MODELIZACIÓN DEL ENVEJECIMIENTO DE LA LITERATURA CIENTÍFICA EN PRESENCIA DE DATOS CENSURADOS

Jesús Basulto Santos
e-mail: basulto@us.es
Francisco Javier Ortega Irizo
e-mail: fjortega@us.es
José Antonio Camúñez Ruíz
e-mail: camunez@us.es
María Dolores Pérez Hidalgo
e-mail: mdperez@us.es

Departamento de Economía Aplicada I
Universidad de Sevilla

Resumen

Ofrecemos un nuevo enfoque para modelizar la variable T que proporciona la antigüedad de las citas recibidas por los trabajos científicos, cuando los datos utilizados son los elaborados por el ISI (Institute of Scientific Investigation) , en los que las citas de antigüedad superior a 10 años aparecen agregadas. Cada dato t_i puede interpretarse como el tiempo que transcurre desde que un trabajo es publicado hasta que recibe la cita que se ha observado; de cada una de las citas agregadas, sólo conocemos que el tiempo transcurrido t_i es superior a 10 años. De esta forma, podemos encuadrar nuestros datos en la situación de los modelos de duración o supervivencia con datos censurados a la derecha, y aprovechar los conocimientos y herramientas ya desarrollados en este campo. Para datos de revistas pertenecientes al ámbito de economía aplicada, exploramos qué modelo(s) resulta(n) más adecuado(s) de entre aquellos que se han usado habitualmente en el campo de datos de supervivencia.

Palabras clave: Bibliometría, Modelos de duración, Datos censurados.

1. Introducción.

En los estudios bibliométricos, la variable T que proporciona la antigüedad de las citas recibidas por los trabajos científicos (es decir, que mide el tiempo transcurrido entre la publicación del artículo y el momento en el que recibe la cita) es de interés para analizar y valorar a las revistas, temáticas, instituciones, grupos de investigación, etc. Los datos correspondientes a esta variable, suelen obtenerse en estudios *retrospectivos*. Por ejemplo, si queremos obtener las observaciones correspondientes a una determinada revista A, consideramos un conjunto fuente de revistas en un año concreto t y en él buscamos todas las citas efectuadas a trabajos anteriormente publicados en A en los años t, t-1, t-2, etc. De esta forma, obtendremos la distribución de la antigüedad de las citas recibidas por la revista A.

Los datos así obtenidos muestran gran similitud con los correspondientes a los modelos de supervivencia, duración o tiempo de fallo. En efecto, como es conocido, en este tipo de modelos se observa el tiempo transcurrido hasta que ocurre un determinado suceso (habitualmente denominado fallo), por ejemplo, se observa el tiempo de funcionamiento de un determinado componente electrónico. Pues bien, cuando observamos que en un año t_1 se ha producido una cita a un artículo publicado en el año t_0 podemos interpretar la antigüedad de la cita como el tiempo transcurrido hasta que ocurre el suceso "ser citado", lo que permitirá utilizar todas las herramientas ya desarrolladas para los modelos de supervivencia en el campo de la bibliometría.

Las bases de datos más importantes y utilizadas para este propósito son las elaboradas por el ISI (Institute of Scientific Investigation); en ellas, todas las citas de antigüedad igual o superior a 10 años aparecen agregadas, lo cual supone una seria limitación para el propósito de obtener un modelo que describa el comportamiento de los datos, por lo que en muchas investigaciones en esta línea (Burell (2002), Egghe y Ravichandra Rao (1992), Gupta (1998)) se acude a otras fuentes de datos menos completas. No obstante, en los estudios de supervivencia es muy habitual la presencia de datos censurados, que aparecen cuando una vez finalizado el experimento existen items para los cuales aún no ha ocurrido el suceso objeto de estudio. Así, si después de observar durante un tiempo L un determinado componente electrónico este aún sigue funcionando, de su tiempo de fallo sólo sabremos que es superior a L, por lo que tendremos un dato censurado. Análogamente, de las citas agregadas en las bases del ISI conocemos que su antigüedad es superior a 10, por lo que podemos considerarlas como datos censurados en nuestro modelo. Por tanto, podemos utilizar las bases del ISI pero teniendo en cuenta que nuestros datos se encuadran en la situación de los modelos de supervivencia con datos censurados a la derecha, pudiendo aprovechar los conocimientos y herramientas ya desarrollados en este campo.

En este trabajo pretendemos explorar qué modelos pueden resultar más adecuados para ajustar los datos de antigüedad de las citas, de entre aquellos que se han usado con mayor frecuencia hasta ahora, como son el log-Normal, log-Logistic y Weibull, así como el Exponencial debido a su simplicidad. Para ello, hemos utilizado un grupo de 10 revistas de referencias, todas ellas pertenecientes al ámbito de Economía Aplicada. El

grado de ajuste conseguido por los modelos ha sido analizado a través de métodos gráficos así como valorando las diferencias entre la función de distribución empírica y la teórica una vez estimados los parámetros con el método de máxima verosimilitud.

2. Revistas analizadas y datos obtenidos.

En la base de datos JCR Social Sciences Edition del ISI, correspondiente al año 2001, se han seleccionado las 10 revistas directamente relacionadas con el ámbito de la Economía Aplicada que se ofrecen en la tabla I.

	ABREVIATURA	NOMBRE COMPLETO
1	ECONOMET THEOR	Econometric Theory
2	ECONOMETRICA	Econometrica
3	INSUR MATH ECON	Insurance Mathematics & Economics
4	J APPL ECONOM	Journal of Applied Econometrics
5	J BUS ECON STAT	Journal of Business & Economic Statistics
6	J ECONOMETRICS	Journal of Econometrics
7	J MATH ECON	Journal of Mathematical Economics
8	J ROY STAT SOC A STA	Journal of the Royal Statistical Society Series A – Statistics in Society
9	OXFORD B ECON STAT	Oxford Bulletin of Economics and Statistics
10	REV ECON STAT	Review of Economics and Statistics

Tabla I. Revistas analizadas

Para todas ellas se han obtenido los datos retrospectivos acerca de la antigüedad de las citas recibidas que se ofrecen en la base de datos (año 2001). La variable T es de naturaleza continua, aunque sólo disponemos de datos anuales, por lo que surge el problema de qué valor de antigüedad otorgar a una cita producida en el año t_1 a un artículo publicado en el año t_0. Siendo coherentes con el método seguido por el ISI para estimar la mediana de la distribución, asignaremos el valor $(t_1-t_0)+0.5$, ya que se asume el reparto uniforme de observaciones en los intervalos (Basulto y Ortega (2002)). Así, la antigüedad de las citas a artículos publicados en 2001 se asignarían al intervalo [0,1), con marca de clase 0.5; para el año 2000, el intervalo es [1,2) con marca de clase 1.5, etc. El último año para el que aparecen las citas sin agregar es 1992, al que le corresponde el intervalo [9,10); el resto de citas corresponderían al intervalo de datos censurados [10,+∞). Los extremos y las marcas de clase de los intervalos de datos no censurados serán pues $[L_{j-1}, L_j)$ y t_j respectivamente, donde $L_j = j$ y $t_j = j - 0.5$, $j = 1,2,...,10$. Los datos obtenidos para las diez revistas se ofrecen en la tabla II.

	Año de la cita	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	Resto	Total
	Intervalo de antigüedad	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	>10	
	Marca de clase	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5		
1	ECONOMET THEOR	7	30	36	33	42	51	76	49	28	40	156	548
2	ECONOMETRICA	20	73	127	162	181	244	178	252	329	261	7080	8907
3	INSUR MATH ECON	3	16	23	27	28	11	14	13	15	12	45	207
4	J APPL ECONOM	1	26	33	27	53	74	77	26	73	49	161	600
5	J BUS ECON STAT	6	14	43	92	66	61	139	115	75	127	376	1114
6	J ECONOMETRICS	9	72	142	145	182	295	214	198	138	261	1560	3216
7	J MATH ECON	7	20	22	20	20	40	16	21	10	18	325	519
8	J ROY STAT SOC A STA	10	27	45	29	28	74	64	46	24	30	597	974

9	OXFORD B ECON STAT	0	11	65	19	26	49	50	15	43	114	259	651
10	REV ECON STAT	12	51	116	152	137	186	108	112	111	111	1780	2876

Tabla II: Datos de antigüedad de las citas

3. Metodología.

El análisis del grado de ajuste que consiguen los distintos modelos a los datos observados lo hemos hecho A) a través de métodos gráficos y B) comparando las funciones de distribución empírica y teórica, una vez estimados los parámetros del modelo por el método de máxima verosimilitud.

Para todos los modelos, hemos buscado una transformación de la función de supervivencia $S(t) = 1 - F(t)$ que tiene un comportamiento lineal frente al tiempo o al logaritmo del tiempo. Posteriormente, representamos la estimación no paramétrica de la función de supervivencia $\hat{S}(t)$ frente a t (o frente a $\log(t)$), siendo un modelo adecuado si los puntos se aproximan a una recta. La obtención de $\hat{S}(t)$ se ha hecho usando el estimador de Kaplan-Meier (Lawless (1982)). Los modelos que hemos utilizado son el Exponencial, Weibull, Log-Normal y Log-Logistic.

Modelo Exponencial: En este caso, $S(t) = e^{-\lambda t}, \lambda > 0$ y por tanto $\log(S(t)) = -\lambda t$. Así, la representación de $\mathrm{Ln}(\hat{S}(t_j))$ frente a t_j debe aproximarse a una recta que pase por el origen si este modelo es adecuado a nuestros datos.

Modelo Weibull: En este caso, $S(t) = e^{-(t/\lambda)^{\beta}}, \lambda, \beta > 0$ y por tanto $\log\left[-\log(S(t))\right] = \beta \log t - \beta \log \lambda$. Así, la representación de $\log\left[-\log(\hat{S}(t_j))\right]$ frente a $\log t_j$ debe aproximarse a una recta si este modelo es adecuado a nuestros datos.

Modelo Log-Normal: En este caso, $S(t) = 1 - \Phi\left((\log t - \mu)/\sigma\right), \mu \in \mathbb{R}, \sigma > 0$, donde $\Phi(\cdot)$ es la función de distribución del modelo Normal tipificado, y por tanto $\Phi^{-1}(1 - S(t)) = (1/\sigma)\log t - \mu/\sigma$. Así, la representación de $\Phi^{-1}(1 - \hat{S}(t_j))$ frente a $\log t_j$ debe aproximarse a una recta si este modelo es adecuado a nuestros datos.

Modelo Log-Logistic: En este caso, $S(t) = 1/\left[1 + (\rho t)^{\gamma}\right], \rho, \gamma > 0$ y por tanto $\log\left((1 - S(t))/S(t)\right) = \gamma \log t + \gamma \log \rho$. Así, la representación de $\log\left((1 - \hat{S}(t_j))/\hat{S}(t_j)\right)$ frente a $\log t_j$ debe aproximarse a una recta si este modelo es adecuado a nuestros datos.

Una vez hecho el análisis gráfico, estimamos los parámetros de los cuatro modelos utilizados para todas las revistas y procedemos a comparar las funciones de distribución empírica y teórica aplicadas en los puntos

L_j. El criterio de comparación elegido ha sido calcular la raíz del error cuadrático medio (RECM), donde los errores e_j se definen como la diferencia entre la función de distribución empírica y la teórica en el punto L_j.

4. Resultados.

En primer lugar se debe aclarar que el grado de ajuste ofrecido por el modelo Exponencial no es comparable al que se consigue con el resto de modelos, ya que en el primer caso nos encontramos ante un modelo uniparamétrico mientras que el resto de caso son biparamétricos. De hecho, el modelo Exponencial es un caso particular del Weibull, por lo que éste último siempre ha de ofrecer un mejor comportamiento. El hecho de incluir el modelo Exponencial es porque, dada su simplicidad, ha sido utilizado en algunos trabajos (Ruiz y Jiménez (1996), Brookes (1974)) y porque puede servir como punto de comparación para analizar la mejora que supone el pasar a los modelos de dos parámetros.

A) ANÁLISIS GRÁFICO

Las gráficas de ajuste correspondiente a cada modelo y revista se ofrecen en el anexo. En el eje OY representamos los valores de las correspondientes transformaciones de $\hat{S}(t_j)$ para cada modelo, mientras que en el eje OX se representan los valores $\log(t_j)$ para los modelos biparamétricos y t_j para el caso exponencial. Para los modelos Log-Normal, Weibull y Log-Logistic hemos representado la recta de regresión mínimo cuadrática así como el coeficiente de bondad de ajuste R^2 de la misma. Para el modelo Exponencial, hemos representado la recta de regresión mínimo cuadrática que pasa por el origen, no ofreciendo ningún coeficiente de bondad de ajuste ya que, como sabemos, en este caso no estaría acotado.

Podemos apreciar con claridad que salvo en el caso de la revista J MATH ECON (y tal vez INSUR MATH ECON) el modelo Exponencial no consigue ajustarse bien a los datos y que el paso a los modelos biparamétricos supone un gran aumento en el grado de ajuste. De hecho, es conocido que una característica común en los histogramas de la variable T es la aparición de una fase inicial de ascenso seguida de otra de descenso, siendo la longitud de dichos períodos y las "velocidades" de aumento y disminución las que caracterizan a las diversas temáticas, revistas, instituciones, etc. Es por ello que el modelo Exponencial no resulta adecuado, sobre todo si la fase de ascenso en el número de citas es prolongada.

Con respecto a los otros tres modelos, acudiendo al criterio de los coeficientes R^2 podemos apreciar que el comportamiento es muy similar (con una pequeña desventaja para el modelo Log-Normal), y que todos ellos pueden servirnos para nuestro propósito de describir el comportamiento de la variable T, aunque siempre teniendo en cuenta las limitaciones de un análisis gráfico de este tipo. Llamando a los modelos Log-Normal, Weibull y Log-Logistic modelos 1, 2 y 3 respectivamente, en los tres casos estamos representado una cierta función $f_i\left(\hat{S}_i(t_j)\right)$, $i=1,2,3$ frente a $\log(t_j)$; los parámetros de las rectas de regresión nos proporcionarían una

primera estimación de los parámetros de los respectivos modelos teóricos, y así las rectas representarían a las funciones teóricas $f_i(S(t))$; los coeficientes de determinación observados (muy próximos a la unidad) indican que las diferencias entre $f_i(\hat{S}_i(t_j))$ y $f_i(S(t_j))$ son pequeñas, pero la magnitud de las diferencias entre $\hat{S}_i(t_j)$ y $S(t_j)$ también dependerán de las funciones $f_i^{-1}(\cdot)$. Es interesante observar que $f_1^{-1}(\cdot)$ sería una composición con la función de distribución de la distribución Normal tipificada, mientras que $f_2^{-1}(\cdot)$ y $f_3^{-1}(\cdot)$ serían composiciones con la función exponencial, por lo que las diferencias observadas en los gráficos entre $f_i(\hat{S}_i(t_j))$ y $f_i(S(t_j))$ podrían aumentar bastante más en los modelos Weibull y Log-Logistic que en el caso Log-Normal al considerar las diferencias entre $\hat{S}_i(t_j)$ y $S(t_j)$ (es decir, que la pequeña desventaja observada en el caso Log-Normal podría compensarse al deshacer las transformaciones y volver al ajuste en base a la función de supervivencia).

Podemos observar también cómo el grado de ajuste conseguido parece depender más de la revista que estemos analizando que del modelo utilizado (excepción hecha del modelo Exponencial). Por ejemplo, la revista OXFORD B ECON STAT presenta un dato anómalo que hace que el grado de ajuste disminuya mucho para los tres modelos; las revista J APPL ECONOM y J BUS ECON STAT también presentan coeficientes de ajuste algo inferiores al resto de revistas. Por el contrario, las revistas que presentan mejor ajuste con respecto al resto lo hacen en los tres modelos (como por ejemplo es el caso de ECONOMETRICA y J ECONOMETRICS). La mayor diferencia entre los distintos modelos se da en las revistas ECONOMET THEOR y J BUS ECON STAT, en las que la bondad de ajuste del modelo Log-Normal es sensiblemente inferior al de los modelos Weibull y Log-Logistic.

En resumen, de este análisis gráfico puede deducirse que el modelo Exponencial resulta insuficiente para ajustar nuestros datos, mientras que el comportamiento de los otros tres modelos es adecuado y muy similar entre sí, si bien parece haber una ligera diferencia a favor de los modelos Weibull y Log-Logistic, que tal vez podría deberse a las diferentes transformaciones de $\hat{S}(t_j)$ que se utilizan para hacer las gráficas.

B) COMPARACIÓN DE LAS FUNCIONES DE DISTRIBUCIÓN EMPÍRICA Y TEÓRICA.

Como ya hemos indicado en el epígrafe 3, procederemos a comparar las funciones de distribución empírica y teórica de cada modelo y cada revista, donde estimamos los parámetros de los distintos modelos por el método de máxima verosimilitud, utilizando el programa LIMDEP 7.0. En el caso del modelo Log-Normal en vez de obtener las estimaciones de μ y σ hemos estimado las transformaciones $\lambda = e^{-\mu}$ y $p = 1/\sigma$. Los parámetros estimados así como sus errores estándar se muestran en la tabla III.

	Log-Normal		Weibull		Log-Logistic		Expon.
	λ	p	λ	β	ρ	γ	λ
ECONOMET THEOR	0.1443 (0.0055)	1.2843 (0.0391)	0.1134 (0.0031)	1.8730 (0.0835)	0.1414 (0.0047)	2.3578 (0.0926)	0.1042 (0.0644)
ECONOMETRICA	0.0414 (0.0010)	0.9423 (0.0158)	0.0450 (0.0010)	1.8406 (0.0436)	0.0498 (0.0011)	1.9451 (0.0438)	0.0220 (0.0005)
INSUR MATH ECON	0.1790 (0.0104)	1.2598 (0.0669)	0.1325 (0.0062)	1.6245 (0.1300)	0.1774 (0.0098)	2.1790 (0.1518)	0.1302 (0.0122)
J APPL ECONOM	0.1399 (0.0042)	1.5429 (0.0474)	0.1134 (0.0025)	2.2039 (0.0949)	0.1373 (0.0036)	2.7925 (0.1065)	0.1032 (0.0064)
J BUS ECON STAT	0.1235 (0.0028)	1.5552 (0.0308)	0.1033 (0.0017)	2.2039 (0.0949)	0.1224 (2.8703)	0.0024 (0.0837)	0.0873 (0.0041)
J ECONOMETRICS	0.1012 (0.0018)	1.2605 (0.0212)	0.0851 (0.0012)	1.9364 (0.0487)	0.1026 (0.0015)	2.2746 (0.0506)	0.0648 (0.0018)
J MATH ECON	0.0671 (0.0054)	0.8077 (0.0454)	0.0569 (0.0042)	1.3225 (.1056)	0.0717 (0.0050)	1.4865 (0.1070)	0.0458 (0.0033)
J ROY STAT SOC A STA	0.0723 (0.0037)	0.9063 (0.0336)	0.0626 (0.0028)	1.4954 (0.0817)	0.0771 (0.0033)	1.6891 (0.0824)	0.0471 (0.0025)
OXFORD B ECON STAT	0.1134 (0.0039)	1.3867 (0.0561)	0.0943 (0.0025)	2.1298 (0.0995)	0.1116 (0.0035)	2.4703 (0.1109)	0.0767 (0.0047)
REV ECON STAT	0.0742 (0.0020)	0.9985 (0.0237)	0.0639 (0.0017)	1.6015 (0.0563)	0.0778 (0.0019)	1.8083 (0.0574)	0.0459 (0.0015)

Tabla III: Parámetros estimados y errores estándar.

Las funciones de distribución teóricas serán $F_i(t) = 1 - S_i(t)$, $i = 1,2,3,4$, donde llamaremos como antes modelos 1, 2 y 3 a los casos Log-Normal, Weibull y Log-Logistic respectivamente y modelo 4 al Exponencial. Para el modelo Log-Normal, se han estimado los parámetros $\lambda = e^{-\mu}$ y $p = 1/\sigma$, por lo que hemos de tener en cuenta que la función de distribución reparametrizada sería:

$$F_i(t) = 1 - S_i(t) = \Phi\left((\log t - \mu)/\sigma\right) = \Phi\left((1/\sigma)\log\left(te^{-\mu}\right)\right) = \Phi\left(p\log(\lambda t)\right)$$

La función de distribución empírica para cada revista se obtiene a partir de la función de supervivencia empírica, la cual se estima utilizando la fórmula de Kaplan-Meier como se indicó en el epígrafe 3.

Así, para cada modelo y revista, definiremos unos "errores" que serán las diferencias entre las funciones de distribución empíricas y las estimaciones paramétricas evaluadas en los puntos L_j, $j = 1,...,10$, calculando posteriormente como medida de bondad de ajuste la raíz cuadrada del error cuadrático medio (RECM). Concretando más, para cada revista, definimos $e_{ij} = \hat{F}(L_j) - F_i(L_j)$, $i = 1,...,4$, $j = 1,...,10$ y calculamos $RECM_i = \sqrt{\sum_{j=1}^{10} e_{ij}^2 n_j \Big/ \sum_{j=1}^{10} n_j}$, que medirá el grado de bondad de ajuste del modelo i-ésimo.

A modo de ejemplo, ofrecemos en la tabla IV los resultados completos para la primera revista de las consideradas, donde puede apreciarse que el modelo que presenta una mejor bondad de ajuste es el Weibull, seguido muy de cerca por los modelos Log-Logistic y Log-Normal. El coeficiente RECM aumenta sensiblemente en el caso del modelo exponencial, volviendo este dato a mostrar que este modelo uniparamétrico no consigue captar bien el comportamiento de los datos.

$L_{j-1} - L_j$	t_j	n_j	$\hat{F}(L_j)$	Log-Normal $F_1(L_j)$	Weibull $F_2(L_j)$	Log-Logistic $F_3(L_j)$	Exponencial $F_4(L_j)$
[0-1]	0.5	3	0,0145	0,0065	0,0168	0,0098	0,0989
[1-2)	1.5	16	0,0918	0,0553	0,0602	0,0484	0,1881
[2-3)	2.5	23	0,2029	0,1412	0,1243	0,1169	0,2684
[3-4)	3.5	27	0,3333	0,2402	0,2035	0,2069	0,3408
[4-5)	4.5	28	0,4686	0,3376	0,2921	0,3063	0,4060
[5-6)	5.5	11	0,5217	0,4267	0,3850	0,4042	0,4647
[6-7)	6.5	14	0,5894	0,5053	0,4774	0,4939	0,5177
[7-8)	7.5	13	0,6522	0,5733	0,5654	0,5721	0,5654
[8-9)	8.5	15	0,7246	0,6315	0,6462	0,6384	0,6084
[9-10)	9.5	12	0,7826	0,6813	0,7179	0,6935	0,6471
Más de 10		45					
Total		207					
RECM				0,0372	0,0183	0,0231	0,0936

Tabla IV: Funciones de Distribución Empírica y ajustadas de la revista ECONOMET THEOR

Los valores de los coeficientes $RECM_i$ para cada una de las revistas estudiadas así como el promedio de todos ellos se ofrece en la tabla V, donde se han resaltado con fondo oscuro los mínimos para cada revista. En la última fila, aparecen los promedios de RECM del conjunto de revistas correspondientes a cada modelo.

	Log –Normal $RECM_1$	Weibull $RECM_2$	Log-Logistic $RECM_3$	Exponencial $RECM_4$
ECONOMET THEOR	0,0372	0,0183	0,0231	0,0936
ECONOMETRICA	0,0056	0,0026	0,0025	0,0290
INSUR MATH ECON	0,0157	0,0331	0,0194	0,0735
J APPL ECONOM	0,0318	0,0174	0,0187	0,1111
J BUS ECON STAT	0,0298	0,0132	0,0189	0,1082
J ECONOMETRICS	0,0113	0,0148	0,0096	0,0695
J MATH ECON	0,0114	0,0185	0,0141	0,0239
J ROY STAT SOC A STA	0,0150	0,0193	0,0153	0,0359
OXFORD B ECON STAT	0,0465	0,0371	0,0415	0,0896
REV ECON STAT	0,0082	0,0180	0,0131	0,0368
Promedios	**0,0213**	**0,0192**	**0,0176**	**0,0671**

Tabla V: RECM para los distintos modelos y revistas.

Como puede apreciarse, y como ya se puso de manifiesto en el análisis gráfico, el modelo exponencial no es adecuado para nuestros datos y el comportamiento de los tres modelos biparamétricos es muy similar y resulta prácticamente imposible decir que uno de ellos consigue ajustar mejor de manera global. Además, podemos ver cómo la desventaja observada para el modelo Log-Normal en el análisis gráfico no se aprecia claramente con el criterio de RECM ya que consigue el mejor ajuste en cuatro de las diez revistas; dicha desventaja, como ya indicamos anteriormente, parece deberse más bien a que la transformación de la función de supervivencia usada en este caso (que se basa en la función de distribución inversa de la Normal tipificada) es "menos linealizadora" que la utilizada en los casos Weibull y Log-Logistic (que se basan en la función logaritmo), si bien también es cierto que es el modelo que presenta un mayor RECM medio para el conjunto, debido sobre todo a que en las revistas ECONOMET THEOR, J APPL ECONOM y J BUS ECON STAT el valor de $RECM_1$ resulta ser sensiblemente superior a $RECM_2$ y $RECM_3$. Lo contrario ocurre en el modelo Log-Logistic, que a pesar de conseguir el mejor ajuste en sólo dos casos, presenta el menor promedio, debido a que en ninguna revista $RECM_3$ es sensiblemente superior a $RECM_1$ y $RECM_2$.

5. Discusión y conclusiones.

El análisis de datos de supervivencia o duración se muestra como una herramienta interesante para el estudio bibliométrico, más concretamente, para la modelización de la variable T que proporciona la antigüedad de las citas recibidas por los trabajos científicos; aunque el uso de datos retrospectivos hace que aparentemente la situación sea distinta de la habitual a la de los modelos de duración (pues primero se observa la cita y luego buscamos el momento de la publicación) al final los datos de antigüedad son igualmente interpretables como el tiempo que transcurre desde que comienza un determinado experimento (momento en el que es publicado el artículo) hasta que ocurre un determinado suceso (el artículo es citado).

Evidentemente, para la búsqueda de un modelo es preferible disponer de un conjunto de datos en los que no exista censura. No obstante, si queremos utilizar las bases de datos elaboradas por el ISI (una de las más completas e importantes que existen actualmente) esto no resulta posible, por lo que debemos aprovechar todas las herramientas previamente desarrolladas para los modelos de duración en presencia de censura para el análisis de nuestros datos de antigüedad de las citas. Observemos que el uso de datos retrospectivos también hace que la interpretación de la censura sea distinta de la habitual, ya que en nuestro caso conocemos cuándo ha ocurrido el suceso (ser citado) y desconocemos el momento de comienzo del experimento (año de publicación), aunque a efectos prácticos el tratamiento es exactamente el mismo.

El análisis gráfico de bondad de ajuste de los modelos se revela útil como primera aproximación simple al problema, aunque siempre teniendo en cuenta las limitaciones del mismo. Este primer procedimiento ya pone de manifiesto que el modelo exponencial, si bien presenta la ventaja de su simplicidad, se presenta como insuficiente para el ajuste de nuestros datos, mientras que los tres modelos biparamétricos estudiados (Log-Normal, Weibull y Log-Logistic) consiguen grados de ajuste aceptables y muy similares entre sí, con una ligera desventaja en contra del modelo Log-Normal, que puede ser debida a la naturaleza de la función utilizada para linealizar $\hat{S}(t)$, ya que en este caso no interviene la función logaritmo, mientras que en los otros dos modelos sí que lo hace.

El estudio de ajuste a través de la función de distribución (que resulta equivalente al ajuste a través de la función de supervivencia) tras estimar por el método de máxima verosimilitud los parámetros de los modelos, nos lleva prácticamente a las mismas conclusiones que el análisis gráfico, siendo los valores de RECM muy parecidos para todas las revistas. Indiquemos también que desde un punto de vista teórico, los modelos Weibull y Log-Logistic presentan la ventaja de que sus funciones de distribución y supervivencia tienen expresiones explícitas simples, cosa que no ocurre en el caso Log-Normal, si bien desde un punto de vista práctico esto no supone gran diferencia ya que podemos evaluar las funciones numéricamente en cualquier punto.

Si nos atenemos al criterio del RECM medio para el conjunto de revistas, hemos de concluir que el modelo que mejor se adapta a este conjunto particular de datos es el Log-Logistic, siendo el modelo Log-

Normal el que muestra un comportamiento peor; además también ocurre que el modelo Log-Logistic no presenta un valor de RECM sensiblemente superior al resto en ninguna de las revistas, hecho que ocurre en el modelo Log-Normal para las revistas ECONOMET THEOR, J APPL ECONOM y J BUS ECON STAT. Por su parte, el modelo Weibull presenta un valor de RECM sensiblemente superior a los otros dos para la revista INSUR MATH ECON. De todos modos, esta última conclusión debe hacerse siempre teniendo en cuenta que en todo caso las diferencias entre los tres modelos son mínimas.

Bibliografía.

1. Basulto, J. y Ortega, F.J. (2002), "Modelización de la antigüedad de las citas en la literatura científica con datos censurados a la derecha", *Revista Española de Documentación Científica*, **25, nº 2**, p. 141 – 150.

2. Brookes, B.C. (1974), "Obsolescence of special library periodical: sampling errors utility contours", *Journal of the American Society for Information Science*, **21, nº 5**, p. 320 – 329.

3. Burrel, Q.L. (2002), "Modelling citation age data: Simple graphical methods from reliability theory", *Scientometrics*, **55, nº 2**, p.273 – 285.

4. Egghe, L. y Ravichandra Rao, I.K. (1992), "Citation age data and the obsolescence functions: Fits explanations", *Information Processing & Management*, **28, nº 2**, p.210 – 217.

5. Gupta, B.M. (1998), "Growth and obsolescence of Literature in theoretical Population Genetics", *Scientometrics*, **42, nº 3**, p. 335 – 347.

6. Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

7. Ruiz, R. y Jiménez, E. (1996), "Envejecimiento de la Literatura Científica en Documentación. Influencia del Origen Nacional de las Revistas. Estudio de una Muestra", *Revista Española de Documentación Científica*, **19, nº 1**, p. 94 – 108.