y tener unos valores no demasiado ajustados en general, el modelo único por categorías es bueno y es un buen candidato para analizar los datos de la red.

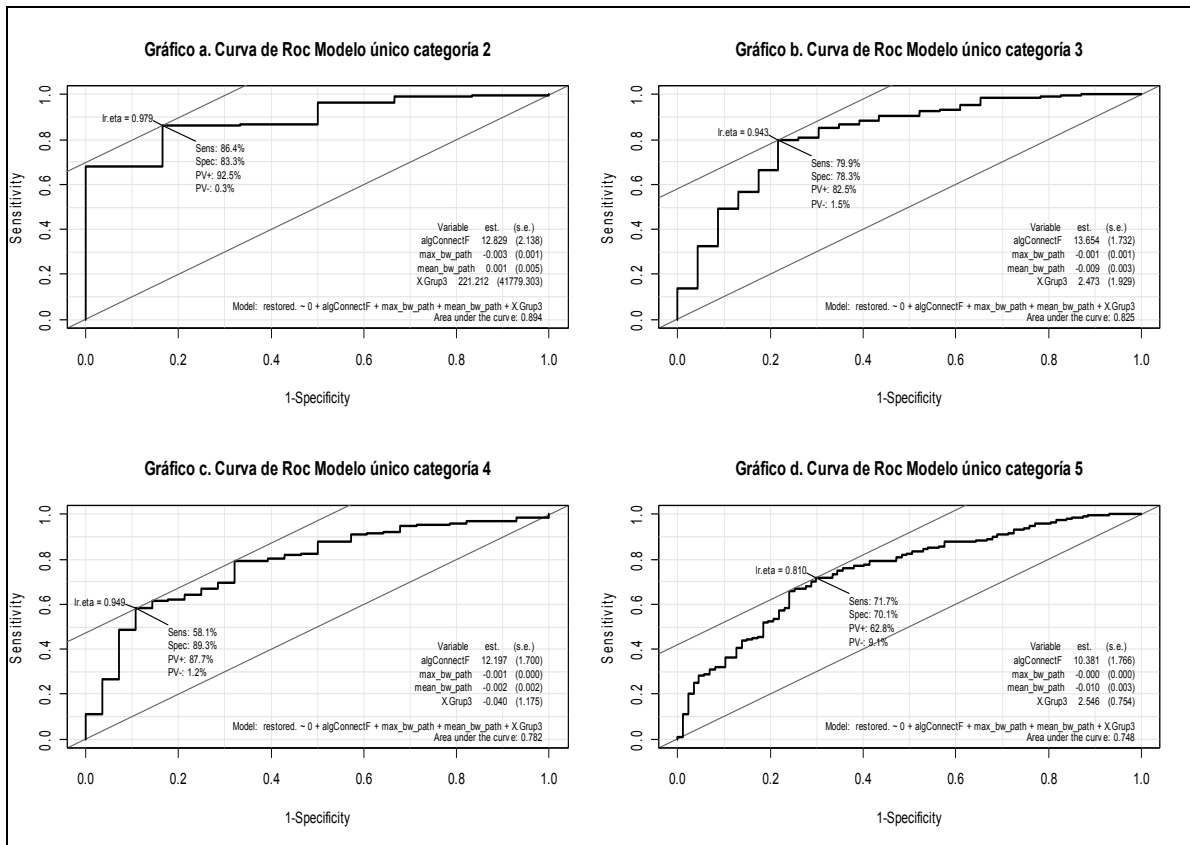


Fig. 6-5 Curva de ROC para la red DT de las categorías

6.3 Modelo de la red TEL

A la hora de buscar los modelos para la red TEL, primero se ha aplicado el modelo único para todos los datos de la red BT tal y como se hizo en la red DT, después se ha buscado un modelo con 4 variables llamado modelo libre siguiendo el procedimiento explicado anteriormente. Dicho modelo se calcula a través de la búsqueda de aquellas variables, hasta un total de cuatro, escogiendo cada una de ellas a través del modelo con menor AIC, es decir, se siguió los mismos pasos para añadir variable a variable pero. Presentamos dos modelos, el modelo único aplicado a la red TEL y el modelo libre. Una vez tenemos cada modelo se interpretan los signos de los modelos, el test de Pearson, el de devianza y, se calcula la escala o parámetro de dispersión. Una vez lo tenemos los resultados, se lleva a cabo la comparación y elección del mejor modelo, el modelo único, a través del anova y se finaliza con la interpretación de los coeficientes del modelo a través de los odds ratio y la curva de ROC del modelo escogido, llamado modelo único.

El modelo único aplicado a la red TEL tiene un AIC igual a 4524.4, los coeficientes de las variables son distintos, para algConnectF es positivo y para las otras tienen el coeficiente negativo. Todas las variables tienen el coeficiente significativo. El test de Pearson es igual a 0.99 y el de devianza a 1. Ambos nos dan un buen ajuste. La escala es de 0.93, valor no muy elevado y sugiriendo una dispersión de los datos no muy elevada. Dichos resultados se muestran a continuación:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x520_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5404   0.2788   0.4771   0.6650   2.2898

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  1.528e+01  5.176e-01  29.521 < 2e-16 ***
max_bw_path -4.596e-04  3.946e-05 -11.646 < 2e-16 ***
mean_bw_path -1.309e-03  2.336e-04  -5.603 2.11e-08 ***
X.Grup3      -1.500e+00  1.707e-01  -8.785 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7429.2 on 5359 degrees of freedom
Residual deviance: 4516.4 on 5355 degrees of freedom
AIC: 4524.4

Number of Fisher Scoring iterations: 5

```

El modelo libre para la red TEL se muestra a continuación. Tiene un AIC igual a 4573.7, los coeficientes de las variables son positivas en todas excepto para max_bw_path. Todas las variables tienen el coeficiente significativo. El modelo presenta max_bw_path al igual del modelo único lo cual indica la importancia a la hora de mejorar la probabilidad de restauración el ancho y lago de la red. El test de Pearson es igual a 0.99 y el de devianza a 1. Ambos nos dan un buen ajuste. La escala es de 0.92, valor no muy elevado lo cual es una dispersión de los datos no muy elevada.

```

glm(formula = restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path +
  algConnectVar, family = binomial, data = x520_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5229   0.2957   0.4837   0.6716   1.5051

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2      1.127e+00  1.880e-01  5.996 2.03e-09 ***
X.Grup1      2.452e+00  6.581e-01  3.725 0.000195 ***
max_bw_path  -5.858e-04  3.407e-05 -17.193 < 2e-16 ***
algConnectVar 2.316e+00  1.576e-01  14.691 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7429.2 on 5359 degrees of freedom
Residual deviance: 4565.7 on 5355 degrees of freedom
AIC: 4573.7

Number of Fisher Scoring iterations: 5

```

Los dos modelos presentados tienen un valor del AIC similar, al igual del resto de resultados siendo difícil escoger uno de ellos. Por este motivo se ha realizado el anova entre los dos modelos para saber cual es mejor. Al ser el p-valor significativo no se puede rechazar la hipótesis de ser el mejor modelo el modelo único presentado, por tanto, al igual de la red anterior, el modelo único para la red BT también lo es para la red DT y para la red TEL.

Analysis of Deviance Table					
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3					
Model 2: restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path + algConnectVar					
	Resid.	Df	Resid. Dev	Df	Deviance Pr(>Chi)
1	5355		4516.4		
2	5355	0	4565.7	0	-49.379 0

Para finalizar vamos a interpretar los coeficientes del modelo en términos de odds ratio, Para algConnectF tenemos un odds ratio superior a la unidad, nos indica para cada unidad

incrementada en la variable, tiene un incremento de (474491)% sobre la variable respuesta y esta se acercará más a conseguir restaurar el fallo.

Odds Ratio de los coeficientes:		
algConnectF	$\exp(1.528e+01)$	= 4325334
max_bw_path	$\exp(-4.596e-04)$	= 0.99
mean_bw_path	$\exp(-1.309e-03)$	= 0.99
X.Grup3	$\exp(-1.500e+00)$	= 0.22

Para max_bw_path y mean_bw_path el odds ratio es de 0.99, inferior a la unidad, por tanto cada unidad de incremento en cada una de las variables, ésta disminuye en 0.01% respectivamente, es decir, nos aleja de obtener una probabilidad cercana a la unidad en la variable respuesta (Y=1) lo cual es positivo al decirnos que para poder restaurar no debe de aumentar el ancho y el largo de la red al ser difícil encontrar rutas para redes más largas a la hora de hacer la restauración. Para X.Grup3, el valor del odds ratio también es negativo, cada unidad que incrementa la variable, esta tiene el mismo efecto comentado anteriormente pero en un 0.78% sobre la variable respuesta, informando de tener redes con menor probabilidad de restauración donde estas no empiecen o acaben en los nodos A y B. Por último algConnectF, al igual de las redes anteriores, tiene un valor positivo por tanto, cada unidad de incremento tendrá un efecto positivo del 4325333% de restaurar el fallo con éxito. Una conectividad algebraica mayor significa mejor conectividad y por tanto mayor posibilidad de encontrar una ruta alternativa cuando se ha producido un fallo en la red.

Para finalizar, se ha realizado la gráfica de la curva de ROC del modelo único mostrada en la Fig. 6.6. Se observa el área bajo la curva de ROC con valor igual a 0.76 indicando un ajuste medio del modelo. No es un ajuste muy bueno, pero ya comentamos en el capítulo 5 la proporción de fallos no restaurados de la red con un 18% frente al 10% en las otras dos redes. Por tanto, todo y tener unos valores no excesivamente idóneos, son lo suficientemente buenos para decidir tener el modelo único como el más bueno a representar la red TEL. Modelo único por categorías de la red TEL.

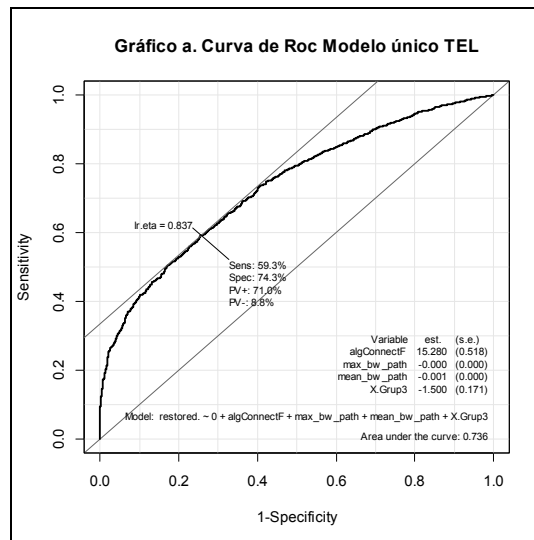


Fig. 6-6 Curva de ROC para el modelo único de la red TEL

6.3.1 Modelo único por categorías de la red TEL

En el siguiente apartado se ha seguido con la misma dinámica de las redes anteriores, primeramente se ha analizado para cada grupo dos modelos, el primer modelo corresponde al

modelo único por categorías de la red BT del apartado 6.1.1. Para ello se han creado las diferentes categorías tal y como se muestra en la tabla 6-12 los puntos de corte de cada categoría a fin de tener para las comprendidas entre la segunda y la quinta con una proporción de datos similar.

Tabla 6-12 Puntos de corte por categorías para la red TEL

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Telefónica	0 – 8.51	8.52 – 9.11	9.12 – 9.55	9.56 – 9.95	9.96 – 13.99	> 14.00

Empezamos el análisis de la categoría 2 para la red TEL, con el modelo único por categorías de la red BT aplicado a los datos correspondientes. Se muestra seguidamente:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x520_1_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4780  0.3374  0.4059  0.4834  1.1225

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  14.3027220  1.1457809  12.483 < 2e-16 ***
max_bw_path  -0.0002944  0.0001232  -2.389  0.01688 *
mean_bw_path -0.0012193  0.0005427  -2.247  0.02466 *
X.Grup3      -1.2917487  0.4618610  -2.797  0.00516 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1515.22  on 1093  degrees of freedom
Residual deviance:  680.42  on 1089  degrees of freedom
AIC: 688.42

Number of Fisher Scoring iterations: 5

```

La categoría 2 contiene 1093 observaciones. El AIC es de 688.42, los coeficientes de las variables son todas negativas excepto algConnectF, todos los coeficientes de las variables son significativos. El test de Pearson es 0.71 y el test de devianza igual a 1, la escala o parámetro de dispersión es de 0.98, valor pequeño informando de poca sobredispersión. Los test indican un buen ajuste del modelo a los datos, además la escala comunica la poca sobredispersión en los datos, por tanto parece un buen modelo.

Continuamos el análisis con la categoría 3, seguidamente se ve el resultado de aplicar el modelo único por categoría a las 1095 observaciones de la tercera categoría de la red TEL. El AIC es de 987.81, el test de Pearson es de 0.65 y el de devianza 0.99, el parámetro de dispersión 0.98, valor pequeño y similar a la categoría anterior, indica poca sobredispersión de los datos. Los test prueban un buen ajuste del modelo. Los coeficientes de las variables son positivo para algConnectF y negativo para el resto. La variable X.Grup3 es significativa al 10% y el resto al 5%.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x520_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1977  0.4728  0.5564  0.6357  1.5778

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  12.1802595  1.2082579  10.081 < 2e-16 ***
max_bw_path  -0.0002196  0.0001049  -2.094  0.03622 *
mean_bw_path -0.0016545  0.0006323  -2.617  0.00887 **
X.Grup3      -0.7500951  0.4046343  -1.854  0.06377 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1517.99  on 1095  degrees of freedom
Residual deviance:  979.81  on 1091  degrees of freedom
AIC: 987.81

Number of Fisher Scoring iterations: 4

```

En la categoría 4, el modelo único presenta un AIC de 1148.6. Recordemos el cambio de categoría, por lo tanto el número de observaciones ahora es de 1094. Con respecto a la categoría anterior, el AIC aumenta, los coeficientes de las variables son iguales a las categorías anteriores. El test de Pearson es de 0.58 y el de devianza 0.14, no significativos pero con valores más próximos a la significación en relación al resto de categorías visto anteriormente, en especial el test de devianza, el de Pearson se mantiene más o menos igual. El parámetro de dispersión es de 0.99, valor mantenido en las otras categorías. El modelo se muestra más a continuación:

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x520_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1137  0.4858  0.6315  0.7362  1.5005

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  1.261e+01  1.514e+00  8.328 < 2e-16 ***
max_bw_path -2.672e-04  9.096e-05  -2.937  0.00331 **
mean_bw_path -2.405e-03  6.253e-04  -3.847  0.00012 ***
X.Grup3      -5.129e-01  4.397e-01  -1.166  0.24343
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1516.6  on 1094  degrees of freedom
Residual deviance: 1140.6  on 1090  degrees of freedom
AIC: 1148.6

Number of Fisher Scoring iterations: 4

```

Para finalizar visualizamos el modelo único para la categoría 5. Contiene 1078 observaciones y tiene el valor del AIC de 1301.4. Los coeficientes de las variables son significativos en algConnectF y mean_bw_path, para las otras dos no. Los signos de los coeficientes son igual al resto de categorías analizadas. Adicionalmente, el test de Pearson es de 0.48 siendo no significativo pero disminuyendo su valor y aproximándose a la significación respecto a la unidad y el test de devianza 4.12e-06 significativo, informando de un mal ajuste del modelo a los datos, por tanto hay contradicción entre los dos test. El parámetro de dispersión es de 1.00. El modelo se ve seguidamente.

```

glm(formula = restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path +
  X.Grup3, family = binomial, data = x520_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8806  -1.3242   0.7401   0.8641   2.1462

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
algConnectF  1.169e+01  1.704e+00   6.861 6.86e-12 ***
max_bw_path  -8.303e-05  8.383e-05  -0.990   0.322
mean_bw_path -3.713e-03  7.266e-04  -5.110 3.21e-07 ***
X.Grup3      -4.519e-01  5.142e-01  -0.879   0.380
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1494.4  on 1078  degrees of freedom
Residual deviance: 1293.4  on 1074  degrees of freedom
AIC: 1301.4

Number of Fisher Scoring iterations: 4

```

6.3.2 Modelo libre por categorías para TEL

El modelo libre por categorías, al igual que en la red DT, se ha escogido el modelo libre para todas las observaciones de la red TEL y se ha aplicado a cada categoría. Para la segunda categoría el modelo libre es el siguiente:

```

glm(formula = restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path +
  algConnectVar, family = binomial, data = x520_l_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4572   0.3360   0.4087   0.4931   0.8880

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2      1.329e+00  4.754e-01   2.796  0.00517 **
X.Grup1      1.589e+00  1.568e+00   1.013  0.31088
max_bw_path  -4.486e-04  9.621e-05  -4.663 3.12e-06 ***
algConnectVar 2.163e+00  3.985e-01   5.429 5.66e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1515.22  on 1093  degrees of freedom
Residual deviance:  686.22  on 1089  degrees of freedom
AIC: 694.22

Number of Fisher Scoring iterations: 5

```

El AIC tiene un valor de 694.22, los coeficientes de las variables son positivos excepto para max_bw_path. El parámetro de dispersión es de 0.99, test de devianza 1 y el de Pearson es 0.97. Ambos no significativos y por tanto muestran un buen ajuste del modelo a los datos. El parámetro de dispersión igual a 0.97 informa de no existir demasiada sobredispersión en los datos.

El modelo libre para la categoría 3 presenta los coeficientes de las variables positivos para todas excepto para max_bw_path. Las variables X.Grup2 y X.Grup1 son no significativas y el resto sí. El modelo libre incluye una variable en común con el modelo único, max_bw_path. El AIC es de 993.99, el test de Pearson 0.57 y el de devianza 0.99, además, el parámetro de dispersión o escala es de 0.99, valor pequeño y similar a la categoría anterior. Los test muestran un buen ajuste del modelo y tiene mayor valor del AIC con respecto al modelo único presentado para la categoría.

```

glm(formula = restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path +
  algConnectVar, family = binomial, data = x520_l_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1513   0.4680   0.5678   0.6459   1.0332

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2      6.233e-01  4.074e-01  1.530   0.126
X.Grup1      2.437e+00  1.800e+00  1.354   0.176
max_bw_path  -3.898e-04  8.242e-05 -4.730 2.25e-06 ***
algConnectVar 2.121e+00  3.398e-01  6.244 4.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1517.99  on 1095  degrees of freedom
Residual deviance:  985.99  on 1091  degrees of freedom
AIC: 993.99

Number of Fisher Scoring iterations: 4

```

El modelo libre de la cuarta categoría se puede ver seguidamente. Las variables X.Grup2 y X.Grup1 son no significativas, además el signo del coeficiente se mantiene igual a las categorías anteriores. El valor del AIC es de 1162.1, el parámetro de dispersión de 0.99, el test de Pearson es de 0.58 y el de devianza 0.09. Aun siendo no significativos, se ha de considerar el test de devianza, para los dos modelos, como un valor próximo a ser significativo.

```

glm(formula = restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path +
  algConnectVar, family = binomial, data = x520_l_3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0563   0.4810   0.6487   0.7416   1.2925

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2      3.982e-01  4.323e-01  0.921   0.357
X.Grup1      3.080e-01  2.065e+00  0.149   0.881
max_bw_path  -4.621e-04  7.646e-05 -6.044 1.51e-09 ***
algConnectVar 2.332e+00  2.981e-01  7.822 5.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1516.6  on 1094  degrees of freedom
Residual deviance: 1154.1  on 1090  degrees of freedom
AIC: 1162.1

Number of Fisher Scoring iterations: 4

```

El modelo libre para la última categoría presenta un AIC de 1327.5. Los coeficientes de las variables son significativas para todas excepto para X.Grup2 y X.Grup1, todos los coeficientes son negativos menos para max_bw_path. El test de Pearson es de 0.5 no significativo y el de devianza 3.6e-07 significativo contradiciendo el buen ajuste del modelo entre los dos test.

El parámetro de dispersión es de 0.99, valor mantenido en las dos categorías anteriores, por lo tanto nos hay demasiada sobredispersión de los datos. El modelo se puede ver a continuación.

```

glm(formula = restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path +
    algConnectVar, family = binomial, data = x520_l_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8106  -1.3956   0.7870   0.8817   1.2427

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
X.Grup2      0.2966853  0.5273443   0.563   0.574
X.Grup1     -2.2039837  2.1641104  -1.018   0.308
max_bw_path -0.0003149  0.0000702  -4.486 7.27e-06 ***
algConnectVar 1.7787061  0.3492152   5.093 3.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1494.4  on 1078  degrees of freedom
Residual deviance: 1319.5  on 1074  degrees of freedom
AIC: 1327.5

Number of Fisher Scoring iterations: 4

```

6.3.3 Elección del mejor modelo por categorías para TEL

En el análisis previo se han hecho dos modelos para cada categoría, para poder recordarlos, se ha creado la tabla 6-13 y 6-14 con el resumen de todos y la información más destacada de ellos. En las diferentes categorías hay similitudes entre el modelo único y el libre, las categorías 4 y 5 nos dan un test de devianza menor a las otras categorías, incluso en la última es significativo tanto para el modelo único como el modelo libre por categoría.

Tabla 6-13 Tabla resumen de modelo único por categorías para la red TEL

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo único DT de la categoría 2	688.42	0.71	1	0.98
Modelo único DT de la categoría 3	987.81	0.65	0.99	0.98
Modelo único DT de la categoría 4	1148.6	0.58	0.14	0.99
Modelo único DT de la categoría 5	1301.4	0.48	4.12e-06	1.00

Tabla 6-14 Tabla resumen de modelo libre por categorías para la red TEL

Modelo	AIC	T. Pearson	T. Devianza	P. dispersión
Modelo libre DT de la categoría 2	694.22	0.72	1	0.97
Modelo libre DT de la categoría 3	993.99	0.57	0.99	0.99
Modelo libre DT de la categoría 4	1162.1	0.58	0.09	0.99
Modelo libre DT de la categoría 5	1327.5	0.5	3.6e-07	0.99

También se ha de tener en cuenta el efecto incremental del AIC a medida que aumentamos de categoría, además de haber pasado también en las otras categorías. Todos los grupos mantienen un parámetro de sobredispersión más o menos igual, además, en todos los casos el modelo libre presenta menor AIC.

Continuamos el análisis para escoger el mejor modelo por categoría para la red TEL. A

continuación se muestra el anova de cada categoría, en todas las categorías el p-valor es significativo. La hipótesis nula del análisis de la devianza nos propone como mejor modelo el modelo 1, es decir, el modelo único por categoría, de igual forma que ocurrió en la red DT, el modelo único del apartado 6.1.1 es el mejor modelo al no poder rechazar la hipótesis nula por ser el p-valor igual significativo e igual a 0.

Analysis of Deviance Table Categoría 2

```
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path + algConnectVar
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1089      680.42
2         1089      686.22  0  -5.7967  0
```

Analysis of Deviance Table Categoría 3

```
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path + algConnectVar
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1091      979.81
2         1091      985.99  0  -6.1802  0
```

Analysis of Deviance Table Categoría 4

```
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path + algConnectVar
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1090      1140.6
2         1090      1154.1  0  -13.434  0
```

Analysis of Deviance Table Categoría 5

```
Model 1: restored. ~ 0 + algConnectF + max_bw_path + mean_bw_path + X.Grup3
Model 2: restored. ~ 0 + X.Grup2 + X.Grup1 + max_bw_path + algConnectVar
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1074      1293.4
2         1074      1319.5  0  -26.154  0
```

Para poder interpretar los coeficientes de las variables del modelo, en la tabla 6.14 se ven los resultados de hacer el odds ratio para cada variable en su conveniente categoría. Para algConnectF, cada unidad de aumento en la variable, tendrá un efecto creciente en un 1623345%, 194851.9%, 299539% y 119371% para las cada una de las categorías sobre la variable respuesta. Para max_bw_path se mantiene igual en todos los casos, es decir, el odds ratio es menor a la unidad y por eso para cada unidad de incremento en cada una de las categorías, esta disminuye en 0.01%, algo bueno porque nos dice que max_bw_path para aumentar la probabilidad de restaurar el fallo no ha de aumentar, es decir, no ha de ser mayor el ancho y largo de la red para tener mejor porcentaje de restauración. Para mean_bw_path, en todos los casos o es igual o menor a la unidad, para la segunda categoría al tener valor igual a 1 nos indica que no tiene efecto por cada unidad de incremento de la variable sobre la variable respuesta, las otras categorías pasa igual a max_bw_path, una unidad de incremento disminuye en un 0.01%. Finalmente, X.Grup3 tiene un odds ratio inferior a la unidad todas las categorías por tanto, cada unidad de aumento de la variable, decrece en un 0.72%, 0.53%, 0.40% y 0.36% respectivamente sobre la variable respuesta.

Para terminar con el análisis de la red TEL, se ejecuta la curva de ROC para cada categoría (Fig. 6-17), tal y como se comentó en las otras redes. Para todas las categorías, el área bajo la curva está entre 0.6 y 0.7 informando de un mal buen ajuste a los datos. El hecho de salir la curva de ROC similar en todas las categorías, nos dice en conclusión, que la red TEL, aun no teniendo un ajuste ideal, al tener en los otros análisis buenos resultados, sigue siendo el modelo único por categorías un modelo válido.

Tabla 6-14 Coeficientes de las variables de los modelos únicos por categoría

Categoría 2	algConnectF	$\exp(14.30) = 1623346$
	max_bw_path	$\exp(-0.0003) = 0.99$
	mean_bw_path	$\exp(-0.001) = 1.00$
	X.Grup3	$\exp(-1.29) = 0.28$
Categoría 3	algConnectF	$\exp(12.18) = 194852.9$
	max_bw_path	$\exp(-0.0002) = 0.99$
	mean_bw_path	$\exp(-0.001) = 0.99$
	X.Grup3	$\exp(-0.75) = 0.47$
Categoría 4	algConnectF	$\exp(1.261e+01) = 299539$
	max_bw_path	$\exp(-2.672e-04) = 0.99$
	mean_bw_path	$\exp(-2.405e-03) = 0.99$
	X.Grup3	$\exp(-5.129e-01) = 0.60$
Categoría 5	algConnectF	$\exp(1.169e+01) = 119372$
	max_bw_path	$\exp(-8.303e-05) = 0.99$
	mean_bw_path	$\exp(-3.713e-03) = 0.99$
	X.Grup3	$\exp(-4.519e-01) = 0.64$

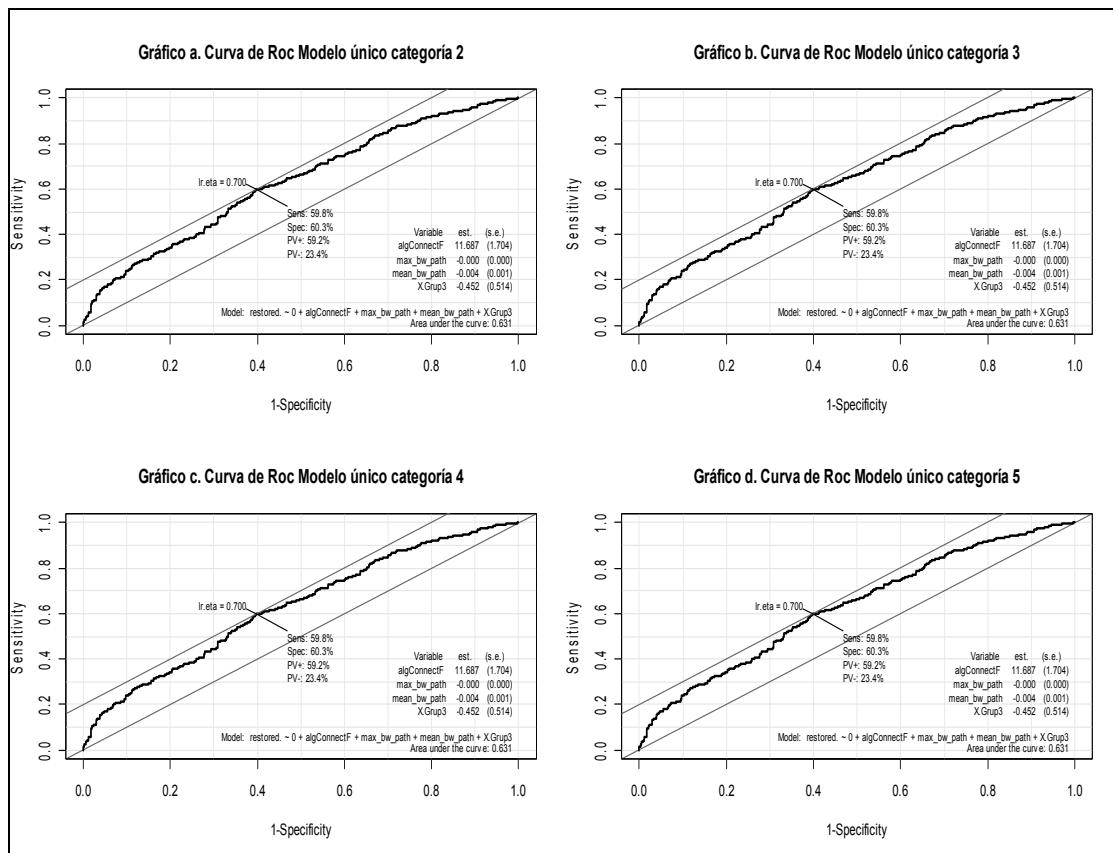


Fig. 6-7 Curva de ROC para las categorías de la red TEL

6.4 Capacidad predictiva de los modelos

A la vista de los resultados similares entre las distintas redes y teniendo como mejor modelo el modelo único siendo el mismo para todos los datos como por categorías, se ha realizado la tabla de contingencias para ver la calidad de las predicciones de los modelos. Recordemos el hecho de tener un valor verdadero no restaurado significa tener un modelo donde los errores son predichos con mayor exactitud y por este motivo elegiremos aquel cuya proporción de verdaderos no restaurados sea mayor, sea el modelo único para todos los datos o el modelo único por categorías.

En la tabla 6-15 se muestran las tablas de contingencia del modelo nulo y de los modelos únicos con todos los datos y por categorías para las tres redes. Podemos ver como el modelo nulo acierta solo un 11% en la red DT y BT y un 18% en la red TEL de las verdaderos 0, lo cual supone una proporción muy baja. Por el contrario, el modelo único con todos los datos mejora esta proporción hasta un 33%, 29% y 38% para las redes DT, TEL y BT respectivamente. Esta valor, aunque puede parecer pequeño, en realidad mejora sustancialmente la capacidad predictiva del modelo nulo (del orden de 3 veces más aciertos). Además, hay que recordar que nos encontramos ante un evento muy poco probable y, por lo tanto, su predicción se hace mucho más complicada. Así que podemos considerar que estos valores de predicción son más que satisfactorios para la aplicación válida de estos modelos. Para el modelo único por categorías las proporciones son similares, si bien mejoran un poco, siendo de 38% para la red DT, 30% para TEL y 41% para BT. Hay que evaluar esta mejora con respecto del modelo único analizando la complejidad que supone estimar la categoría a la que pertenece una observación que queramos predecir.

Tabla 6-15 Predicciones para las tres redes

		Modelo Nulo			Modelo único todos los datos			Modelo único categorías		
	restored.	Predicción			Predicción			Predicción		
		0	1	total	0	1	total	0	1	total
Red BT	0	12	101	113	43	70	113	46	67	113
		0,11	0,89		0,38	0,62		0,41	0,59	
	1	101	868	969	77	892	969	63	906	969
		0,10	0,90		0,08	0,92		0,07	0,93	
Red DT	restored.	0	1	total	0	1	total	0	1	total
		22	187		209	70		139	209	
	0	0,11	0,89	0,33		0,67	0,38	0,62		
	1	187	1580	1767	136	1631	1767	125	1642	1767
0,11		0,89	0,08		0,92	0,07		0,93		
Red TEL	restored.	0	1	total	0	1	total	0	1	total
		179	800		979	287		692	979	
	0	0,18	0,82	0,29		0,71	0,30	0,70		
	1	799	3581	4380	688	3692	4380	680	3700	4380
0,18		0,82	0,16		0,84	0,16		0,84		

6.5 Conclusiones y elección del mejor modelo

En el presente capítulo se han encontrado dos modelos, el modelo único para todas las observaciones y el modelo único por categoría para cada una de las redes, validados después de probar varias alternativas.

Una vez obtenidos los mejores modelos para las tres redes, hemos demostrado tener un modelo único que contiene las mismas variables para distintas redes, ya tanto para el modelo que contempla todas las observaciones como para el modelo que separa la variable respuesta por categorías. Las variables implicadas nos informan de diversas características de las redes, que pasaremos brevemente a resumir.

Una de ellas pertenece a las características de la topología en ausencia del enlace fallado, a saber, `algConnectF`. Para todos siempre tiene un valor positivo y por tanto, a mayor conectividad algebraica mayor será la probabilidad de restaurar un fallo producido en la red, lo cual tiene mucho sentido.

Respecto a las variables `max_bw_path` y `mean_bw_path`, que nos informan sobre ancho de banda total ocupado por las conexiones que usaban el enlace antes del fallo, vemos que tienen un odds ratio negativo, es decir a mayor valor de la variable en el modelo, la restauración será más difícil, ya que las conexiones a restaurar requieren de mucha más capacidad disponible libre en los enlaces que siguen operativos.

En tercer lugar, `X.Grup3` nos da características de cómo son las conexiones que utilizaban el enlace antes del fallo en cuanto a los nodos origen y destino de éstas. A diferencia de las anteriores, esta variable presenta distinto comportamiento según la topología. Así, en las redes BT y TEL tiene un odds ratio superior a la unidad, lo que nos está informando que tener menor proporción de conexiones que no empiezan o terminan en los nodos adyacentes al enlace fallado, afecta negativamente a la restauración. En cambio para la red DT sucede lo contrario; un incremento en la variable `X.Grup3` aumenta la probabilidad de tener una restauración completa del tráfico afectado por el corte.

Tanto de los resultados de validación como de las tablas de predicción podemos decir que el modelo por categorías mejora el modelo con todos los datos. Así que, en la búsqueda del modelo con mayor precisión en la predicción de la restorabilidad, nos decantaríamos por la primera opción. De todas formas, la creación de las categorías, sobre todo cuando se quiere utilizar el modelo para predicción, resulta una tarea complicada y que está fuera del alcance de este proyecto. Es por este motivo que debido a la simplicidad del modelo único con todos los datos, proponemos su utilización como herramienta útil para el diseño de redes ópticas dinámicas sujetas a fallos. No obstante, no podemos descartar el uso del modelo por categorías si bien debemos puntualizar que requerirá de una aplicación más complicada y menos directa que el modelo propuesto.

En el siguiente capítulo introduciremos un caso concreto de planificación periódica de red dinámica que hace uso del modelo aquí obtenido.

Capítulo 7. Caso de aplicación práctica

En este capítulo, introducimos un caso práctico de aplicación del modelo estadístico para el diseño periódico de redes dinámicas en operación. En primer lugar presentamos la arquitectura de red considerada para albergar tal diseño para, después, proponer el uso del modelo estadístico de vulnerabilidad en el contexto de un problema específico de planificación de red.

7.1 Diseño periodico en redes dinámicas

En la Fig. 7-1 se muestra el diagrama de flujo del esquema de re-planificación periódica de red estudiado en esta sección (reproducido de [Ru14]). Los elementos más importantes que se pueden ver en esta figura son:

- Una red dinámica automatizada mediante un plano de control distribuido y un controlador centralizado basado en tecnología SDN (de *Software Defined Networks*). El controlador SDN incluye varios elementos como el elemento de cálculo de rutas (PCE) usado para establecer conexiones ante la llegada de una nueva petición o para encontrar las rutas alternativas en caso de restauración.
- El sistema de gestión de red (*NMS*), encargado de gestionar la red de forma global e implementar operaciones relacionadas con configuración, administración, seguridad, etc.
- El departamento de planificación de red (*Planning Department*) encargado de administrar el proceso de diseño de la red. Concretamente, monitoriza y analiza el estado de la red, detecta cuando una acción debe ser ejecutada, lanza procesos de planificación de la red, etc
- El departamento de ingeniería (*Engineering Department*) que realiza las acciones relacionadas con la instalación y configuración de nuevos equipos en la red.
- Una herramienta de planificación (*Planning Tool*) encargada de computar soluciones de cada problema de optimización relacionado con la planificación periódica de la red. Estos problemas cubren desde reconfiguraciones de la red existente, hasta ampliaciones y migraciones tecnológicas para cubrir escenarios distintos al actual.

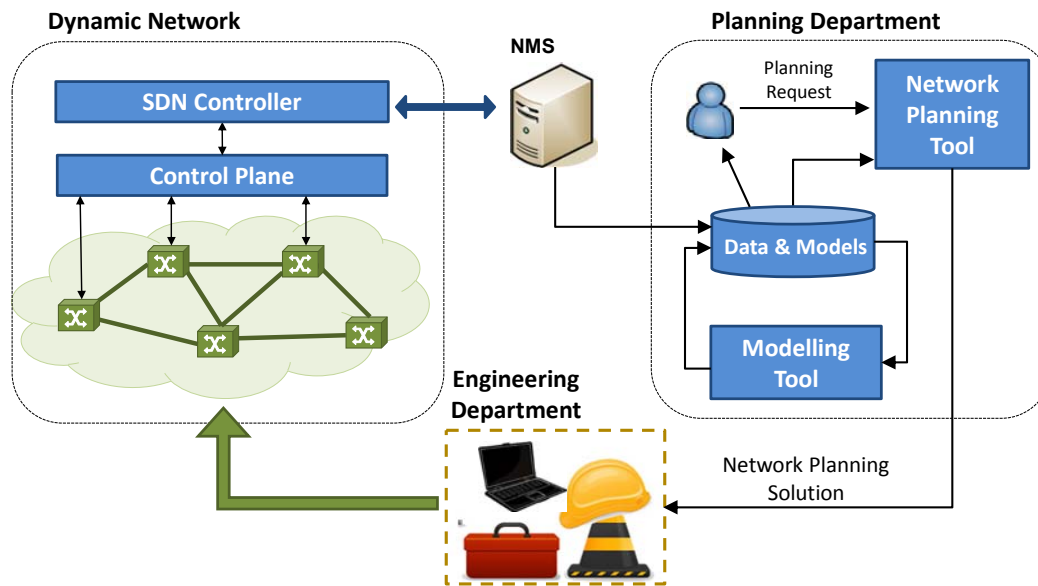


Fig. 7-1 Actores en la re-planificación periódica de la red

- Una herramienta de modelización (*Modelling tool*) encargada del ajuste y validación de modelos estadísticos para las distintas variables aleatorias relacionadas con redes dinámicas (e.g. evolución del tráfico, vulnerabilidad, etc). Esta herramienta requiere de intervención humana para el ajuste de los modelos pero puede operar de forma semi-automática con el fin de adaptar modelos existentes a las nuevas condiciones de tráfico.
- Las bases de datos con el histórico de datos y medidas obtenidas de la red y con el histórico de modelos estadísticos, que serán usados tanto de entrada a los problemas de planificación como para el reajuste y creación de nuevos modelos estadísticos.

Así, el departamento de planificación se encarga de ejecutar la resolución de problemas de reconfiguración y ampliación de la red cuando detecta que alguna condición necesaria no se está cumpliendo o la evolución del tráfico lleva a una degradación de la red que puede causar problemas en un futuro. Después de que el departamento de planificación cree una solución a dicho problema de optimización, ésta se envía al departamento de ingeniería para que organice y planifique las operaciones necesarias para implementar la solución en la red. Una vez los cambios son implementados y verificado su correcto funcionamiento, entonces la red puede continuar operando con los nuevos recursos instalados.

En la siguiente sección mostramos un caso detallado de problema de planificación para reducir la vulnerabilidad de una red en deterioro debido a la evolución del tráfico.

7.2 Ejemplo de problema de re-planificación dinámica

El problema que vamos a utilizar como caso ilustrativo de aplicación de los modelos consiste en un problema de re-planificación de una red en operación. Específicamente, el problema se lanza cuando, ante los datos estimados de tráfico futuro, se prevé que la vulnerabilidad de alguno de los enlaces de la red llegará a una cota no deseada (en términos del grado de restorabilidad). Ante esta circunstancia, el problema de re-planificación se ejecuta con el fin de ampliar la topología convenientemente (añadiendo enlaces no existentes) y así aumentar la capacidad en aquellas zonas más vulnerables. De este modo, reduciremos la vulnerabilidad de

la red y la prepararemos para afrontar en mejores condiciones el escenario de tráfico futuro.

El enunciado del problema es como sigue:

Dado:

- Un topología representada por el grafo (N,E) , donde N representa el conjunto de nodos y E el conjunto de enlaces de fibra.
- Un conjunto E' con aquellos enlaces no presentes en E y que pueden ser añadidos a la red.
- Las características del tráfico esperado para cada enlace, caracterizado por la tupla $t_e = \{\text{meanBW_Path}, \text{maxBW_Path}, X.\text{Grup3}\}$.
- El grado de vulnerabilidad máximo permitido para cada enlace

Salida:

- El conjunto de enlaces E^* , formado por elementos en E' , y que serán añadidos a la topología G

Objetivo: Minimizar el coste de despliegue de los nuevos enlaces en E'

Sujeto a: el nivel de vulnerabilidad de cada enlace no supere el máximo fijado.

En este escenario de re-planificación periódica, el papel que juegan los modelos de predicción definidos en este proyecto es doble, como se detalla a continuación.

7.2.1 Activación del ciclo de planificación

El departamento de planificación de la red tiene acceso a varios modelos que modelan el comportamiento del tráfico dinámico. Podemos asumir que algunos de ellos se encargan de predecir la evolución del tráfico en el tiempo, con el fin de anticipar los cambios que puedan surgir. Sin entrar en muchos detalles, podemos considerar que el uso de técnicas basadas en modelos de series temporales o técnicas de “machine learning” podrían ser utilizadas para obtener dichos modelos.

A partir de las estimaciones producidas por estos modelos de evolución del tráfico, se pueden estimar los valores de las variables referentes al tráfico presentes en el modelo de vulnerabilidad (es decir, mean_bw_path , max_BW_Path , y $X.\text{Grup3}$) a un plazo vista de pocas semanas o meses. Con estos valores de tráfico estimado, se puede calcular la vulnerabilidad de cada uno de los enlaces de la red actual con el modelo único obtenido en el Capítulo 6 y ver si alguno de ellos podría superar, en el plazo temporal estudiado, el nivel de vulnerabilidad crítico. Si es así, entonces el departamento de planificación lanzará la re-planificación de la red.

7.2.2 Computo de la vulnerabilidad durante la optimización

Tal y como se ha descrito en el problema de re-planificación detallado anteriormente, la solución deberá asegurar que la red resultante cumpla con los requisitos de vulnerabilidad fijados. Gracias a los modelos obtenidos, se pueden obtener estimaciones de esta vulnerabilidad para una topología resultante de añadir algunos nuevos enlaces a la red ya

existente. Es por ello que el método de resolución del problema de re-planificación debe hacer uso de los modelos propuestos.

De forma muy genérica podemos decir que la metodología de resolución de este tipo de problemas estaría basada fundamentalmente en dos técnicas:

- *Modelos de programación matemática* que aseguren la condición de máxima vulnerabilidad a partir del uso del modelo como una restricción del problema. Dado que el modelo propuesto no cumple con las características de una ecuación lineal (p.e. la conectividad algebraica requiere del cálculo de valores propios de una matriz), esta formulación pertenecería a la familia de modelos de programación matemática no lineales. Por todo esto, esta metodología, aunque permitiría formular formalmente el problema para obtener la solución óptima, es poco práctico a la hora de obtener soluciones.
- *Heurísticas* que permitan integrar modelos no lineales como el modelo de vulnerabilidad y otros posibles modelos de predicción de tráfico de una forma más manejable. El objetivo de este método sería el de obtener soluciones de buena calidad de una forma eficiente. Ya sea bien mediante heurísticas deterministas o aleatorizadas o, incluso, el uso de meta-heurísticas, el problema puede ser resuelto con un esfuerzo computacional asumible.

Nótese que estos cambios impactarán claramente en el funcionamiento de los algoritmos dinámicos encargados de enrutar nuevas conexiones y restaurar el tráfico afectado por un corte. Además, hay que recordar que si ha habido una re-planificación de la red quiere decir que ha habido cambios sustanciales en las condiciones de tráfico. Por lo tanto, está claro que el proceso de ajuste de los modelos estadísticos debe de ser realizado de forma continua para adaptarse a estos cambios. Este proceso continuo permite fortalecer el potencial de los modelos para predecir nuevos escenarios de tráfico.

Capítulo 8. Conclusiones

8.1 Contribuciones y próximos pasos

En este proyecto se han planteado la obtención de modelos estadísticos para la estimación del grado de vulnerabilidad de una red óptica dinámica sujeta a fallos de enlace. Concretamente, se han buscado modelos capaces de predecir, a partir de las condiciones de tráfico y las características del enlace, si la restauración del tráfico afectado por un corte podría ser o no realizada por completo.

Siguiendo el esquema clásico de un trabajo de este estilo, hemos diseñado el experimento necesario para obtener datos a partir de simulación que sirvieran para la fase de modelización. Para ello hemos contado con tres redes reales de operadores nacionales, con el objetivo de obtener modelos específicos para cada una de ellas. Después de esta obtención de datos y como previo paso a la modelización, se ha realizado un análisis descriptivo de los datos con el fin de validar estos para el proceso de estimación de modelos.

Sin duda, la máxima aportación de este trabajo ha sido el proceso de estimación de los modelos, ya que se han probado varias alternativas con el fin de buscar modelos precisos con un número reducido de variables. Tras un exhaustivo análisis, hemos obtenido dos modelos válidos para cada red, uno único para todo el rango de observaciones y otro separado por categorías de una variable creada para tal efecto. Ambos modelos presentan unos coeficientes sencillos y directos de analizar y logran mejorar sustancialmente la predicción básica obtenida por el modelo nulo. Hay que destacar que todos los modelos presentan el mismo conjunto de variables pero con coeficientes distintos según el caso. Este resultado era uno de los objetivos previstos de este proyecto, ya que esta condición abre la puerta a plantear la obtención de un modelo más complejo que pueda adaptarse a varias redes similares a las aquí estudiadas.

Finalmente, a modo de comprobar que los modelos se pueden usar para su finalidad, se presenta un caso práctico de planificación periódica de la red en la que el modelo propuesto no sólo sirve para decidir cuándo hay que aplicar la optimización, sino que es propuesto como parte del método de resolución del problema.

A partir de este punto y como trabajo futuro, existen dos vías claras a explorar. La primera es el diseño e implementación del problema propuesto en el Capítulo 7 para validar su eficiencia en la mejora de redes ópticas dinámicas expuestas a una evolución del tráfico. En segundo lugar, la continuación de la obtención de datos con otras redes con el fin de tener datos suficientes como para afrontar la obtención de un único modelo válido para un conjunto amplio de topologías. En este sentido, el trabajo realizado en este proyecto ha allanado el camino para conseguir esta importante contribución.

8.2 Valoración personal

A nivel personal, la realización del presente proyecto ha implicado la necesidad de introducirme conceptualmente en el mundo de las redes ópticas. Así, entender los conceptos teóricos de las redes ópticas y de la temática del proyecto, ha sido fundamental a fin de conocer las variables consideradas en este proyecto. Las nociones recibidas fueron tanto a través de mis directores de proyecto como de algunos de los integrantes del Grupo de Comunicaciones Ópticas (GCO) de la UPC.

Por otro lado, el planteamiento inicial a resolver teniendo una variable endógena binomial, me hizo profundizar mis conocimientos estadísticos sobre los modelos logit ya vistos anteriormente a nivel académico. Especialmente, el cómo validarlos y hacer predicciones con ellos me hizo buscar información adicional. Además, debido a que el proceso de obtención de los datos se hizo de forma secuencial a lo largo de la evolución temporal del proyecto, eso me creó la necesidad de irme adaptando a los datos a medida que estos se iban obteniendo. El tiempo requerido para ello y la escasez de tiempo para la realización del proyecto conllevan una dedicación intensa.

A nivel más técnico, durante el PFC he utilizado varios entornos de software para distintos objetivos, lo que me ha permitido mejorar mis habilidades en esta materia. Así, he mejorado en el uso de R, fundamental para la aplicación de los modelos juntamente con su análisis, validación, etc. Además, he empezado a usar el software OMNeT de simulación de redes de comunicaciones para lanzar las simulaciones y recoger los resultados obtenidos para realizar los modelos.

8.3 Reconocimientos

Parte de este trabajo ha sido aceptado para ser presentado en el congreso de redes ópticas denominado International Conference on Transparent Optical Networks (ICTON) que se celebrará en Graz (Austria) en Julio de 2014. El título del trabajo a presentar (como ponencia invitada) es “Vulnerability Modelling for Periodical Flexgrid Network Planning”

Los métodos y resultados obtenidos en este proyecto serán incluidos como parte de la investigación realizada en el marco del proyecto europeo IDEALIST, subvencionado por el 7º Programa Marco 2007/2013 bajo el contrato nº 317999.

Apéndice A. Código implementado

```
#####  
## Modificación de las variables  
#####  
x300_<-read.csv2("F:/TFM Redes/Datos/BT_300_DATA_LINK_2.csv",header=T,dec=",")  
##### convertir en numérica variables  
x300_!$X.Grup1<-as.numeric(as.character(x300_!$X.Grup1))  
x300_!$X.Grup2<-as.numeric(as.character(x300_!$X.Grup2))  
x300_!$X.Grup3<-as.numeric(as.character(x300_!$X.Grup3))  
x300_!$avgLbetF<-as.numeric(as.character(x300_!$avgLbetF))  
x300_!$algConnectF<-as.numeric(as.character(x300_!$algConnectF))  
x300_!$algConnectVar<-as.numeric(as.character(x300_!$algConnectVar))  
x300_!$spectralRadiusVar<-as.numeric(as.character(x300_!$spectralRadiusVar))  
x300_!$mean_bw_demand<-as.numeric(as.character(x300_!$mean_bw_demand))  
x300_!$mean_path_length_before<-  
as.numeric(as.character(x300_!$mean_path_length_before))  
x300_!$mean_bw_path<-as.numeric(as.character(x300_!$mean_bw_path))  
x300_!$diameterVar<-as.numeric(as.character(x300_!$diameterVar))  
x300_!$radiusVar<-as.numeric(as.character(x300_!$radiusVar))  
x300_!$avgDegree<-as.numeric(as.character(x300_!$avgDegree))  
x300_!$spectralRadiusF<-as.numeric(as.character(x300_!$spectralRadiusF))  
#####  
##### Red TEL #####
```

```
#####
```

```
x520_l<-read.csv2("E:/TFM Redes/Datos/TEL_520_LINKDATA.csv",header=T,dec=".")
```

```
### crear variable log.bw_path_demand. de la red TEL
```

```
i= 1:5359
```

```
x520_l$log.bw_path_demand.[i]=log(x520_l$mean_bw_path[i]*x520_l$Total[i])/x520_l$algCo  
nnectVar[i]
```

```
x520_l$log.bw_path_demand.<-as.numeric(as.character(x520_l$log.bw_path_demand.))
```

```
#####
```

```
##### modelo nulo con las que no se han podido restaurar:
```

```
#####
```

```
mod_nul<-glm( restored.~0, family=binomial(link = "logit"),data=x520_l)
```

```
summary(mod_nul)
```

```
#####
```

```
##### agrupar variable log.bw_path_demand. #####
```

```
#####
```

```
install.packages("car")
```

```
library(car)
```

```
x520_l$log_bw_p_demand_1 <- recode(x520_l$log.bw_path_demand., "0:8.529999=1;  
8.530000:9.119999=2; 9.120000:9.560999=3; 9.561000:9.960999=4; 9.961000:14.000000=5;  
14.000001:16.000000=0")
```

```
summary(x520_l)
```

```
x520_l_0=subset(x520_l[x520_l$log_bw_p_demand_1=="1",c(1:24)])
```

```
x520_l_1=subset(x520_l[x520_l$log_bw_p_demand_1=="2",c(1:24)])
```

```
x520_l_2=subset(x520_l[x520_l$log_bw_p_demand_1=="3",c(1:24)])
```

b

```

x520_l_3=subset(x520_l[x520_l$log_bw_p_demand_1=="4",c(1:24)])
x520_l_4=subset(x520_l[x520_l$log_bw_p_demand_1=="5",c(1:24)])
x520_l_5=subset(x520_l[x520_l$log_bw_p_demand_1=="0",c(1:24)])

### Histogramas

### para Y=1 y Y=0

par(mfrow=c(3,2))

hist(res300_l$algConnectVar, main="conectividad algebraica VAR Y=1",xlab="algConnectVar")
hist(res300$algConnectVar, main="conectividad algebraica VAR Y=0",xlab="algConnectVar")
hist(res300_l$link_node_B, main="nodo B Y=1",xlab="link_node_B")
hist(res300$link_node_B, main="nodo B Y=0",xlab="link_node_B")
hist(res300_l$X.Grup2, main="Porcentaje Grup2 Y=1",xlab="% Grup2")
hist(res300$X.Grup2, main="Porcentaje Grup2 Y=0",xlab="% Grup2")
hist(res300_l$X.Grup3, main="Porcentaje Grup3 Y=1",xlab="% Grup3")
hist(res300$X.Grup3, main="Porcentaje Grup3 Y=0",xlab="% Grup3")
hist(res300_l$mean_bw_path, main="media bw_path Y=1",xlab="mean_bw_path")
hist(res300$mean_bw_path, main="media bw_path Y=0",xlab="mean_bw_path")
hist(res300_l$max_bw_path, main="valor máximo bw_path Y=1",xlab="max_bw_path")
hist(res300$max_bw_path, main="valor máximo bw_path Y=0",xlab="max_bw_path")

#####

##### Mirar modelo unico

#####

## M4

M3_25c<-glm(          restored.~0+algConnectF+max_bw_path+mean_bw_path+X.Grup3,
family=binomial, data=x520_l)

summary(M3_25c)

```

```
#### bondad de ajuste y sobredispersión del modelo
```

```
## estadístico de Pearson
```

```
M2pea<-sum(residuals(M3_25c,type="pearson")^2)
```

```
M2pea
```

```
1-pchisq(M2pea,M3_25c$df.residual)
```

```
#estadístico de devianza
```

```
1-pchisq(deviance(M3_25c),M3_25c$df.residual)
```

```
#### estimar parámetro de dispersión
```

```
escala<-(sum(residuals(M3_25c,type="pearson")^2))/(M3_25c$df.residual)
```

```
escala
```

```
##### PREDICCIÓN
```

```
x520_1<- x520_1[-c(1)]
```

```
datos_M1<-data.frame(x1=red_1_300$algConnectF, x2=red_1_300$max_bw_path,  
x3=red_1_300$mean_bw_path, x4=red_1_300$X.Grup3)
```

```
colnames(datos_M1) <- c("algConnectF", "max_bw_path", "mean_bw_path", "X.Grup3")
```

```
pred<- predict(M3_25c,type="response", newdata=datos_M1)
```

```
red_1_300$predict<-pred
```

```
summary(red_1_300)
```

```
### generar numero aleatorio
```

```
red_1_300$random<-runif(5359,0,1)
```

```
##### crear variable predict_1
```

```
i=1:5359
```

```
d
```

```

red_1_300$predict_1[i] <- red_1_300$predict[i] - red_1_300$random[i]

#### recodificar variable predict_1

for (i in 1:5359){

if (red_1_300$predict_1[i]>0) red_1_300$predict_1[i]=1

else red_1_300$predict_1[i]=0

}

### tabla de predicción

install.packages("gmodels")

library(gmodels)

CrossTable(red_1_300$restored.,red_1_300$predict_1)

#####

## Modelo libre

#####

M3_9d<-glm( restored.~0+X.Grup2+X.Grup1+max_bw_path+algConnectVar, family=binomial,
data=x520_l)

summary(M3_9d)

### bondad de ajuste y sobredispersión del modelo

## estadístico de Pearson

M2pea<-sum(residuals(M3_9d,type="pearson")^2)

1-pchisq(M2pea,M3_9d$df.residual)

#estadístico de devianza

1-pchisq(deviance(M3_9d),M3_9d$df.residual)

### estimar parámetro de dispersión

escala<-((sum(residuals(M3_9d,type="pearson")^2))/(M3_9d$df.residual)

```

escala

PREDICCIÓN

```
datos_M1<-data.frame(x1=red_2_300$X.Grup2, x2=red_2_300$X.Grup1,  
x3=red_2_300$max_bw_path, x4=red_2_300$algConnectVar)
```

```
colnames(datos_M1) <- c("X.Grup2", "X.Grup1", "max_bw_path", "algConnectVar")
```

```
pred<- predict(M3_9d,type="response", newdata=datos_M1)
```

```
red_2_300$predict<-pred
```

```
summary(red_2_300)
```

generar numero aleatorio

```
red_2_300$random<-runif(5359,0,1)
```

crear variable predict_1

```
i=1:5359
```

```
red_2_300$predict_1[i] <- red_2_300$predict[i] - red_2_300$random[i]
```

recodificar variable predict_1

```
for (i in 1:5359){
```

```
if (red_2_300$predict_1[i]>0) red_2_300$predict_1[i]=1
```

```
else red_2_300$predict_1[i]=0
```

```
}
```

```
CrossTable(red_2_300$restored.,red_2_300$predict_1)
```

anova modelos

```
anova(M3_25c,M3_9d,test="Chisq")
```

Curva de ROC

```
install.packages("Epi")
```

```
library(Epi)
```

```
ROC( form = restored. ~0+algConnectF+max_bw_path+mean_bw_path+X.Grup3, plot="ROC",  
PV=T, MX=T, AUC=T, data=x520_l, main='Gráfico a. Curva de ROC Modelo único TEL')
```

f

```

##### ACP

install.packages("amap")

library(amap)

## matriz correlaciones

a<-round(cor(x300_l),2)

#determinante matriz correlaciones

det(cor(x300_l))

acp=prcomp(x300_l,scale=TRUE)

acp

summary(acp)

C=cor(x300_l,predict(acp)[,1:3]) ## coge las primeras 3 componentes

##### ACP gráfico

par(mfrow=c(1,2))

(corvar <- acp$rotation %*% diag(acp$sdev))

plot(-1:1, -1:1, type='n', asp=1, xlab='CP1', ylab='CP2')

#abline(h=0, v=0, lty = 2, col=8)

symbols(0, 0, 1, inches=F, add=T)

## Dibuja los vectores y coloca los nombres

arrows(0, 0, corvar[,1], corvar[,2], length=.1)

text(corvar[,1], corvar[,2], colnames(x300_l), pos=2, offset=.8,col=3, font=1)

(corvar <- acp$rotation %*% diag(acp$sdev))

plot(-1:1, -1:1, type='n', asp=1, xlab='CP1', ylab='CP2')

abline(h=0, v=0, lty = 2, col=8)

## Dibuja los vectores y coloca los nombres

arrows(0, 0, corvar[,1], corvar[,2], length=.1)

```

```
text(corvar[,1], corvar[,2], colnames(x300_l), pos=2, offset=.8,col=3, font=1)
```

```
#####
```

```
##### Buscar modelo por categorías
```

```
#####
```

```
## Una variable
```

```
spectralRadiusVar
```

```
M3_1<-glm( restored.~0+spectralRadiusVar, family=binomial, data=x520_l_2)
```

```
summary(M3_1)
```

```
## avgDegree
```

```
M3_2<-glm( restored.~0+avgDegree, family=binomial, data=x520_l_2)
```

```
summary(M3_2)
```

```
## diameterF
```

```
M3_8<-glm( restored.~0+diameterF, family=binomial, data=x520_l_2)
```

```
summary(M3_8)
```

```
## avgLbetF
```

```
M3_9<-glm( restored.~0+avgLbetF, family=binomial, data=x520_l_2)
```

```
summary(M3_9)
```

```
## spectralRadiusF
```

```
M3_10<-glm( restored.~0+spectralRadiusF, family=binomial, data=x520_l_2)
```

```
summary(M3_10)
```

```
## radiusVar
```

```
M3_12<-glm( restored.~0+radiusVar, family=binomial, data=x520_l_2)
```

```
summary(M3_12)
```

```
## diameterVar
```

```
M3_13<-glm( restored.~0+diameterVar, family=binomial, data=x520_l_2)
```

```
summary(M3_13)
```

```
## algConnectVar
```

```
h
```



```
M3_15<-glm( restored.~0+algConnectVar, family=binomial, data=x520_I_2)
summary(M3_15)
## link_node_A
M3_16<-glm( restored.~0+link_node_A, family=binomial, data=x520_I_2)
summary(M3_16)
## link_node_B
M3_17<-glm( restored.~0+link_node_B, family=binomial, data=x520_I_2)
summary(M3_17)
## Grup1
M3_18<-glm( restored.~0+Grup1, family=binomial, data=x520_I_2)
summary(M3_18)
## Grup3
M3_19<-glm( restored.~0+Grup3, family=binomial, data=x520_I_2)
summary(M3_19)
## X.Grup1
M3_20<-glm( restored.~0+X.Grup1, family=binomial, data=x520_I_2)
summary(M3_20)
## X.Grup2
M3_21<-glm( restored.~0+X.Grup2, family=binomial, data=x520_I_2)
summary(M3_21)
## X.Grup3
M3_22<-glm( restored.~0+X.Grup3, family=binomial, data=x520_I_2)
summary(M3_22)
## min_path_length_before
M3_23<-glm( restored.~0+min_path_length_before, family=binomial, data=x520_I_2)
summary(M3_23)
## max_path_length_before
```

```
M3_24<-glm( restored.~0+max_path_length_before, family=binomial, data=x520_l_2)
summary(M3_24)
## max_bw_path

M3_25<-glm( restored.~0+max_bw_path, family=binomial, data=x520_l_2)
summary(M3_25)
## mean_bw_path

M3_26<-glm( restored.~0+mean_bw_path, family=binomial, data=x520_l_2)
summary(M3_26)
## mean_path_length_before

M3_27<-glm( restored.~0+mean_path_length_before, family=binomial, data=x520_l_2)
summary(M3_27)
## Grup2

M3_28<-glm( restored.~0+Grup2, family=binomial, data=x520_l_2)
summary(M3_28)
## min_bw_path

M3_29<-glm( restored.~0+min_bw_path, family=binomial, data=x520_l_2)
summary(M3_29)
## algConnectF

M3_30<-glm( restored.~0+algConnectF, family=binomial, data=x520_l_2)
summary(M3_30)
## min_bw_path

M3_31<-glm( restored.~0+mean_bw_path, family=binomial, data=x520_l_2)
summary(M3_31)

#####

##### 2 variables

#####

##spectralRadiusVar

j
```

```
M3_1a<-glm( restored.~0+Grup2+spectralRadiusVar, family=binomial, data=x520_I_2)
```

```
summary(M3_1a)
```

```
## algConnectF
```

```
M3_21a<-glm( restored.~0+Grup2+algConnectF, family=binomial, data=x520_I_2)
```

```
summary(M3_21a)
```

```
## avgDegree
```

```
M3_2a<-glm( restored.~0+Grup2+avgDegree, family=binomial, data=x520_I_2)
```

```
summary(M3_2a)
```

```
## avgLbetF
```

```
M3_9a<-glm( restored.~0+Grup2+avgLbetF, family=binomial, data=x520_I_2)
```

```
summary(M3_9a)
```

```
## spectralRadiusF
```

```
M3_10a<-glm( restored.~0+Grup2+spectralRadiusF, family=binomial, data=x520_I_2)
```

```
summary(M3_10a)
```

```
## radiusVar
```

```
M3_12a<-glm( restored.~0+Grup2+radiusVar, family=binomial, data=x520_I_2)
```

```
summary(M3_12a)
```

```
## diameterVar
```

```
M3_13a<-glm( restored.~0+Grup2+diameterVar, family=binomial, data=x520_I_2)
```

```
summary(M3_13a)
```

```
## algConnectVar
```

```
M3_15a<-glm( restored.~0+Grup2+algConnectVar, family=binomial, data=x520_I_2)
```

```
summary(M3_15a)
```

```
## link_node_A
```

```
M3_16a<-glm( restored.~0+Grup2+link_node_A, family=binomial, data=x520_I_2)
```

```
summary(M3_16a)
```

```
## link_node_B
```

```
M3_17a<-glm( restored.~0+Grup2+link_node_B, family=binomial, data=x520_l_2)
summary(M3_17a)
## Grup1
M3_18a<-glm( restored.~0+Grup2+Grup1, family=binomial, data=x520_l_2)
summary(M3_18a)
## Grup3
M3_19a<-glm( restored.~0+Grup2+Grup3, family=binomial, data=x520_l_2)
summary(M3_19a)
## X.Grup1
M3_20a<-glm( restored.~0+Grup2+X.Grup1, family=binomial, data=x520_l_2)
summary(M3_20a)
## X.Grup3
M3_22a<-glm( restored.~0+Grup2+X.Grup3, family=binomial, data=x520_l_2)
summary(M3_22a)
## min_path_length_before
M3_23a<-glm( restored.~0+Grup2+min_path_length_before, family=binomial, data=x520_l_2)
summary(M3_23a)
## max_path_length_before
M3_24a<-glm( restored.~0+Grup2+max_path_length_before, family=binomial, data=x520_l_2)
summary(M3_24a)
## max_bw_path
M3_25a<-glm( restored.~0+Grup2+max_bw_path, family=binomial, data=x520_l_2)
summary(M3_25a)
## mean_bw_path
M3_26a<-glm( restored.~0+Grup2+mean_bw_path, family=binomial, data=x520_l_2)
summary(M3_26a)
## mean_path_length_before
```

```

M3_27a<-glm( restored.~0+Grup2+mean_path_length_before, family=binomial,
data=x520_l_2)

summary(M3_27a)

## X.Grup2

M3_28a<-glm( restored.~0+Grup2+X.Grup2, family=binomial, data=x520_l_2)

summary(M3_28a)

## min_bw_path

M3_29a<-glm( restored.~0+Grup2+min_bw_path, family=binomial, data=x520_l_2)

summary(M3_29a)

#####

# 3 variables

##spectralRadiusVar

M3_1b<-glm( restored.~0+Grup2+link_node_A+spectralRadiusVar, family=binomial,
data=x520_l_2)

summary(M3_1b)

## avgDegree

M3_2b<-glm( restored.~0+Grup2+link_node_A+avgDegree, family=binomial, data=x520_l_2)

summary(M3_2b)

## max_bw_path

M3_4b<-glm( restored.~0+Grup2+link_node_A+max_bw_path, family=binomial,
data=x520_l_2)

summary(M3_4b)

## radiusVar

M3_12b<-glm( restored.~0+Grup2+link_node_A+radiusVar, family=binomial, data=x520_l_2)

summary(M3_12b)

## diameterVar

M3_13b<-glm( restored.~0+Grup2+link_node_A+diameterVar, family=binomial,
data=x520_l_2)

summary(M3_13b)

```

```
## algConnectVar
```

```
M3_15b<-glm( restored.~0+Grup2+link_node_A+algConnectVar, family=binomial,  
data=x520_l_2)
```

```
summary(M3_15b)
```

```
## Grup3
```

```
M3_16b<-glm( restored.~0+Grup2+link_node_A+Grup3, family=binomial, data=x520_l_2)
```

```
summary(M3_16b)
```

```
## link_node_B
```

```
M3_17b<-glm( restored.~0+Grup2+link_node_A+link_node_B, family=binomial,  
data=x520_l_2)
```

```
summary(M3_17b)
```

```
## Grup1
```

```
M3_18b<-glm( restored.~0+Grup2+link_node_A+Grup1, family=binomial, data=x520_l_2)
```

```
summary(M3_18b)
```

```
## X.Grup1
```

```
M3_20b<-glm( restored.~0+Grup2+link_node_A+X.Grup1, family=binomial, data=x520_l_2)
```

```
summary(M3_20b)
```

```
## algConnectF
```

```
M3_21b<-glm( restored.~0+Grup2+link_node_A+algConnectF, family=binomial,  
data=x520_l_2)
```

```
summary(M3_21b)
```

```
## X.Grup3
```

```
M3_22b<-glm( restored.~0+Grup2+link_node_A+X.Grup3, family=binomial, data=x520_l_2)
```

```
summary(M3_22b)
```

```
## min_path_length_before
```

```
M3_23b<-glm( restored.~0+Grup2+link_node_A+min_path_length_before, family=binomial,  
data=x520_l_2)
```

```
summary(M3_23b)
```

```
## max_path_length_before
```

```
n
```

```
M3_24b<-glm( restored.~0+Grup2+link_node_A+max_path_length_before, family=binomial,
data=x520_l_2)
```

```
summary(M3_24b)
```

```
## mean_bw_path
```

```
M3_25b<-glm( restored.~0+Grup2+link_node_A+mean_bw_path, family=binomial,
data=x520_l_2)
```

```
summary(M3_25b)
```

```
# mean_path_length_before
```

```
M3_26b<-glm( restored.~0+Grup2+link_node_A+mean_path_length_before, family=binomial,
data=x520_l_2)
```

```
summary(M3_26b)
```

```
## X.Grup2
```

```
M3_27b<-glm( restored.~0+Grup2+link_node_A+X.Grup2, family=binomial, data=x520_l_2)
```

```
summary(M3_27b)
```

```
## min_bw_path
```

```
M3_28b<-glm( restored.~0+Grup2+link_node_A+min_bw_path, family=binomial,
data=x520_l_2)
```

```
summary(M3_28b)
```

```
#####
```

```
### 4 variables
```

```
#####
```

```
##spectralRadiusVar
```

```
M3_1c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+spectralRadiusVar,
family=binomial, data=x520_l_2)
```

```
summary(M3_1c)
```

```
## avgLbetF
```

```
M3_9c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+avgLbetF, family=binomial,
data=x520_l_2)
```

```
summary(M3_9c)
```

```
## spectralRadiusF
```

```
M3_10c<-glm(          restored.~0+Grup2+link_node_A+max_bw_path+spectralRadiusF,  
family=binomial, data=x520_l_2)
```

```
summary(M3_10c)
```

```
## radiusVar
```

```
M3_12c<-glm(  restored.~0+Grup2+link_node_A+max_bw_path+radiusVar,  family=binomial,  
data=x520_l_2)
```

```
summary(M3_12c)
```

```
## diameterVar
```

```
M3_13c<-glm(          restored.~0+Grup2+link_node_A+max_bw_path+diameterVar,  
family=binomial, data=x520_l_2)
```

```
summary(M3_13c)
```

```
## algConnectVar
```

```
M3_15c<-glm(          restored.~0+Grup2+link_node_A+max_bw_path+algConnectVar,  
family=binomial, data=x520_l_2)
```

```
summary(M3_15c)
```

```
## Grup3
```

```
M3_16c<-glm(  restored.~0+Grup2+link_node_A+max_bw_path+Grup3,  family=binomial,  
data=x520_l_2)
```

```
summary(M3_16c)
```

```
## link_node_B
```

```
M3_17c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+link_node_B, family=binomial,  
data=x520_l_2)
```

```
summary(M3_17c)
```

```
## Grup1
```

```
M3_18c<-glm(  restored.~0+Grup2+link_node_A+max_bw_path+Grup1,  family=binomial,  
data=x520_l_2)
```

```
summary(M3_18c)
```

```
## X.Grup1
```

```
M3_20c<-glm(  restored.~0+Grup2+link_node_A+max_bw_path+X.Grup1,  family=binomial,  
data=x520_l_2)
```

```
summary(M3_20c)
```

p


```
## X.Grup2
```

```
M3_21c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+X.Grup2, family=binomial,  
data=x520_l_2)
```

```
summary(M3_21c)
```

```
## X.Grup3
```

```
M3_22c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+X.Grup3, family=binomial,  
data=x520_l_2)
```

```
summary(M3_22c)
```

```
## min_path_length_before
```

```
M3_23c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+min_path_length_before,  
family=binomial, data=x520_l_2)
```

```
summary(M3_23c)
```

```
## max_path_length_before
```

```
M3_24c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+max_path_length_before,  
family=binomial, data=x520_l_2)
```

```
summary(M3_24c)
```

```
## mean_bw_path
```

```
M3_25c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+mean_bw_path,  
family=binomial, data=x520_l_2)
```

```
summary(M3_25c)
```

```
## X.Grup2
```

```
M3_27c<-glm( restored.~0+Grup2+link_node_A+max_bw_path+X.Grup2, family=binomial,  
data=x520_l_2)
```


Bibliografia

- [Ca10] L. Cayuela, Modelos Lineales Generalizados, Universidad de Granada, 2010 <http://158.49.96.73:8080/documenta/bitstream/00000001/24/1/3-Modelos%20lineales%20generalizados.pdf>.
- [Ca12] A. Castro, L. Velasco, M. Ruiz, M. Klinkowski, J. P. Fernández-Palacios, and D. Careglio, "Dynamic Routing and Spectrum (Re)Allocation in Future Flexgrid Optical Networks," Elsevier Computers Networks, vol. 56, pp. 2869-2883, 2012.
- [Ca14] A. Castro, L. Velasco, J. Comellas, and G. Junyent, "On the benefits of Multi-path Recovery in Flexgrid Optical Networks," accepted in Springer Photonic Network Communications, 2014.
- [Ch54] Chernoff, H.; Lehmann E.L. "The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit," The Annals of Mathematical Statistics 25: pp. 579–586, 1954.
- [Ge14] Generalitat de Catalunya. *Actualització del Programa d'identificació visual del Servei Català de la Salut. [en elaboració]*, 2014
- [Ha01] Harrell, Frank E., Jr. Regression modeling strategies. New York: Springer-Verlag; Section 9.3.1., 2001.
- [Hi74] Akaike, Hitotug. A new look at the statistical model identification. *IEEE Transactions on Automatic; Control* 19 (6): 716–723, 1974.
- [ITU05] Telecom Development Bureau, Teletraffic Engineering Handbook, ITU, 2005.
- [Li11] Y. Li, F. Zhang, and R. Casellas. Flexible grid label format in wavelength switched optical network; IETF RFC Draft, Jul 2011.
- [Ma92] Martínez de Sousa, J. Diccionario de la tipografía y el libro. 3a ed. Madrid: Paraninfo, 1992.
- [Me03] E. Medina, Modelos de elección discreta. http://www.uam.es/personal_pdi/economicas/eva/pdf/logit.pdf., 2013
- [Me95] J.M. Mestres et al., *Manual d'estil: la redacció i l'edició de textos*. Barcelona: Eumo Editorial; Universitat de Barcelona; Universitat Pompeu Fabra; Associació de Mestres Rosa Sensat, 1995.
- [OMNeT] OMNeT++. <http://www.omnetpp.org/>
- [Ru14] M. Ruiz and L.Velasco, "Vulnerability Modelling for Periodical Flexgrid Network Planning," Invited paper in 16th International Conference on Transparent Networks, 2014.

- [Sa13] A. Salazar, *Modelos de respuesta discreta en R y aplicación con datos reales. Máster oficial en Estadística Aplicada*. Universidad de Granada. <file:///C:/Users/Noem%C3%AD/Downloads/Modelos%20de%20respuesta%20discreta%20en%20R%20y%20aplicaci%C3%B3n.pdf>.
- [Sw96] J. Swets, "Signal detection theory and ROC analysis in psychology and diagnostics". Collected papers, Lawrence Erlbaum Associates, Mahwah; NJ, 1996
- [Ve09] L. Velasco, "Recovery Mechanisms in ASON/GMPLS Networks," PhD Thesis, UPC, 2009.
- [Vm11] P. Van Mieghem, *Graph Spectra of Complex Networks*, Cambridge University Press, 2011.