

Extensiones del análisis de supervivencia en la investigación en management: las curvas ROC.

Autores: Ana Debón Aucejo¹
Francisco Puig Blanco²

RESUMEN:

El objetivo de este trabajo es mostrar la aplicabilidad de curvas ROC (*Receiver Operating Characteristic*) en la investigación en management. A pesar de que la utilización de esta técnica es relativamente común en áreas como biomedicina, ingeniería, su uso en otras disciplinas como el marketing o la organización de empresas es más bien escasa. Por tanto, este trabajo persigue describir algunas situaciones en la que la utilización de curvas ROC es apropiada para la realización de análisis empíricos, comparar los modelos disponibles para el investigador y proponer ámbitos en los que su aplicación es posible.

Palabras clave: Curvas ROC, Modelo de Cox, Modelos Lineales Generalizados.

¹ Dep. de Estadística I.O. Aplicadas y Calidad. Universidad Politécnica de Valencia (España).

² Dep. de Dirección de Empresas “Juan J. Renau Piqueras”. Universitat de València (España). francisco.puig@uv.es.

1. INTRODUCCION

El estudio de los factores que influyen en las decisiones estratégicas de las empresas ha sido siempre de interés para el management. Desde distintas perspectivas y metodologías estadísticas se ha considerado el problema de determinar qué factores tienen un efecto significativo en la internacionalización o no de una empresa, en su quiebra o liquidación, y en general en cualquier estrategia de tipo corporativo. La investigación de esta cuestión ha hecho uso frecuente de los métodos econométricos tradicionales: Regresión Lineal, o modelos Logit o Probit. Más recientemente se han considerado otros modelos procedentes del Análisis de Supervivencia (Fuentedelsaz et al., 2004). En este caso las técnicas consideradas han ido desde el modelo de Cox (1972) a otros modelos estadísticos de más reciente creación como los Modelos Lineales Generalizados (entre otros).

El propósito de esas técnicas ha sido explicativa, sin embargo en raras ocasiones se han comparado los modelos aplicados a partir de sus resultados, más bien la decisión de un ajuste sobre otro se ha basado en medidas de bondad de ajuste. Es en este punto donde la curva ROC pueden servir de ayuda al investigador ya que la capacidad o validez del modelo aplicado se juzga en función de la clasificación certera o exitosa de las empresas a partir de la probabilidad ajustada por el propio modelo.

Con el objetivo de mostrar una aplicación práctica de la curva ROC en una investigación de management este trabajo se ha estructurado en cinco apartados: en el siguiente se explicita teóricamente; en el tercero se describe el diseño del estudio y en el cuarto se muestran los principales resultados obtenidos; el quinto y último epígrafe se dedica a las conclusiones del trabajo.

2. LA CURVA ROC.

Una curva ROC (Receiver Operating Characteristic) es un gráfico o técnica para visualizar, organizar y seleccionar clasificadores basados en su performance (Fawcett, 2006). Usualmente los gráficos ROC han sido utilizados para decisiones vinculadas al campo de la medicina y en años más recientes se están utilizados de forma creciente en investigación *machine learning* y *data mining*.

Para su comprensión consideremos problemas donde las observaciones (empresas) sólo pertenecen a 2 clases (internacionalizadas o no; supervivientes o no, etc.) y algunos modelos de clasificación o clasificadores (por ej. una regresión lineal o la regresión de Cox) que producen un output continuo (por ej., la probabilidad de internacionalización o la de quiebra). De acuerdo a ese clasificador según el punto de corte se obtendría una predicción de la clase a la que pertenece la observación (empresa).

De esa forma, para cada empresa tendríamos dos tipos de resultados: 1) la predicción del modelo y 2) la verdadera clasificación de la empresa. Por ello, conforme a la tabla (1) y dado un clasificador y un conjunto de empresas es posible distinguir 4 posibles resultados.

Tabla 1: Resultados en forma de tabla.

		Verdadera clase a la que pertenece la empresas	
		A	B
Clase predicha por el modelo	A	Verdaderos positivos	Falsos positivos
	B	Falsos negativos	Verdaderos negativos
		Total positivos	Total negativos

Siendo, el ratio de verdaderos positivos (tp rate) de un clasificador el cociente siguiente:

- $tp\ rate = \text{verdaderos positivos} / \text{total positivos}$,

y el ratio de falsos positivos (fp rate) el siguiente:

- $fp\ rate = \text{falsos positivos} / \text{total negativos}$,

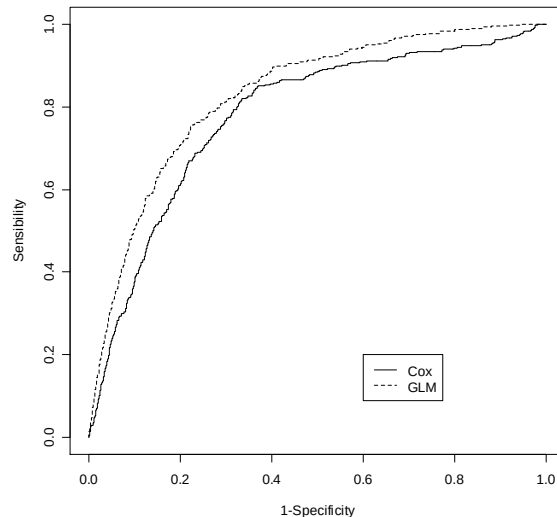
de acuerdo a los cuales, cuanto mayor es el tp rate y menor el fp rate mejor clasificador o modelo ha empleado el investigador.

Otros términos asociados con las curvas ROC son la sensibilidad y la especificidad de un modelo, medido como sigue:

- $\text{sensibilidad} = tp\ rate$
- $\text{especificidad} = 1 - fp\ rate$

Los gráficos ROC son gráficos bidimensionales en los cuales el tp rate (sensibilidad) se dibuja en el eje Y (vertical) y el fp rate (1-especificidad) en el X (horizontal). En la figura (1) se puede observar una representación de este gráfico para dos modelos diferentes como los pueden ser el de Cox y el GLM.

Figura 1: Ejemplo de gráfico ROC.



Fuente: elaboración propia.

De esta manera, un gráfico ROC representa el equilibrio relativo entre beneficios (verdaderos positivos) y costes (falsos negativos). Donde la diagonal $Y=X$ representa la estrategia de elegir la clase aleatoriamente y el punto (0,1) (vértice superior izquierdo) representa la clasificación perfecta. Por ello son preferibles los clasificadores o modelos más cercanos a la zona de arriba a la izquierda³. De acuerdo a ello, podemos afirmar que en la figura (1) el modelo GLM sería preferible al de Cox. Por todo ello, el modelo será mejor cuanto más discrimine entre las empresas que sí se han presentado un suceso (internacionalización, quiebra,...) y las que no.

Otras ventajas de la curva ROC son que proporciona una representación global de la exactitud del modelo y que ofrece el “optimal operating point”. Se dice que la exactitud de un modelo se incrementa conforme la curva se acerca a la esquina superior izquierda. Este indicador es útil para comparar clasificadores mediante un escalar obtenido del área bajo la curva de ROC (AUC). Como AUC es la proporción de área relativa a 1, toma valores entre 0 y 1. Dado que el procedimiento de elección aleatorio del grupo correspondería al área 0.5, cuanto mayor es el AUC mejor es el clasificador.

Por otra parte, el “optimal operating point” de la curva de ROC es el punto perteneciente a dicha curva más cercano a la esquina superior izquierda. Ese punto proporciona el valor umbral del riesgo que puede ser utilizado para detectar observaciones (empresas) que efectivamente van a hacer algo (por ej. internacionalizarse o están en riesgo alto de presentar irregularidades como una quiebra).

En la literatura y paquetes estadísticos como SPSS es posible encontrar otras técnicas que también permiten estimar la capacidad predictiva de un modelo como es el análisis discriminante o las tablas de clasificación. Sin embargo, a diferencias de éstas, las curvas ROC tienen la ventaja de que permiten decidir sobre el punto de corte o riesgo para la clasificación (en los otros casos suele

³ En algunos casos es posible que las curvas se corten. En esas situaciones lo que indica las curvas son las regiones concretas en las que un clasificador tiene mejores resultados que otro.

situarse en el 0.5). Esta característica les permite ser utilizadas por decisores con mucha o poca aversión al riesgo según el punto óptimo que elijan.

En definitiva, las curvas ROC son una técnica estadística que por medio de indicadores gráficos y escalares permiten, vía comparación, evaluar la capacidad predictiva de los modelos aplicados en una investigación. Dado que nuestro objetivo es mostrar una aplicación práctica de esta técnica en una investigación de management y sobre dos técnicas diferentes como son: el modelo de Cox y los modelos lineales generalizados, en el siguiente epígrafe vamos a exponer el diseño del estudio y posteriormente se llevará a cabo el análisis y discusión de resultados.

3. DISEÑO DEL ESTUDIO

La Fuente de Información Estadística utilizada ha sido la base de datos: SABI⁴. La población objeto de estudio lo configuran las empresas españolas que pertenecen al sector textil-confección: códigos 17 y 18 de la CNAE 93 y NACE Rev. 1. El criterio de selección de la muestra se ha basado en la declaración de actividad principal efectuada por parte de la empresa; de forma que sólo se han considerado aquellas empresas recogidas bajo el código primario relacionado con la rama textil (epígrafes 17.1 a 17.7) y con la rama de la confección (epígrafes 18.2). De esa manera, no se han considerado otras empresas que aún realizando actividades de este sector, han declarado como actividad principal otras, como por ejemplo las dedicadas a la comercialización u otros sectores relacionados como la maquinaria, el químico o el curtido.

El último periodo del que se disponen de datos de SABI es del año 2007. Según la extracción efectuada a finales de abril de 2009, el total de empresas del textil-confección que configuraban la población ascendían a un total de 11.741 empresas.

De acuerdo con los planteamientos expuestos, el total de variables empleadas han sido 5, las cuales han sido clasificadas en dos grupos: dependientes e independientes. La variable dependiente del análisis la hemos denominado Supervivencia. Por otra parte, las variables independientes son 4: Subsector de actividad, Productividad, Tamaño y Coeficiente de Especialización.

La variable **Supervivencia** toma dos valores: 0 = si la empresa está en quiebra, concurso, extinguida o en suspensión de pagos. Y 1 = si está activa.

El **Subsector** de actividad lo conforman las 5 actividades principales o subsectores siguientes: (1) Hilaturas (epígrafes 17.1 y 17.2), (2) Acabados (17.3), (3) Hogar y otras aplicaciones textiles (17.4 y 17.5), (4) Punto (17.6 y 17.7) y (5) Vestuario (18.2).

La **Productividad** (PR) es un ratio de actividad que está definido como la contribución de los empleados en el resultado final de la gestión. Concretamente, este ratio es el cociente resultante de la siguiente ecuación:

$$(\text{Cifra Vtas.} + \text{Otros Ingr.} - \text{Consumos} - \text{Otros Gtos. Exp.}) / \text{Gtos. de Personal}$$

El **Tamaño** está medido por el número de empleados (NE).

El **Coeficiente** de Especialización (CE) es un estadístico que sirve para valorar la presencia de una actividad (en nuestro caso el sector textil-confección) en un territorio (provincia) respecto a la presencia de ese sector en el conjunto de referencia (nación). Cuanto mayor es ese dato más significativa es la presencia de la actividad en el territorio considerado.

Su expresión matemática es la siguiente:
$$CE_{ij} = \frac{E_{ij} / E_j}{E_i / E_n}.$$

Donde:

E_{ij} se refiere al volumen de empleo u ocupados de la rama i en la entidad territorial j,

E_j indica el total de empleo en la entidad territorial j,

E_i es la variable que muestra el total del empleo de la rama i,

E_n recoge el volumen total de empleo referida al ámbito nacional.

Una aplicación práctica sobre el cálculo y obtención de coeficiente de especialización en la industria textil española puede encontrarse en Puig et al. (2007).

⁴ El Sistema de Análisis de Balances Ibérico (SABI) se nutre de los datos recogidos en el Registro Mercantil. Esta base en CD-Rom recoge información sobre más de 550.000 empresas españolas y 65.000 empresas portuguesas. Su utilización nos proporciona importantes ventajas, sobre todo respecto a su desagregación individualizada a nivel territorial y sectorial.

Las técnicas de análisis que se han empleado están en función de los objetivos de la investigación y de la naturaleza de los datos (Hair, et al. 2000). Dado que el propósito del presente trabajo se basa en un análisis exploratorio tendente a evaluar la capacidad de ajuste del Modelo de Cox frente a los Modelos Lineales Generalizados se ha empleado la técnica de las curvas ROC.

4. ANALISIS Y DISCUSIÓN DE RESULTADOS.

4.1 Modelo de Cox

Como es habitual en las técnicas de supervivencia la estimación de los efectos de las covariables (subsector, productividad, número de empleados, coeficiente de especialización) se analiza sobre el riesgo de irregularidades de la empresa (hazard rate) asociado al tiempo de vida o duración de dicha empresa.

Comenzamos aplicando el modelo de riesgos proporcionales (Proportional Hazard model), también llamado regresión de Cox (1972). Su expresión matemática dada en la ecuación es:

$$h(t; x) = h_0(t) \exp(\beta' x) \quad (1)$$

donde $h_0(t)$ es una función base no especificada y β es el vector de coeficientes de la regresión, asociado a x . Por analogía en este trabajo el riesgo (hazard rate) es la probabilidad que una empresa, que sobrevive hasta el instante t , presente irregularidades dentro del siguiente intervalo de tiempo. Por lo que en nuestro caso, la función de riesgo (hazard function) $h(t)$ mide el riesgo de que una empresa presente irregularidades, condicionada por el hecho que no las ha presentado y ha sobrevivido hasta el instante t .

Además, x_i es el vector de covariables para el i -ésimo caso. La principal hipótesis es que el riesgo en todos los casos es un múltiplo de un ratio base no especificado (baseline hazard rate) $h_0(t)$. El modelo de riesgos proporcionales de Cox ha sido ajustado en R mediante el comando *coxph* utilizando la librería *survival* Therneau(2009). Detalles sobre los argumentos y estimación pueden encontrarse en Venables y Ripley (2002). Estudios de management empleando esta técnica puede verse en Fuentelsaz et al. (2003) o Pla-Barber y Puig (2009).

Los resultados de la regresión de Cox se muestran en la Tabla 2, que incluye el valor de los coeficientes para cada una de las covariables, la exponencial del valor del coeficiente (que expresa el efecto de la correspondiente covariable en el riesgo, hazard rate), el error estándar y la significatividad para cada uno de los coeficientes. La categoría de referencia de la variable cualitativa o factor subsector, confección, no aparece en dichas tablas.

Tabla 2: Regresión de Cox.

Covariables	β	$\exp(\beta)$	$se(\beta)$	sig.
Hilados	-0,302	0,739	0,085	0,000
Acabados	-0,167	0,846	0,120	0,163
Hogar-técnicos	-0,471	0,625	0,104	0,000
Punto	-0,112	0,894	0,104	0,285
PR	0,003	1,003	0,002	0,219
NE	-0,006	0,994	0,001	0,000
CE	0,195	1,215	0,035	0,000

Fuente: elaboración propia.

La interpretación del coeficiente positivo β que corresponde al CE (coeficiente de especialización de una empresa) (0,195) se puede interpretar de la siguiente manera: el riesgo de quiebra que corresponde a una empresa es mayor cuando aumenta su coeficiente de especialización. De forma contraria, el significado del coeficiente negativo que corresponde al NE (número de empleados) (-0,006), en este caso el riesgo disminuye conforme aumenta el número de empleados de la misma.

Por otra parte, si consideramos de forma aislada el efecto del subsector al que pertenece la empresa, de acuerdo con nuestros resultados, sólo son significativos “hilados” y “hogar-técnicos”, y ambos tienen coeficientes negativos (-0,302 y -0,471), de forma que su riesgo es menor que en la categoría de referencia que es confección.

A partir de dicha tabla en la segunda columna cada valor individual del coeficiente puede ser interpretado de la siguiente forma: para la variable explicativa CE el incremento de No Supervivencia para un incremento de una unidad es de 21,5%, en el caso de la variable NE el decrecimiento en el riesgo para un aumento de un empleado es del 0,6%, el riesgo para las empresas de “acabados” y “punto” no es significativamente diferente de las del subsector “confección”, mientras que “hilados” y “hogar-técnico” disminuyen dicho riesgo en 26,1% y 37,5% respectivamente, si el resto de variables permanecen constantes.

4.2 Modelo lineal generalizado.

Para analizar la consistencia de los resultados anteriores, vamos a utilizar una técnica diferente: los modelos lineales generalizados (GLM). Esta técnica permite analizar datos bajo la hipótesis de que el número de empresas que presentan irregularidades se distribuye Poisson. Los GLM son una extensión de los modelos lineales para variables respuesta con distribuciones no normales y transformaciones no lineales (McCullagh y Nelder, 1989).

Un modelo de regresión constituye una especificación de la variable media m , en términos de un número pequeño de parámetros $\beta_0, \beta_1, \dots, \beta_p$ correspondientes a covariables.

En el caso particular de modelos lineales, deseamos encontrar una función lineal tal que $E(Y|X) = m = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, con la varianza de Y que se supone constante, $\text{var}(Y) = \sigma^2$.

De forma análoga, un GLM proporciona un método para la estimación de una función de la media de la variable respuesta como una combinación lineal de un conjunto de variables predictivas, esto es:

$$l(E(Y / X)) = l(m) = \eta = \beta_0 + \sum_{i=1}^p \beta_i x_i.$$

La función de la media de la variable respuesta, $l(m)$, se llama función link, y se considera una función lineal de los predictores, η , y se llama predictor lineal. Cada componente y_i tiene distribución Binomial, Poisson o Gamma.

Un referencia a consultar para comprender los GLM es McCullagh y Nelder (1989) y con predictores cualitativos en Agresti (2002). Un aplicación práctica de esta técnica puede verse en Debón et al. (2005).

Nuestro modelo es formalmente un modelo lineal generalizado Poisson con una función de enlace (link) logarítmica, lo hemos ajustado en R con el comando *glm* de la librería *stats*. Detalles sobre los argumentos y estimación pueden encontrarse en Venables y Ripley (2002).

En el modelo lineal generalizado hemos introducido la variable tiempo como explicativa, a diferencia del modelo de Cox donde aparece como parte del riesgo según su expresión (1). Los resultados del GLM se muestran en la tabla (3).

Tabla 3: Modelo Lineal Generalizado (GLM).

Covariables	β	$\exp(\beta)$	$se(\beta)$	sig.
Intercepto	-2,482	0,084	0,089	0,000
Tiempo	0,000	1,000	0,000	0,043
Hilados	0,131	1,140	0,084	0,120
Acabados	0,034	1,035	0,012	0,773
Hogar-técnicos	-0,392	0,676	0,010	0,000
Punto	0,183	1,200	0,010	0,080
PR	0,003	1,003	0,002	0,163
NE	0,000	1,000	0,000	0,827
CE	0,213	1,237	0,003	0,000

Fuente: elaboración propia.

De la lectura de dicha tabla podemos observar algunas diferencias comparando con las obtenidas por el modelo de Cox (tabla 2). Concretamente, en la variable subsector, donde solo resulta significativo “hogar-técnicos” y donde la variable NE no resulta significativa. La significatividad para el resto de variables es similar en ambos.

Respecto al signo de las variables y sus coeficientes, hogar-técnicos ($\beta = -0,392$), y CE ($\beta = 0,213$), que como hemos visto resultan significativas en ambos modelos, permanece igual.

4.3 Aplicación de la curva ROC.

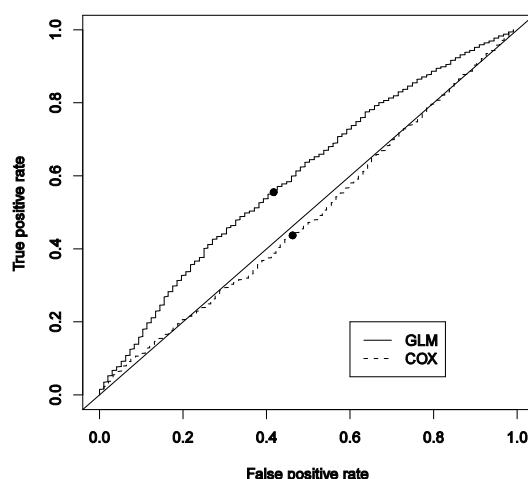
Los modelos anteriores han sido diseñados para mostrar el impacto de cada variable en el riesgo de una empresa de presentar irregularidades (en nuestro caso de supervivencia o no). Para evaluar el poder discriminante de esos modelos y poder compararlos vamos a representar la curva ROC.

Se utilizan en un test en el cual hay dos posibles resultados que son de interés para el investigador. El test proporciona resultados correctos positivos (A) o resultados correctos negativos (B). En este caso corresponde al hecho de que una empresa presente irregularidades (A) o no (B). De esta forma, la curva ROC obtenida puede verse como un resumen, en el sentido de que utiliza todos los posibles puntos de corte de riesgo para el test y, a partir de ello, crea o dibuja la curva correspondiente⁵.

La curva ROC resultante de nuestro análisis se muestra en la figura (2). La misma se ha obtenido en R utilizando la librería *ROCR*. Más detalles sobre esa aplicación puede encontrarse en Sing et al. (2007).

⁵ Notar que al comparar curvas ROC, los test con áreas grandes son considerados como los mejores test, pero es posible que dos test con la misma área no sean equivalentes. Por ello, es importante comparar las curvas ROC y sus áreas.

Figura 2: curvas ROC.



Fuente: elaboración propia.

De la visualización de dicha figura puede deducirse que el riesgo (hazard rate) obtenido con el GLM discrimina mejor entre las empresas que presentan irregularidades y las que no, puesto que su curva está siempre por encima de del modelo de Cox.

Además, el AUC correspondiente al riesgo obtenido con el GLM es 0,594, el cual es superior al obtenido por el modelo de Cox (0,486). Esto confirma la superioridad del GLM.

Asimismo, los puntos de curvas muestran el óptimo operativo (optimal operating point) para cada modelo sobre su curva de ROC.

En definitiva, si el investigador hubiera empleado la técnica de Cox para estudiar qué factores tienen un efecto significativo en la probabilidad de supervivencia de su empresa, éste hubiera podido concluir con que el tamaño (NE), la especialización productiva en los subsectores “hilados” y “técnicos-hogar” y la localización (CE) son significativos. Sin embargo, si su análisis lo hubiera desarrollado empleando la técnica estadística del GLM hubiera concluido con que NE e “hilados” no son significativos. Llegado a ese punto, podría preguntarse cuál de las conclusiones es más adecuada, pudiendo encontrar respuesta en la curva ROC: de nuestro análisis se desprende que el GLM, indicándonos además el óptimo operativo.

5. CONCLUSIONES

La cada vez mayor utilización en management de las técnicas estadísticas basadas en el análisis de supervivencia están permitiendo a los investigadores de esta área de conocimiento el empleo y resolución de problemas con una metodología más adecuada al problema. No obstante, la validez de esas y otras técnicas suele basarse en simples medidas de bondad de ajuste o tablas de clasificación.

En este trabajo hemos demostrado que si bien el análisis de supervivencia es útil, ya que los resultados obtenidos con la regresión de Cox o vía GLM son bastante coincidentes, su validez debe realizarse con técnicas basadas en la curva ROC. De ese análisis y para el caso concreto aquí analizado (la supervivencia de empresas textiles) podemos concluir, con un criterio simple y objetivo, que el modelo de Cox clasifica peor que el GLM.

En definitiva, las extensiones que esta técnica tiene en management son variadas: a) proporcionan una medida de clasificación a la vez que son complementarias a otras medidas

escalares de bondad de ajuste; b) definen un umbral sobre el que apoyar la toma de decisiones del directivo y c) aunque en este trabajo se hayan aplicado sobre el caso de la supervivencia, también puede hacerse en otros aspectos como la aceleración de la internacionalización de una empresa o la duración de un acuerdo de cooperación.

BIBLIOGRAFIA

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, 2ª Edición.

Cox, D. R. (1972). *Regression Models and Life-Tables*, *Journal of the Royal Statistical Society*, 34/2, 187-202.

Debón, A., Montes, F. y Sala, R. (2005). A comparison of parametric models for mortality graduation application to mortality data of the Valencia region (Spain). *Statistics and Operations Research Transactions*, 29/2, 269-287.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27.

Fuentelsaz, L., Gómez, J. y Polo, Y. (2004). Aplicaciones del análisis de supervivencia a la investigación en economía de la empresa. *Cuadernos de Economía y Dirección de la Empresa (CEDE)*, 19, 81-114.

Fuentelsaz, L.; Gómez, J. y Polo, Y. (2003). Intrafirm diffusion of new technologies: an empirical application, *Research Policy*, 32/4, 533-551.

Hair, J. F.; Anderson, R. E. Tatham, R. L. Y Black, W. C. (2000) *Multivariate Data Analysis with Readings*. Prentice Hall: New Jersey.

McCullagh, P. y Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Pla-Barber, J. y Puig, F. (2007): "Is the Influence of the Industrial District on International Activities being eroded by Globalization? Evidence from a Traditional Manufacturing Industry", *International Business Review*, vol. 18, pp. 435-455.

Puig, F.; Pla, J. y Linares, E. (2008): "Los distritos industriales como unidad de análisis: una propuesta metodológica para su delimitación", *Revista ICADE*, núm. 73, enero-abril.

Sing, T.; Sander, O.; Beerenwinkel, N. y Thomas Lengauer (2007). *ROCR: Visualizing the performance of scoring classifiers*. R package version 1.0-2. <http://rocr.bioinf.mpi-sb.mpg.de/>

Therneau, T (2009). *Survival analysis, including penalised likelihood*. R package version 2.35-4.

<http://CRAN.R-project.org/package=survival>

Venables, W., Ripley, B.D. (2002). *Modern Applied Statistics with S-PLUS*. Springer, Nueva York.