DOCUMENTACIÓN TÉCNICA septiembre | 2009

Metodología de Análisis de Variables de Influencia para los indicadores del CMBD del SNS. Una técnica genérica basada en Redes Bayesianas

Dr. José Manuel Gutiérrez
Instituto de Física de Cantabria (IFCA)
CSIC - Universidad de Cantabria (UC)
gutierjm@unican.es

Dr. Antonio S. Cofiño
Dpto. de Matemática Aplicada y C.C.
Universidad de Cantabria (UC)
cofinoa@unican.es

CONTENIDOS:

RESUMEN: En este informe se describe la metodología propuesta para el análisis de dependencias entre los indicadores y las variables de influencia (o factores) del CMBD, de cara a definir modelos de ajustes por riesgo apropiados para los indicadores (*ver documentación técnica sobre metodología de ajuste*). Todos los indicadores y variables consideradas son discretas, salvo la complejidad (factor) y la estancia (indicador), para las cuales se consideran sus transformadas normalizadas (según Box-Cox) descritas en el informe descriptivo de los factores e indicadores (ver documentación técnica). En un primer análisis se concluye que la aproximación normal no es apropiada en este problema ni para los factores (complejidad) ni para los indicadores (estancias). Por tanto, se aborda el estudio de dependencias considerando las variables discretizadas (*ver documentación técnica sobre análisis descriptivo de variables de influencia e indicadores*); en este caso se aplica el test estándar de diferencias de proporciones (para los indicadores en forma de tasa) y los test no paramétricos U de Mann-Whitney (paralelo a la prueba t de Student) y el de Kruskal-Wallis (paralelo al ANOVA de un factor) para las diferencias de estadísticos promedio (estancias). La conclusión de este estudio es que, dado el gran tamaño de la muestra, todas las relaciones son significativas y, por tanto, es necesario cuantificar la magnitud relativa de las mismas en base a la variación que producen en el indicador para poder valorar su relevancia práctica. Para ello, se utilizan modelos gráficos probabilísticos (redes Bayesianas) para definir una probabilidad conjunta de todas las variables (indicador junto con las variables de influencia) que contenga todas las dependencias relevantes del problema para la población disponible. Esta probabilidad permite cuantificar las dependencias en función de la variación que producen.

Para una descripción detallada de los distintos test estadísticos utilizados en este trabajo, se remite al lector por ejemplo, a G.J. Hahn y W.Q. Meeker. 1991. *Statistical Intervals. A Guide for Practicioners*. Wiley Interscience.

1. Introducción

El registro de altas hospitalarias, también conocido como CMBD (Conjunto Mínimo Básico de Datos al alta), constituye la mayor base de datos administrativa sobre pacientes hospitalizados (casi 27 millones de registros, a razón de 3,5 millones/año aproximadamente), siendo la principal fuente de información sobre morbilidad atendida, con información muy valiosa sobre múltiples aspectos de la actividad hospitalaria, incluyendo la calidad y variabilidad de la práctica asistencial.

1. Estancia Media (1)
2. Estancia Media Preoperatoria (1)
3. Tasa de Mortalidad (17)
4. Tasa de Reingresos (1)
5. Tasa de Infección Nosocomial (1)
6. Tasa de Cesáreas (1)
7. Tasa de Complicaciones (15)
8. Tasa de Ambulatorización Quirúrgica (2)
9. Frecuentación en Hospitalización (1)
10. Tasa de Realización (11)
Tabla 1. Conjunto de indicadores

Esta base de datos recopila la información de 283 hospitales del Sistema Nacional de Salud (SNS).

El Ministerio de Sanidad y Política Social es responsable de la gestión del CMBD estatal, generando con periodicidad anual diversas estadísticas oficiales. Recientemente, se ha desarrollado un modelo de explotación de esta información basado en un conjunto reducido de indicadores de máximo valor explicativo que permita profundizar en el análisis de las características de la atención hospitalaria de los pacientes ingresados en el Sistema Nacional de Salud. Este modelo se basa en 51 indicadores, agrupados en diez familias genéricas mostradas en la Tabla 1 (el número de indicadores en cada familia se muestra entre paréntesis; ver documento de *"descripción del modelo de indicadores"* para más información). Algunos de estos indicadores se refieren a resultados continuos, como número de días, mientras que otros se refieren a resultados discretos, como ocurrencias.

Para poder comparar el valor de los indicadores en los distintos hospitales primero es necesario corregir, o ajustar, los resultados brutos teniendo en cuenta las distintas casuísticas de los pacientes tratados en los distintos hospitales (case-mix), considerando así la complejidad de los servicios prestados en cada caso (*ver documentación técnica sobre metodología de ajuste para más detalles*).

RELACIONADOS CON LA ENFERMEDAD
1. Complejidad (peso español GRD-AP v18)
2. Severidad (GRD refinados)
3. Riesgo de Mortalidad, ROM
4. Cat. Diagnostica Mayor (GRD-AP v18)
5. Tipo de GRD: médico o quirúrgico
PACIENTE / HOSPITAL
6. Edad
7. Sexo
8. Tipo de ingreso
9. Tipo de alta
10. Tipo de hospital
11. Edad de la madre (sólo para Tasa de Cesáreas)
Tabla 2. Factores de riesgo/influencia

El CMBD incluye un conjunto genérico de variables de influencia, o factores de riesgo, que influyen en esta casuística y están relacionados tanto con la enfermedad y su diagnóstico (basados en GRDs), como con el paciente y con el funcionamiento hospitalario (Tabla 2).

Con la excepción de la complejidad, todas las variables de influencia son discretas, o se utilizan en una variante discretizada (como los grupos de edad).

Estas variables se han utilizado en el proyecto para llevar a cabo una primera corrección genérica de los indicadores considerando modelos gráficos probabilísticos (redes Bayesianas) que permiten estimar las variaciones que producen los factores de riesgo en un indicador dado; así, se puede ajustar el valor de los distintos hospitales según la casuística particular de los factores de riesgo en cada uno de ellos. Sin embargo, antes de construir los modelos de ajuste es necesario analizar cuales de los factores influencian en mayor medida el valor del indicador y sirven, por tanto, para crear modelos óptimos para realizar el ajuste.

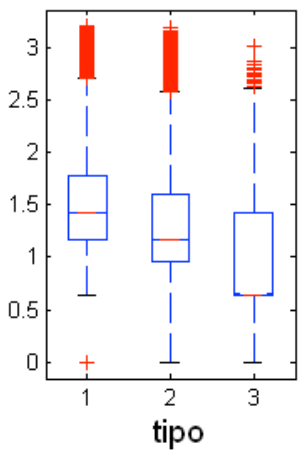
2. Metodología de Análisis: Contrastes de Hipótesis

La forma estándar de cuantificar el grado de dependencia o relación entre dos variables es a través de contrastes de hipótesis adecuados, donde se obtenga la significación estadística de la variación que produce una variable en la otra. En nuestro problema, se tiene una variable a predecir (el indicador) y otras predictoras (los factores de influencia), con el objetivo de obtener modelos apropiados de predicción del indicador en base a las variables de influencia relevantes en cada caso. Por tanto, se trata de determinar cuáles de las variables de influencia producen una variación significativa del valor del indicador, para ser así incluidas en el modelo.

Los contrastes de hipótesis permiten cuantificar el nivel de significación de estas relaciones, separando aquellas que pueden deberse al azar de aquellas que no tienen una explicación aleatoria y, por tanto, son el reflejo de alguna relación de dependencia entre las variables. El carácter discreto/continuo del indicador y de las variables de influencia determina el tipo concreto de test que es necesario aplicar. Los resultados mostrados en este informe corresponden a las altas del año 2005, pero los resultados son idénticos para el resto de años (2004-2007), con variaciones casi idénticas de gráficos y resultados.

2.1. CASO CONTINUO

En primer lugar se consideran indicador *estancia media* (EM) en su forma continua y se analizan las dependencias de variables de influencia discretas utilizando técnicas de análisis de la varianza (ANOVA) que son las apropiadas en este caso. En primer lugar se considera la transformada normalizada de EM para el análisis, según la transformación de Box-Cox (*ver documentación técnica, "análisis descriptivo de indicadores"*). A modo ilustrativo, se considera la variable de influencia *Tipo de ingreso*, aplicando un Análisis de la Varianza (ANOVA) estándar de un factor de efectos fijos y completamente aleatorizados.



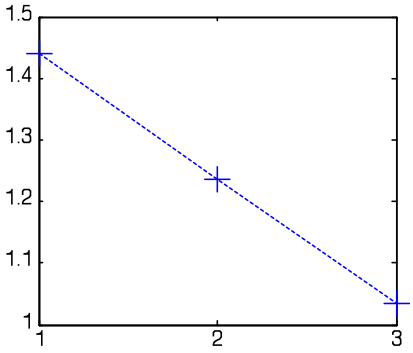
Esta técnica permite contrastar la diferencia de las medias de *EM* para los distintos grupos en que las categorías del factor separan a la población. Por ejemplo, la figura izquierda muestra, mediante diagramas de cajas, las distribuciones de valores de estancia media transformados Box-Cox para los distintos estados de *tipo de ingreso* (1-urgentes, 2-programados y 3-otros).

Se considera la hipótesis nula H_0: $m_1 = m_2 = ... = m_k$ (donde m_i es el valor de estancia media para la muestra de la población en la que la variable de influencia toma el valor *k*) y se aplica ANOVA para estudiar su significancia. En la siguiente figura se muestran los diagramas de cajas de los valores *EM* para los distintos valores del factor *tipo de ingreso* (con tres categorías, *ver documentación técnica "análisis descriptivo de*

variables de influencia"), así como una comparación múltiple de las diferencias de medias entre los distintos grupos utilizando como valor crítico el mínimo de los índices dados por los tests de tukey-kramer, bonferroni y scheffé. En todos los casos la significancia, *pvalor,* es muy alta (valores inferiores a 0,0001) debido al gran número de datos disponibles en la población.

Tabla ANOVA para *Estancia Media* con factor tipo de ingreso

Source	***SS***	***grad. libertad***	***MS***	***F***	***pvalor***
Groups	32186,04	2	16093,02	70430,30	<0,0001
Error	798874,68	3496235	0,22		
Total	831060,72	3496237			

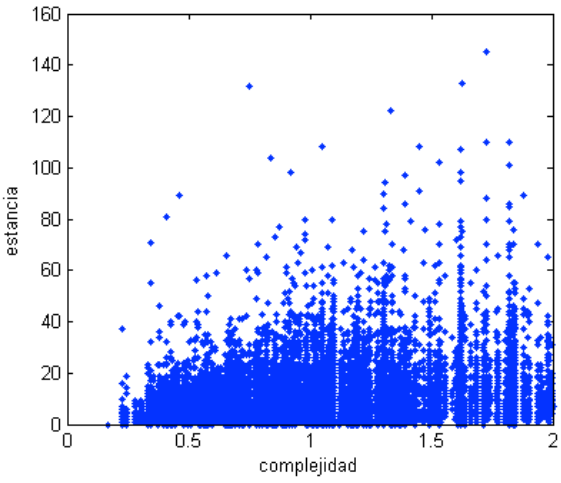


Ingresos	*T(EM)*	*EM*
1-Urgentes	1,4275 (s=3 10^{-4})	**5,0412**
2-Programados	1,2321 (s=5 10^{-4})	**3,4214**
3-Otros	1,0432 (s=5 10^{-3})	**2,3421**

Comparación múltiple de diferencia de medias:

Valores	*diferencia*	*Intervalo confianza (99%)*
1-2	0,1954	(0,202,0,205)
1-3	0,3843	(0,370,0,408)
2-3	0,1889	(0,176, 0198)

donde T(EM) es el valor de la estancia media transformado según Box-Cox y EM es el valor correspondiente anti-transformado. Aunque los análisis realizados ponen de manifiesto las diferencias significativas entre los valores medios de los distintos grupos (correspondiente a distintos valores de las variables de influencia), se observa que los valores de estancia media transformados a la escala original no muestran unos valores razonables (si se comparan, por ejemplo, con los obtenidos directamente de los datos originales). Este hecho pone de manifiesto que la transformación de Box-Cox no permite aproximar de forma razonable la distribución de la estancia media, que tiene dos regímenes distintos, para valores de estancia inferiores a 20 días, y para valores superiores. Esta falta de normalidad en la variable transformada hace que los análisis estándar de ANOVA y los modelos lineales generalizados resulten inapropiados para este problema.



En el caso del factor de influencia *complejidad*, considerado como variable continua transformada según Box-Cox, se ve que existe una relación con la *estancia media* para valores altos de ésta; sin embargo, esta relación es muy ruidosa como se muestra en la figura izquierda y no permite obtener conclusiones razonables a partir de un análisis de regresión.

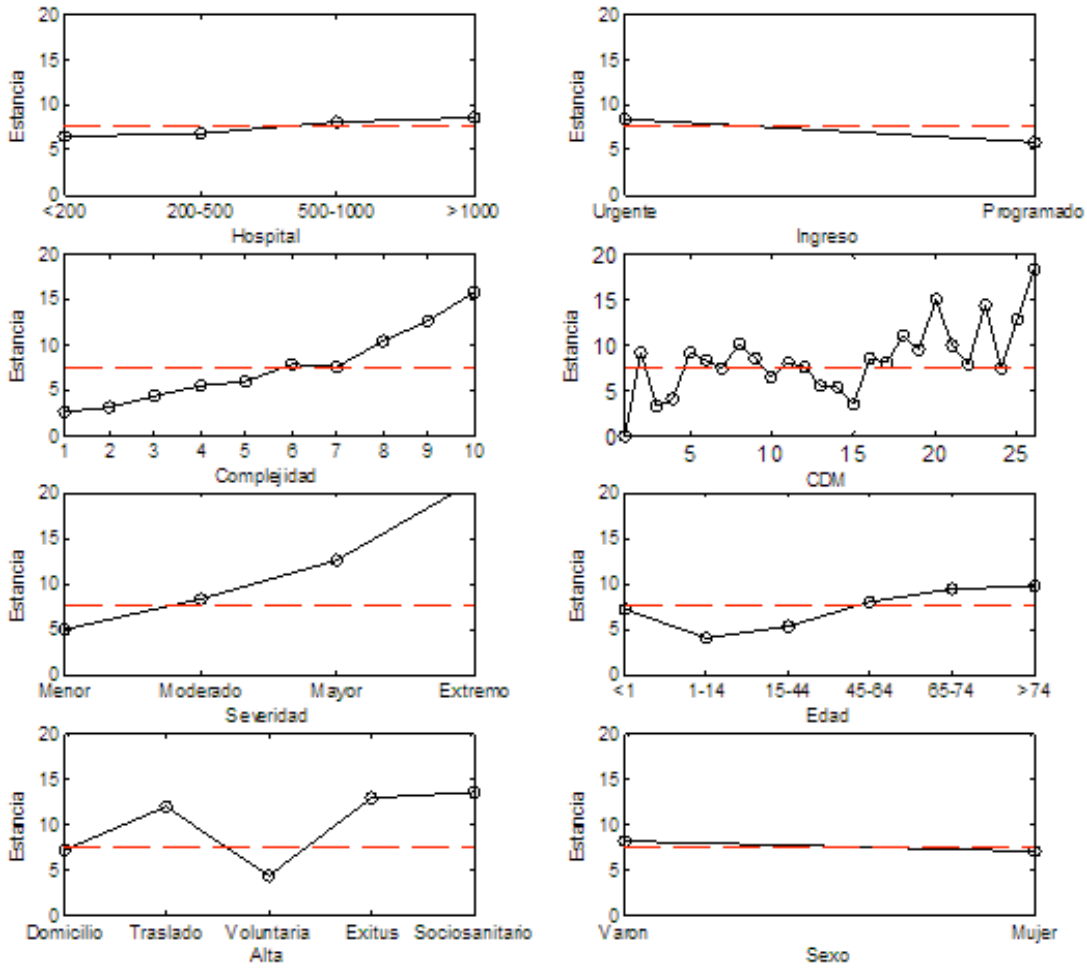
Estos resultados motivaron plantear el problema utilizando un modelo discreto, discretizando previamente las variables continuas.

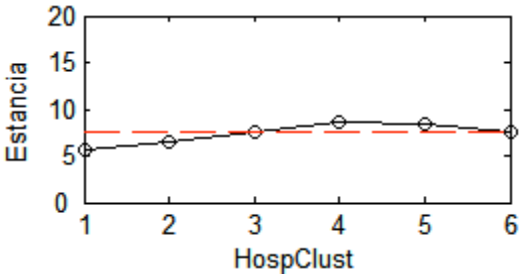
2.2. CASO DISCRETO

En el caso discreto, todas las variables son consideradas como variables discretas, lo que facilita y simplifica el análisis, evitando transformaciones intermedias. Por tanto, estancia y complejidad son discretizadas en un número suficiente de estados para dar cuenta de su distribución de valores (*ver documentación técnica "análisis descriptivo de indicadores y factores"*). De esa manera se tiene que *X* posee *k* estados distintos *1, 2, ..., k*, cada uno asociado a un cierto intervalo de estancias, y con marcas de clase $x_1, x_2,.....,x_k$ (la media de valores del intervalo). Así, de la misma forma que en el caso anterior, se puede calcular la media de la estancia condicionada a cada una de las categorías de la variable indicador. Por ejemplo, se puede calcular la probabilidad a posteriori de las distintas categorías de la *estancia* dado que el ingreso ha sido urgente: *P(EM=x | ingreso=urgente).* Estas probabilidades permiten calcular el valor medio de la estancia para los ingresos urgentes mediante

$$\Sigma_x \, x^*P(EM=x \mid ingreso=urgente).$$

De esta manera se han obtenido las *estancias* medias que se muestran en la siguiente figura, para todos los posibles estados de los factores de influencia mostrados en la Tabla 2. La línea roja indica el valor medio del total de altas (indicador no condicionado).





Esta figura muestra que existe una fuerte relación entre la *complejidad* y la *estancia media*; además esta relación es creciente de forma aproximadamente lineal (un aumento de complejidad produce un aumento proporcional del tiempo medio de estancia. También existe una relación similar entre la *severidad* y la *estancia* (en particular, esta factor produce la mayor variación de la estancia media, asociada con la severidad extrema). Por otra parte, existen relaciones no lineales entre el *tipo de alta* y la *estancia*, así como entre el CDM y también edad. La variabilidad de los restantes factores es de menor magnitud, aunque pueden ser significativas.

Para comprobar si estas diferencias de estancias medias para distintos factores son o no significativas se ha aplicando una prueba t de Student (para comparar dos medias distintas) o un análisis de la varianza ANOVA (para comparar las medias asociadas a distintos valores de un factor). Todas las diferencias obtenidas con estas pruebas paramétricas son significativas con pvalor inferior a 0,0001. En este caso, también se ha aplicado el test de Kruskal-Wallis para comprobar la hipótesis de la diferencia de medianas entre grupos distintos de la población (correspondientes a distintos variables de los factores de influencia). Este test es una versión no paramétrica del análisis de la varianza para comparar la diferencia entre estadísticos centrales asociados a grupos distintos de la población y se basa en la distribución chi-cuadrado. En este caso también se obtuvieron significaciones superiores al 99,99% (pvalor inferior a 0,0001). Es normal que con tamaños de muestras tan grandes, las diferencias resulten significativas, aunque estas diferencias puedan corresponden a variaciones que no tienen importancia desde el punto de vista práctico. Por ejemplo, al aplicar el test a los dos grupos dados por la variable *sexo (varón y mujer)*, nos encontramos con el siguiente resultado:

Origen	***SS***	***df***	***MS***	***chi-sq***	***p-valor***
Grupos	2,455e+016	1	2,45504e+016	24348,14	<0.0001
Error	3,499e+018	3495475	100128 e+012		
Total	3,524e+018	3495476			

Que muestra una diferencia significativa al 99,99%. La significación es tan alta que no vale la pena utilizar pruebas más precisas para abordar el problema en más detalle, sino que se hace necesario estudiar la relevancia práctica de las dependencias (para construir un modelo de ajuste de los indicadores) desde otro punto de vista. Por tanto, resulta fundamental cuantificar cuáles de estas dependencias son importantes para discriminar variaciones notables del indicador.

3. Metodología Basada en Redes Bayesianas

Para poder modelizar las dependencias entre los factores y el indicador es necesario analizar tanto las dependencias incondicionales como las dependencias condicionales, que involucran varios factores, ya que las influencias sobre el indicador pueden darse

a través de distintas combinaciones de los factores. Para ello es necesario trabajar con la probabilidad conjunta de todas las variables:

P(indicador,cdm,complejidad,sexo,edad,ingreso,hospital,alta,severidad) *(1)*

Sin embargo, la especificación completa de esta probabilidad conjunta requeriría más de 50 millones de parámetros (el producto de todos los posibles estados de las variables) y, por tanto, se hace necesario simplificar esta función eliminando aquellas dependencias que no sean relevantes en el conjunto de datos a estudiar y que permitan definir la probabilidad conjunta con un número menor de parámetros a través de una factorización adecuada. Para ello, se han considerado los modelos gráficos probabilísticos (en concreto las redes Bayesianas) como una metodología sólida y extendida para abordar este problema (ver por ejemplo Castilli, E. Hadi. A.S. y Gutiérrez, J.M., 1997: "Expert Systems and Probabilistic Network Models". Springer). Estos modelos utilizan un grafo dirigido acíclico para representar de forma cualitativa las dependencias entre variables (cada enlace en el grafo indica una dependencia directa entre variables y los distintos caminos, indican dependencias indirectas según un criterio conocido como criterio de d-separación, ver la referencia anterior para más detalles). El grafo resultante permite definir la probabilidad (1) utilizando una factorización de la misma como el producto de las probabilidades condicionadas de cada variable dados sus padres (variables que apuntan a la variable en cuestión) en el grafo. Esta factorización permite expresar la probabilidad conjunta de forma compacta, requiriendo un número reducido de parámetros. Además, la probabilidad conjunta resultante contiene todas las dependencias representadas por el grafo. Por tanto, con el modelo resultante se pueden calcular de forma eficiente las probabilidades marginales de las variables, o las condicionadas a una evidencia cualquiera. Estas probabilidades proporcionan toda la información necesaria para poder llevar a cabo los análisis de dependencias requeridos considerando un modelo global que será utilizado también más adelante para el ajuste de los indicadores utilizando los factores más relevantes (que serán dados como evidencia, o información conocida).

Por tanto, todo el análisis de este estudio se realiza con un mismo modelo conjunto que relaciona el indicador con los factores correspondientes.

Para definir las redes Bayesianas para cada indicador se ha procedido en dos etapas. Primero, dado que el indicador es la variable objetivo, se parte de una estructura de clasificador Bayes ingenuo (el indicador es padre de todos los factores). De esa forma, se garantiza que las dependencias directas entre el indicador y cada uno de los factores de influencia estarán caracterizadas de forma exacta en el modelo resultante. A continuación, aplicando una técnica de aprendizaje automático, se infieren las dependencias incondicionales y condicionales relevantes para el conjunto de datos, es decir el grafo más adecuado para representar las relaciones entre las variables. Para ello se han utilizado y comparado distintos algoritmos de aprendizaje automático, basados en una búsqueda iterativa del grafo óptimo. Finalmente, tras las pruebas realizadas (los resultados comparativos se explican en detalle para el primer indicador: la *estancia media*), se ha elegido un algoritmo voraz denominado "algoritmo B", que es el más potente y también el más costoso computacionalmente de los algoritmos comparados. Este algoritmo va añadiendo sucesivamente los enlaces que mejor explican globalmente los datos (redes de mejor calidad), penalizando modelos demasiado complejos que se pueden sobreajustar a los datos disponibles (una descripción completa de los métodos de aprendizaje utilizados puede consultarse en Castillo, E., Hadi, A.S. y Gutiérrez, J.M., 1997: "Expert Systems and Probabilistic Network Models". Springer).

El grafo resultante permite definir la probabilidad conjunta (1) a partir de un número reducido de parámetros, en particular a partir de las probabilidades de cada variable condicionada al conjunto de sus padres. La red Bayesiana resultante (grafo + factorización de la probabilidad) permite calcular la variación producida en un indicador por efecto de un cierto factor o conjunto de factores; para ello se calculan las probabilidades *P(indicador=x | factores=valores)* para los distintos estados *x* del indicador dados los distintos estados de los factores considerados (por ejemplo, *severidad = extrema*). Las diferencias de estas probabilidades con las probabilidades iniciales *P(indicador=x)* indican la variación que produce la combinación de factores correspondientes en el valor del indicador. Si el indicador es una proporción, dado por una variable binaria (por ejemplo, exitus=si o no), entonces la variación absoluta del indicador producida por los valores concretos asignados a los factores viene dada directamente por:

$$Variación = P(indicador=si \mid factores=valores) - P(indicador=si). \quad (2)$$

Si el indicador es un promedio (*estancia media* y *estancia media preoperatorio*), entonces la variación absoluta del valor del indicador debida a los valores concretos asignados a los factores puede obtenerse de la siguiente forma:

$$Variación = \Sigma_x x*P(indicador=x \mid factores= valores) - \Sigma_x x*P(indicador=x). \quad (3)$$

Obsérvese que el primer sumando en la ecuación anterior corresponde al valor medio condicionado a los casos compatibles con los valores indicados de los factores, mientras que el segundo sumando corresponde al valor medio del indicador en la población total. Por tanto, una vez aprendido el grafo, se puede cuantificar la variación que produce cada factor, y cada pareja de factores, en el valor del indicador correspondiente utilizando (2) o (3), según el indicador sea una proporción o un valor medio, respectivamente.

Esta metodología se aplicó a los distintos indicadores (Tabla 1, hasta el indicador 8, ya que los dos últimos son especiales y requieren un tratamiento específico), considerando las variables de influencia de la Tabla 2 (previo el filtrado de variables que son el efecto y no la causa del indicador). La siguiente tabla muestra las variables que han sido consideradas para cada indicador.

	Seve-ridad	ROM	Sexo	CDM	Edad	Comple-jidad	Tipo Ingreso	Hospital	Tipo Alta	Tipo GRD	Edad madre
Estancia	■	■	■	■	■	■	■	■	■		
Est. preop.	■	■	■	■	■	■	■	■			
Mortalidad	■		■	■	■	■	■	■			
Reingesos	■	■	■	■	■	■	■	■	■		
Infección	■	■	■	■	■	■	■	■		■	
Cesáreas								■			■
Complica.		■	■	■	■	■	■	■	■		
Ambulat.			■	■	■	■		■			

Tabla 4. Indicadores y factores de riesgo considerados en el ajuste

La aplicación para el análisis y explotación de altas del CMBD muestra las variables

de influencia y los estados que dan lugar a una variación máxima del indicador, aplicando las fórmulas (2) y (3), según corresponda.

4. Conclusiones

En este informe se describe la metodología aplicada para cuanticar las variables de influencia más relevantes para cada indicador, con el objeto de aplicarlas al ajuste de los indicadores. Dado el elevado número de datos disponibles, todas las relaciones entre indicadores y variables de influencia son significativas, sin que ello indique que las variables puedan tener un impacto práctico importante en el valor del indicador (es decir, incluso pequeñas variaciones resultan significativas). Además, la existencia de variables contínuas (estancia y complejidad) dificulta el análisis pues las distribuciones de las mismas son complejas y no se ajustan de forma significativa a una distribución normal, ni tampoco pueden ser transformadas a una distribución normal.

Por tanto, en este trabajo se discretizan el indicador y la variable de influencia continuas y se procede a realizar un análisis basado en la distribución conjunta de probabilidad (que es la que contiene todas las dependencias posibles entre variables). Para ello se utilizan modelos simplificados conocidos como redes Bayesianas, que permiten definir una distribución conjunta con un reducido número de parámetros, conservando todas las dependencias relevantes.

Los modelos resultantes son utilizados para cuantificar la variación bruta del indicador ante distintas situaciones (evidencias) dadas por los factores. Por tanto, este análisis permite determinar los factores y estados que explican una mayor variabilidad del indicador y pueden, por tanto, ser utilizados para ajustar los valores del mismo. Este análisis se presenta en la documentación técnica "*metodología de ajuste de indicadores*".