EL ANÁLISIS DE DATOS MULTIVARIADOS EN ACCIÓN: BUSCANDO PATRONES DE CONSUMO DE ALIMENTOS EN EUROPA OCCIDENTAL

[1,2]Robert A. Cazar, [3]Roberto Todeschini

[1]Escuela Superior Politécnica del Chimborazo, Facultad de Ciencias. Panamericana Sur Km 1 ½, Riobamba, Ecuador. [2]Grupo Ecuatoriano para el Estudio Experimental y Teórico de Nanosistemas, GETNano. [3]Dipartimento di Scienze dell'Ambiente e del Territorio e di Scienze della Terra, Università degli Studi di Milano - Bicocca, Italy

Resumen

Se detalla un análisis de datos multivariados que busca identificar patrones del consumo de alimentos en Europa Occidental. Un conjunto de datos que presenta la frecuencia de consumo de 17 productos alimenticios comunes en los hogares de 16 países europeos occidentales fue estudiado. Análisis de componentes principales, análisis de agrupamientos y algoritmos de clasificación fueron aplicados para identificar y caracterizar grupos de países con comportamiento similar respecto al consumo de estos alimentos. Los resultados muestran que los países en estudio se distribuyen en cuatro agrupamientos bien separados. Estos agrupamientos reflejan las similitudes en cultura, historia y tradición, así como la proximidad geográfica de los países incluidos en ellos. El objetivo principal del trabajo ha sido demostrar el potencial del análisis de datos multivariados para extraer información relevante de conjuntos de datos complejos.

Palabras claves: análisis de datos multivariados, consumo de alimentos, Europa Occidental

Abstract

A multivariate data analysis that seeks to identify patterns of food consumption in Western Europe is detailed. A data set that presents the frequency of consumption of 17 common food products in the households of 16 Western European countries was studied. Principal component analysis, cluster analysis and classification algorithms were applied to identify and characterize groups of countries with similar behavior regarding food consumption. The results show that the countries under study distribute in four well-separated clusters. Those clusters reflect the similarities in culture, history and traditions as well as the geographical proximity of the countries included in them. The main objective of this work has been to show the potential of multivariate data analysis to glean relevant information from complex data sets.

Keywords: multivariate data analysis, food consumption, Western Europe

INTRODUCCIÓN

El análisis de datos multivariados (MDA) comprende un arsenal de métodos estadísticos y matemáticos que asisten a un experimentador en la extracción de información a partir de conjuntos de datos complejos (1). El análisis de componentes principales (PCA), el análisis de agrupamientos y los algoritmos de clasificación se encuentran entre los métodos más versátiles y útiles del MDA. Estos ayudan, entre otras tareas, a localizar grupos de objetos con característica similares, reducir la dimensión de un conjunto de datos, y clasificar objetos en categorías previamente establecidas (2, 3, 4). En este trabajo, un conjunto de datos obte-

nido de una encuesta sobre la frecuencia de consumo de 17 alimentos comunes en los hogares de 16 países de Europa Occidental (5) ha sido analizado con tales métodos. Se busca identificar, dentro de estos países, grupos con comportamiento similar en cuanto al consumo de alimentos y se intenta explicar las razones detrás de tales comportamientos. Es esperable que tales grupos reflejen ciertas costumbres de alimentación que existen en Europa Occidental producto de la similitud en historia, hábitos y tradiciones de ciertos países. Entre estas se encuentra la dieta mediterránea –compartida por naciones como Italia, Portugal y España– que se caracteriza por un alto consumo de vegetales, legumbre, frutas, nueces y cereales, moderado consumo de pescado, alto consumo de grasas insaturadas (aceite de oliva) y un bajo consumo de grasas saturadas así como una ingesta baja a moderada de productos lácteos (6). La dieta del norte de Europa, que se practica en el Reino Unido de Gran Bretaña (Inglaterra, Escocia, Gales e Irlanda del Norte), la República de Irlanda y Francia, en cambio, consiste de una alta ingesta de carne o pescado y productos lácteos, bajo consumo de vegetales, cereales y frutas (7). Los países nórdicos, a su turno, consumen una dieta muy saludable que se basa en una alta ingesta de pescado, patatas, vegetales frescos y pan negro; así como frutas propias de la región (7). Finalmente, los países que ocupan la parte central del continente –como Alemania, Suiza y Bélgica– comparten una dieta en la que se consume principalmente patatas, vegetales y carne acompañado de una importante ingesta de diversas variedades de pan, café y productos dulces como mermeladas y pasteles (7).

En cuanto a trabajos similares al presentado en este artículo, se ha encontrado dos reportes publicados en 2009 y 2014, respectivamente, en los que se discuten aplicaciones del MDA para descubrir patrones de hábitos de consumo de alimentos en individuos de países europeos. El primero de ellos (8) se refiere a una aplicación del PCA para identificar patrones de consumo de alimentos en adolescentes griegos y asociarlos con su estilo de vida y condiciones socioeconómicas. El segundo (9) utiliza métodos de análisis de agrupamientos y algoritmos de clasificación sobre los datos producidos por una encuesta de consumo de alimentos en individuos jóvenes de Suecia y Dinamarca; el propósito del estudio es establecer diferencias en los patrones de consumo de estos dos grupos y evaluar la calidad de sus dietas. Como se puede ver estos trabajos involucran situaciones más específicas que la descrita en este.

MATERIALES Y MÉTODOS

Los datos provienen de una encuesta conducida hace unos años en 16 de países de Europa Occidental. En la encuesta, una muestra representativa de hogares de cada

País	Café molid.	Café instan.	té	Sopa sobre	Sopa lata	Papa	Pesca. cong.	manzana	naranja	fruta seca	mermela.	Ajo	mantequilla	margarina	Aceite oliva	yogur	Pan
Alemania	90	49	88	51	19	21	27	81	75	44	71	22	91	85	74	30	26
Italia	82	10	60	41	3	2	4	67	71	9	46	80	66	24	94	5	18
Francia	88	42	63	53	11	23	11	87	84	40	45	88	94	47	36	57	3
Holanda	96	62	98	67	43	7	14	83	89	61	81	15	31	97	13	53	15
Bélgica	94	38	48	37	23	9	13	76	76	42	57	29	84	80	83	20	5
Luxenburgo	97	61	86	73	12	7	26	85	94	83	20	91	94	94	84	31	24
Inglaterra	27	86	99	55	76	17	20	76	68	89	91	11	95	94	57	11	28
Portugal	72	26	77	34	1	5	20	22	51	8	16	89	65	78	92	6	9
Austria	55	31	61	33	1	5	15	49	42	14	41	51	51	72	28	13	11
Suiza	73	72	85	69	10	17	19	79	70	46	61	64	82	48	61	48	30
Suecia	97	13	93	43	43	39	54	56	78	53	75	9	68	32	48	2	93
Dinamarca	96	17	92	32	17	11	51	81	72	50	64	11	92	91	30	11	34
Noruega	92	17	83	51	4	17	30	61	72	34	51	11	63	94	28	2	62
Finlandia	98	12	84	27	10	8	18	50	57	22	37	15	96	94	17	5	64
España	70	40	40	43	2	14	23	59	77	30	38	86	44	51	91	16	13
Irlanda	30	52	99	75	18	2	5	57	52	46	89	5	97	25	31	3	9

Cuadro 1. Porcentaje de hogares en 16 países de Europa Occidental que consumen regularmente 17 artículos alimenticos comunes

país fue interrogado respecto a su frecuencia de consumo de 17 artículos alimenticios comunes. Los datos fueron organizados en una matriz de dimensiones 16×17 en la que cada columna corresponde a un producto alimenticio (variable) y cada fila a un país (objeto). Los valores registrados corresponden a los porcentajes de hogares que consumen regularmente los productos alimenticios en consideración. Los datos se muestran en cuadro 1.

En primer lugar, se efectuó un análisis exploratorio de los datos usando PCA. Este método produce nuevas variables no correlacionadas entre sí, llamadas componentes principales, que son combinaciones lineales de las variables originales. Los componentes se calculan de tal manera que la primera componente principal (PC1) sigue la dirección de máxima variabilidad de los datos y por consiguiente explica la mayor cantidad de información presente en los datos; cada componente principal sucesiva (PC2, PC3, etc.) es ortogonal a la anterior y explica la mayor cantidad de información residual. El conjunto de datos puede entonces ser proyectado sobre las dos o tres primeras componentes principales para visualizar su estructura conservando la máxima cantidad de información posible para este reducido número de dimensiones. Matemáticamente, las componentes principales son los autovectores de la matriz de correlación de las variables originales. Los coeficientes de las variables originales en las combinaciones lineales que constituyen las componentes principales se denominan *loadings* y representan la contribución de cada variable original en una componente principal dada. Mientras más alto es el *loading* (se toma en cuenta su valor absoluto) de cierta variable en una determinada componente principal más aporta aquella en esta. Este hecho permite asociar a las componentes principales con factores que caracterizan a los datos.

Variable	PC1	PC2	PC3	PC4	PC5
Café molido	0,033	0,115	**-0,538**	0,232	-0,096
Café instántaneo	-0,287	**-0,335**	0,158	-0,011	0,127
Té	**-0,312**	0,182	0,163	0,135	0,089
Sopa de sobre	-0,275	-0,282	0,079	-0,101	-0,113
Sopa de lata	**-0,357**	0,078	0,129	-0,100	0,073
Papa	-0,186	0,213	-0,270	**-0,437**	-0,118
Pescado congel.	-0,119	**0,363**	**-0,281**	-0,114	0,174
Manzanas	**-0,303**	-0,215	-0,237	0,064	-0,069
Naranjas	0,227	-0,156	**-0,457**	-0,049	-0,046
Frutos secos	**-0,407**	-0,078	-0,046	-0,021	0,262
Mermelada	**-0,331**	0,095	0,250	-0,148	**-0,298**
Ajo	0,239	**-0,345**	-0,206	-0,174	0,136
Mantequilla	-0,133	0,012	0,064	-0,089	**0,553**
Margarina	-0,108	0,063	-0,100	**0,655**	0,365
Aceite de oliva	0,154	-0,221	-0,137	**-0,408**	**0,442**
Yogurt	-0,168	**-0,349**	-0,234	0,157	-0,295
Pan	-0,099	**0,446**	-0,157	-0,130	0,001

Cuadro 2. Loadings de las 5 primeras componentes principales (aquellas con autovalores mayores de 1). Las contribuciones más importantes (valores absolutos más altos) se indican en negrillas

Componente principal	Autovalor	Varianza (%)	Varianza acumulada (%)
PC1	4,9603	29,2	29,2
PC2	3,4776	20,5	49,6
PC3	2,6387	15,5	65,2
PC4	1,4326	8,4	73,6
PC5	1,1718	6,9	80,5

Cuadro 3. Autovalores, varianzas individuales y varianzas acumulativas de las cinco primeras componentes principales

En segundo lugar, un método de agrupamiento jerárquico aglomerativo fue aplicado sobre los datos (10). Tal técnica trabaja directamente sobre el espacio multidimensional de las variables originales para buscar agrupamientos de objetos con comportamiento similar. El criterio básico para agrupar los objetos es su similitud, esto es, los objetos son agrupados en virtud de su proximidad en el espacio. Las técnicas de agrupamiento jerárquico aglomerativo inicialmente consideran a cada objeto como un agrupamiento individual; estos agrupamientos se combinan progresivamente empleando una métrica de similitud, usualmente una distancia, y en consecuencia se obtiene agrupamientos más grandes en cada etapa del procedimiento hasta que se obtiene un solo gran agrupamiento que contiene todos los objetos. El producto es un diagrama de árbol en el cual se visualiza la progresiva asociación de los objetos en virtud de su cercanía y que puede ser "cortado" a cual-

quier nivel de similitud para separar los objetos en un determinado número de grupos.

Finalmente se aplicó un algoritmo de clasificación sobre los datos. Este método genera una función matemática denominada "regla de clasificación" que se aplica sobre los grupos identificados en la etapa de agrupamiento para evaluar cuán bien se diferencian estas categorías. Adicionalmente, la regla de clasificación permite predecir la membresía de categoría de un objeto de identidad desconocida.

Todos los cálculos fueron efectuados con el paquete de *software* SCAN (11).

RESULTADOS Y DISCUSIÓN

Los resultados del análisis de componentes principales se muestran en los cuadros 2 y 3. El cuadro 2 presenta la composición de las cinco primeras componentes principales, las cuales poseen autovalores mayores que 1. En conjunto, ellas incluyen el 80,5% de la información presente en los datos.

De la inspección del cuadro 2, se puede observar que las variables originales que más contribuyen en la primera componente principal son frutos secos, sopa en lata, mermelada, té y manzanas. Una interpretación plausible de esta componente no es inmediata, pero parece contener elementos de la dieta mediterránea y de la dieta del norte del continente por lo que podría ser útil para distinguir a los grupos de países que consumen tales dietas. En la segunda componente, las variables con mayor contribución son pan, pescado congelado, yogurt, ajo y café instantáneo. Esta componente se podría asociar con la dieta nórdica por lo que sería útil para caracterizar al grupo de países escandinavos. Un análisis similar puede efectuarse para el resto de componentes.

El cuadro 3 presenta el porcentaje de información retenido por cada componente.

De la inspección del cuadro 3, se puede observar que primera componente (PC1) retiene 29,2% de la información contenida en los datos. PC2, a su turno, retiene 20,5% de la información residual. Por consiguiente, las dos primeras componentes conservan el 49,6% de la información total de los datos.

La figura 1 muestra la proyección de los objetos sobre las dos primeras componentes principales. De la observación de este gráfico se puede hacer una aproximación preliminar respecto a la distribución de los objetos. Por ejemplo, los objetos 11, 12, 13 y 14 (Suecia, Dinamarca, Noruega y Finlandia) forman un grupo discreto que se separa a lo largo de la segunda componente principal. Similarmente, los objetos 2, 8, 9 y 15 (Italia, Portugal, Austria y España) constituyen otro grupo bien definido que se separa a lo largo de la primera componente principal. No existe una agrupación clara del resto de objetos. Estas aproximaciones validan la interpretación de las dos primeras componentes efectuada anteriormente.

La figura 2 muestra la proyección de las variables sobre las dos primeras componentes principales. Dado que la proximidad en el espacio es sinónimo de similitud, se

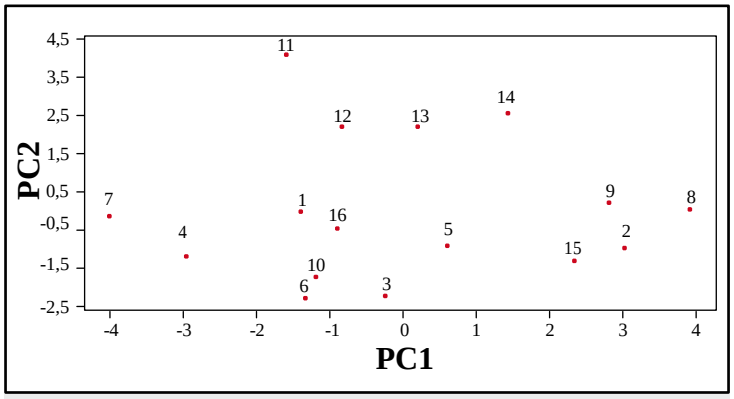

Figura 1. Proyección de los objetos en el espacio PC1 – PC2; 49.6% de la información total es retenida en este gráfico

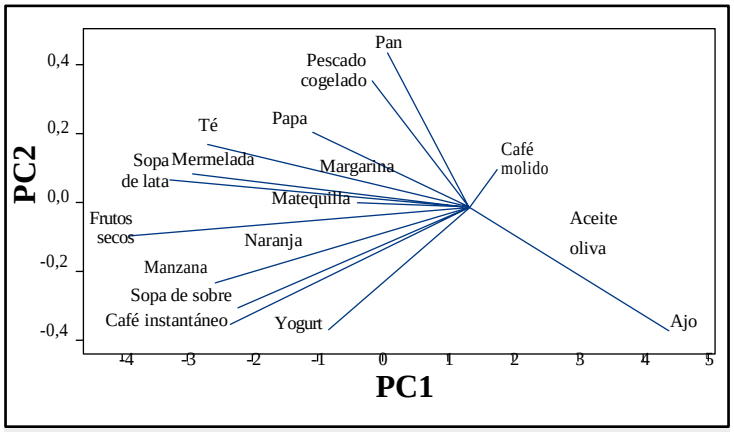

Figura 2. Proyección de las variables en el espacio PC1 – PC2; 49,6% de la información total es retenida en este gráfico

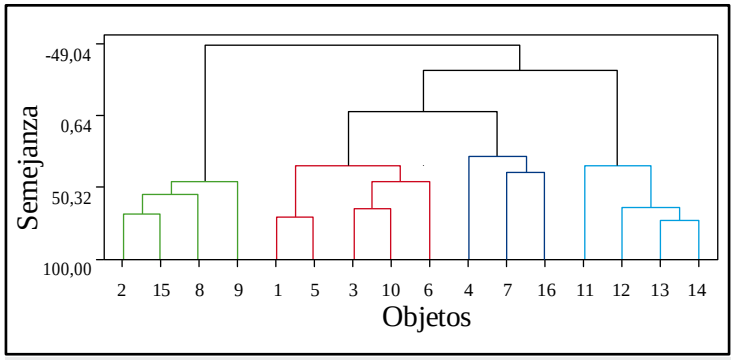

Figura 3. Dendograma del agrupamiento jerárquico aglomerativo de los objetos

puede inferir que las variables que forman agrupamientos portan información similar. Tal hecho puede proveer una manera de reducir el número de variables empleadas en un estudio, ya que el analista podría eliminar alguna o algunas variables de cierto grupo ya que todas portan la misma información. Por ejemplo, este sería el caso del grupo integrado por sopa enlatada, mermelada y té y de aquel integrado por naranjas, manzanas, sopa en polvo y café instantáneo. Sin embargo, en este estudio se ha decidido conservar todas las variables para efectuar los siguientes análisis.

La figura 3 muestra el dendograma resultante de la aplicación de un método de agrupamiento jerárquico aglomerativo sobre el conjunto de datos original de 16 objetos y 17 variables. El método de Ward fue utilizado puesto que se conoce que produce agrupamientos significativos. El dendograma proporciona una división de los objetos en cuatro agrupamientos bien definidos. Desde la izquierda, el primer grupo incluye los objetos 2, 8, 9 y 15 (Italia, Portugal, Austria y España). El segundo está integrado por los objetos 1, 3, 5, 6 y 10 (Alemania, Francia, Bélgica, Luxemburgo y Suiza). El tercero contiene los objetos 4, 7 y 16 (Holanda, Inglaterra e Irlanda). Por último, el cuarto agrupamiento incluye los objetos 11, 12, 13 y 14 (Suecia, Dinamarca, Noruega y Finlandia). El cuadro 4 resume la conformación de los agrupamientos.

Agrupamiento	Conformación	Interpretación
A	Italia, Portugal, Austria y España	Países mediterráneos más Austria
B	Alemania, Francia, Bélgica, Luxemburgo y Suiza	Países centrales
C	Holanda, Inglaterra e Irlanda	Países del Reino Unido más Holanda
D	Suecia, Noruega y Finlandia	Países nórdicos

Cuadro 4. Agrupamientos observados en el dendograma

La interpretación de los agrupamientos es simple. El agrupamiento A corresponde a los países mediterráneos más Austria. Al observar el cuadro 1, ellos comparten una alta ingesta de aceite de oliva, pan, frutas y ajo, consumo moderado de pescado y bajo consumo de yogur y mantequilla. El agrupamiento B está integrado por los países del centro del continente. Ellos comparten una dieta con alta ingesta de fruta fresca, mantequilla y café molido, consumo moderado de yogur, pan, patatas y vegetales. El agrupamiento C está integrado por los países del Reino Unido más Holanda; ellos se caracterizan por un alto consumo de té, mermelada y fruta, tanto fresca como seca, moderado consumo de lácteos y bajo consumo de vegetales. El agrupamiento D reúne a los países nórdicos. Ellos consumen una dieta baja en vegetales, alta en productos lácteos, fruta fresca, café molido y té y moderada en pescado y aceite de oliva.

Como se esperaba, la distribución de los grupos refleja similitudes en cultura, historia y tradiciones así como la proximidad geográfica, lo cual se traduce en similares hábitos de alimentación (12). Una relación no prevista que surge de los resultados es que Austria se agrupa con los países mediterráneos. Esto parecería extraño a primera vista, pero es importante recordar que Austria ocupó una gran porción del territorio italiano de 1815 a 1859 (13) y en ese lapso es posible que los austriacos hayan adoptado algunas costumbres locales, entre ellas, seguramente, ciertos hábitos de alimentación. Otra relación interesante es la similitud entre Holanda y los países del Reino Unido, esta se podría explicar a partir del hecho que Holanda y el Reino Unido mantienen fuertes lazos políticos y económicos desde la era Napoleónica; por ejemplo, más de 40 ciudades de Holanda tienen sus gemelas británicas; además, el holandés y el inglés son idiomas germánicos occidentales –incidentalmente, un 87% de los habitantes de Holanda sostiene que habla

el inglés– (14). Esto podría explicar sus hábitos alimenticios similares.

Finalmente, las membrecías de los objetos a las clases encontradas en el análisis de agrupamiento fueron verificadas usando el método de clasificación llamado SIMCA (Modelamiento independiente de analogía de clases). Este método construye un modelo de cada clase sobre la base de unos pocos componentes principales de cada una. Estas envolturas multidimensionales permiten a un analista determinar si un objeto dado pertenece o no a una clase preestablecida. El cuadro 5 presenta la matriz de clasificación resultante de tal procedimiento.

La matriz de clasificación revela que SIMCA asigna correctamente todos los los objetos en sus clases previamente identificadas. El desempeño del método en fase clasificación es perfecto lo cual confirma que las clases derivadas del análisis de agrupamientos forman grupos discretos en el espacio multidimensional de las variables.

Clase preestablecida	Total	Clase asignada			
		A	B	C	D
A	4	4	0	0	0
B	5	0	5	0	0
C	3	0	0	3	0
D	4	0	0	0	4
Errores de clasificación: 0					

Cuadro 5. Matriz de clasificación SIMCA

CONCLUSIONES

Se ha aplicado el análisis de datos multivariados para identificar patrones de comportamiento relativos a consumo de alimentos en 16 países de Europa Occidental. De los resultados se concluye los hábitos de consumo de alimentos permiten distribuir a los países bajo estudio en cuatro grupos bien definidos. Estos grupos son: países mediterráneos (A), países nórdicos (B), países centrales (C) y países del Reino Unido (D). La similitud de costumbres, cultura e historia, así como la cercanía geográfica explican la constitución de tales grupos. El estudio ilustra el gran potencial del analisis de datos multivariados para encontrar patrones de similitud de comportamiento y otras relaciones más sutiles en conjuntos de datos complejos.

Referencias

1. Hair JF. Multivariate Data Analysis: A Global Perspective, Prentice Hall: Upper Saddle River; 2009.
2. Jolliffe IT. Principal Component Analysis. Springer: New York; 2002.
3. Everitt B, Landau S, Leese M. Cluster Analysis, Wiley: New York; 2011.
4. Otto, M. Chemometrics, Wiley: New York; 2007.
5. http://www.umetrics.com/downloads (Fecha de acceso: diciembre 15, 2014).
6. Tognon G. *et al.* 2014. Adherence to a Mediterranean-like dietary pattern in children from eight European countries. The IDEFICS study. International Journal of Obesity 38: S108-S114.
7. Kittler PG, Sucher KP. Food and Culture: 3rd edition. Stamford, CT: Wadsworth; 2001
8. Kourlaba, G. *et al.* 2009. Dietary patterns in relation to socio-economic and lifestyle characteristics among Greek adolescents: a multivariate analysis." Public health nutrition 12.09 (2009): 1366-1372.
9. Hammerling U. *et al.* 2014. Identifying food consumption patterns among young consumers by unsupervised and supervised multivariate data analysis. European Journal of Nutrition and Food Safety. 4.4: 392-403.
10. Miller JN, Miller JC. Statistics and chemometrics for analytical chemistry. Pearson Education: Harlow; 2005.
11. Brown SD. 1994. SCAN (Software for Chemometric Analysis), available from JerII, Inc., North American Office: 790 Esplanada, Stanford, CA 94305, U.S.A. (telephone (415) 856-3401); European Office: via V. Pisani 13, 20124 Milano, Italy (telephone 39-2-26603247. Journal of Chemometrics. 8: 95–96.
12. Capacci S. *et al.* 2012. Policies to promote healthy eating in Europe: a structured review of policies and their effectiveness. Nutrition reviews. 70.3: 188-200.
13. Gritti S. 1988. Italia, Instituto Geografico de Agostini: Novara; 1988.
14. Ashton N, Hellema D. (Eds.). Unspoken allies: Anglo-Dutch relations since 1780. Amsterdam University Press: Amsterdam; 2001.
1. 2013. Disminución de la dosis de radiación en el radiodiagnóstico. Scielo. 19(1).