Scriptum Digital
Vol. 3 (2014), pp. 29-63

DATACIÓN CRONO-GEOGRÁFICA DE DOCUMENTOS MEDIEVALES ESPAÑOLES*

YOSHIFUMI KAWASAKI
Universidad de Tokio

RESUMEN:
Este trabajo plantea un nuevo método para determinar la procedencia espacio-temporal de documentos medievales españoles, tratando a la vez tanto la variación cronológica como la geográfica. La datación se realiza en función del coeficiente de correlación calculado a partir del patrón global de coincidencia de rasgos lingüísticos entre documentos. La fecha estimada de composición del documento se computa por el promedio ponderado de la *data chronica* de *k* (≥ 1) documentos más afines a este en el uso lingüístico. Hemos logrado datar más del 60 % de documentos con un margen de error de ±20 años. Por otra parte, la adscripción diatópica ha conseguido un 60 % de precisión a nivel de región.
PALABRAS CLAVE: datación, historia de la lengua española, documentos archivísticos, lingüística de corpus, estadística

ABSTRACT:
This paper presents a new process for surmising dates and locations of medieval Spanish documents, while dealing with the diachronic and geographic linguistic variation simultaneously. The dating is carried out based on a correlation coefficient calculated by examining the coincidence between the texts in the linguistic characteristics. The date in which a document was issued is inferred by a weighted average of dates from the *k* (≥ 1) documents showing the strongest correlation. Through this method, more than 60 % of the documents were successfully dated within the margin of error of ±20 years. On the other hand, with respect to the region where the document was issued, a predictability of 60 % was achieved.
KEY WORDS: Dating, History of the Spanish language, Archival documents, Corpus linguistics, Statistics

1. INTRODUCCIÓN

Este trabajo tiene como objetivo plantear un nuevo procedimiento para datar automáticamente documentos medievales españoles con el método *k-NN* (*k-Nearest Neighbors*), que nos ha brindado no solo una capacidad predictiva bastante mejorada respecto a propuestas anteriores, sino también la posibilidad de realizar simultáneamente tanto la determinación cronológica como la adscripción geográfica del documento[1].

* Agradecemos al profesor Pedro Sánchez-Prieto Borja de la Universidad de Alcalá por facilitarnos la transcripción paleográfica de los documentos incorporados al *Corpus de Documentos Españoles Anteriores a 1700* (*CODEA*). Nuestro agradecimiento es también para el profesor Hiroto Ueda de la Universidad de Tokio y dos evaluadores anónimos que revisaron la primera versión de este artículo por sus valiosos comentarios y sugerencias que me han ayudado a mejorar el contenido del mismo. Para la revisión del artículo en español se ha contado con la ayuda de la profesora Ana Isabel García de la Universidad de Tokio. Este trabajo ha sido subvencionado por la *JSPS* (*Japan Society for the Promotion of Science*) KAKENHI Grant Number 13J03408 (This work was supported by JSPS KAKENHI Grant Number 13J03408).

[1] Hasta donde sabemos, el estudio de Azofra (2009: 201-204) constituye el primer intento en el ámbito hispánico de establecer el procedimiento para la datación. En Kawasaki (en prensa a) hemos empleado el método de concentración propuesto por Ueda (en prensa). Por otra parte, en Kawasaki (en prensa b, d) nos hemos servido del método de máxima verosimilitud a base de la distribución probabilística de parámetros lingüísticos, aunque contábamos con menor cantidad de parámetros.

Fecha de recepción: 09/2013/2013
Fecha de aceptación: 14/01/2014 ISSN: 2014-640X

Por la datación no solo se interesan los filólogos sino también historiadores y estadísticos. El equipo del proyecto *DEEDS*[2] (*Documents of Early England Data Set*) de la Universidad de Toronto ha desarrollado varios procedimientos estadísticos, de los cuales es de destacar el denominado *Maximum Prevalence* (*MP*), para datar automáticamente documentos (*charters*) medievales ingleses escritos en latín, que dispone como parámetros de todas las *k* palabras consecutivas (k-*shingle*) en toda la documentación (Feuerverger *et al.* 2005, 2008; Fiallos 1997, 2000; Gervers 1997, 2000a, 2000b; Tilahun 2011; Tilahun *et al.* 2012).

Ahora bien, la originalidad de nuestra investigación radica en la aproximación tanto filológica como estadística, que, por una parte, se propone establecer parámetros lingüísticos fundamentados en la gramática histórica y, por otra, aplicarles un tratamiento matemático, pues lo que mayor interés presenta para los historiadores de la lengua es realizar datación en función de los rasgos lingüísticos, a saber, paleográficos, gráfico-fonéticos, morfosintácticos y léxicos[3]. Por tanto, siendo idéntica la meta final de datación que es fechar correctamente textos antiguos sin *data chronica* para su indagación en campos relevantes, difiere nuestra aproximación de la del equipo canadiense en la manera de establecer parámetros con los que se realiza la datación[4]. Otra cosa que de nuestro procedimiento es digna de mencionar es el uso de la transcripción paleográfica, a diferencia del proyecto *DEEDS* que utiliza la versión crítica, realizada con rigor académico de acuerdo con los criterios establecidos por la red *CHARTA*[5] (*Corpus Hispánico y Americano en la Red: Textos Antiguos*), lo que permite minimizar la posible distorsión textual por parte de editores modernos. Esto tiene relevancia para los filólogos, pues, como señala Ueda (2013a), incluso la abreviación encierra cierta información sobre la circunstancia en la que se compuso el documento.

2. CORPUS

Para el presente estudio hemos utilizado documentos en su versión paleográfica incorporados al *Corpus de Documentos Españoles Anteriores a 1700* (*CODEA*) del *Grupo de Investigación de Textos para la Historia del Español* (*GITHE*) de la Universidad de Alcalá, dirigido por el Prof. Pedro Sánchez-Prieto Borja. De los 1502 documentos que componen dicho corpus, hemos seleccionado para el presente estudio 1026 que presentan tanto *data chronica* como *data tópica* con la exclusión de aquellos carentes de una de las dos o de ambas. Se observa en la Figura 1 un número elevado de textos en el siglo XIII que se contrasta con la escasez documental en los siglos XII y XVII. En lo que concierne a la distribución geográfica, si bien hay documentación de casi todo el espacio ocupado hoy día por el español y sus variedades relacionadas, son Ávila, Burgos, Guadalajara, León,

[2] http://deeds.library.utoronto.ca/

[3] Sobre la estilometría en el ámbito japonés, incluyendo temas como la datación, la identificación de autores y la falsificación de textos, véanse Jin (2009) y Murakami (1994, 2002 y 2006).

[4] Es de suponer que, de no ser estadísticamente, la datación de documentos latinos resultaría infactible, ya que, siendo una lengua muerta, presentaría escasísima, si no nula, variación cronológica, como lo hace la lengua viva, de manera que no se puedan establecer parámetros lingüísticos.

[5] http://www.charta.es/criterios-de-edicion-/

Scriptum Digital
Vol. 3 (2014), pp. 29-63

Madrid, Salamanca, Sevilla, Toledo, Valladolid y Zaragoza las provincias que proporcionan mayor cantidad de documentos.

	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Álava											1																1
Albacete																				3							3
Almería																				1							1
Asturias					2		3			1	1		1	1		1					1						11
Ávila				1	1	24	23				1						2										52
Badajoz									1										1		1						3
Burgos	2		4	8	7	9	5	2		2	1	1		3		3	8	5			1						61
Cáceres						1		1	5	4	2	1	1	1				1			1						18
Cádiz					1		1					2			1	1	2				7						15
Cantabria		1			1	2	1		1	2	6	2			3	4	5	1									29
Córdoba				1		2								1		1	2		1								8
Cuenca			1																			2	1	4			8
Granada															1	2		4		4			4	2			17
Guadalajara				1		4	2				1	1	1	1	3	2	3	2	1	9	2	5	8	6	1	6	59
Guipúzcoa												1					1			2							4
Huelva																		1			2						3
Huesca	1				3	1	1			2	10		1	1		2											22
Islas Baleares																					1						1
Italia																					13	1					14
Jaén								1							1	3		1		1			4				11
La Rioja			1	1	4	3	2	1	1	1		3	1	2		1	2	2									25
León	1		2	4	6	1	1				8	2	1	3	3	1	10					2	2	1	1		49
Lugo			2																								2
Madrid						1		1	1		4	2			1	4	7	6	6	9	19	2	5	4	3	2	77
Málaga																			2		2		1				5
Murcia							1											1			2						4
Navarra					11	5	5	2	3		1	2		1		1	4	1									36
País Vasco					1	3												2									6
Palencia	1			2	2	2						3		2	3		2					1					18
Portugal											1										2						3
Salamanca			2	5			10	1	5	3	1	4	4	1	9	1	4				2		1	5			58
Segovia			1	1	3	6						2			6		4	2									25
Sevilla			1	9	3	3	1	1	3	4				1	1	1	2	8	7		12			1			58
Soria					1				1							1				1							4
Teruel							1		2	7	5	11	4	4	3			2									39
Toledo	1				3	10		4	3	2	1		6	4	6	2	1	12	6				1				62
Valencia																										1	1
Valladolid				2	2	28	7		3	5	2	9		1	2	10	11	11	6	5		5	3				112
Vizcaya																	4										4
Zamora			1	1		4	1		2	1	3	2	5	2				1	1	1							25
Zaragoza			2	3	1	1	2	2	12	4	6	21	2	3		4	3	3		1				1	1		72
Total	6	1	17	39	52	110	67	16	43	38	55	69	27	32	43	45	77	66	31	37	68	18	30	24	6	9	1026

Figura 1. Distribución espacio-temporal de los documentos (número de piezas)

Por otra parte, la Figura 2 muestra la distribución cronológica de los textos de acuerdo con la tipología documental: Cancilleresco, Eclesiástico, Judicial, Municipal y Particular. Los textos cancillerescos, por la naturaleza itinerante de la cancillería, se considera que no reflejan la modalidad lingüística de la localidad donde se emitió el documento, mientras que los otros tres tipos sí lo hacen, de ahí que estos contribuyan en mayor medida a estudiar la dialectología histórica.

	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Cancilleresco	1		6	17	13	60	16	7	11	8	6	11	1	8	13	14	21	15	14	8	6		1	2	3	1	263
Eclesiástico	4	1	8	17	26	21	13	7	17	20	31	31	14	12	16	17	17	7	3	2	5	13	13	11	2	1	329
Judicial			1		1	4	3		3		1			3	1	2	11	13	5	7	16	4	5	5	1	5	91
Municipal				1	2	1			2		2	4	4	2	5	2	9	16	7	8	3		6				74
Particular	1		2	4	10	24	35	2	10	10	15	23	8	7	8	10	19	15	2	12	38	1	5	6		2	269
Total	6	1	17	39	52	110	67	16	43	38	55	69	27	32	43	45	77	66	31	37	68	18	30	24	6	9	1026

Figura 2. Distribución cronológica de los documentos según tipología documental

Debemos hacer notar que la predecibilidad depende en gran medida de la distribución tanto crono-geográfica como tipológica de los documentos incluidos en el corpus. Al aplicar el método *k-NN* al *Corpus de documentos de Cancillería Real* (*CODCAR*) elaborado por el *Grupo de Estudio de Documentos Históricos y Textos Antiguos de la Universidad de Salamanca* (*GEDHYTAS*), dirigido por la Profa. María Nieves Sánchez González de Herrero, cerca del 80 % de los 538 documentos resultaron datados con un margen de error de ±10 años respecto a la fecha verdadera, siendo el promedio, la mediana y la media cuadrática del margen de error absoluto de 6, 4 y 10 años, respectivamente (Kawasaki en prensa c). Es de señalar que esta elevada predecibilidad se debe al corto periodo abarcado por el corpus, menos una centuria, la relativamente alta densidad documental y la homogeneidad tipológica.

A continuación reproducimos la estadística descriptiva de los tres corpus junto con la de la datación (Figura 3)[6]. Se aprecia la escasa densidad documental del *CODEA* frente a los otros dos, lo que podría incidir en la predecibilidad aunque se desconoce cómo y cuánto.

	CODEA	*CODCAR*	*DEEDS*
Periodo	1109-1697	1223-1311	1089-1438
Documentos datados	1026	538	3353
Densidad	1.8	6.2	9.6
Data chronica			
Promedio	1431	1272	1237
Mediana	1421	1272	1237[7]
Desviación estándar	124	17	46
Método de datación	k-NN6	k-NN4	MP2
Margen de error absoluto			
Promedio	21.3	6.4	9.0
Mediana	14.0	4.0	6.0
Media cuadrática	31.3	9.5	14.7

Figura 3. Estadística descriptiva de los corpus y de la datación

[6] En realidad, no es apropiado utilizar el promedio, mediana y desviación estándar para la descripción de *data chronica*, ya que tanto el *CODEA* como el *CODCAR* presentan una distribución multimodal, a diferencia del DEEDS que la presenta unimodal.

[7] Aunque Tilahun *et al.* (2012) no registra explícitamente la mediana de la *data chronica* de los documentos en el *DEEDS*, dada la distribución normal de los mismos, no es arriesgado conjeturar que la mediana no se aparta mucho del promedio que es de 1237.

3. PARÁMETROS

Los parámetros son rasgos lingüísticos que consideramos útiles a la hora de realizar la datación. Son de índole variada, a saber, gráfica, abreviativa, fonética, morfosintáctica, léxica y formularia. Por ahora contamos con cerca de 300 parámetros, la mayoría de los cuales han sido establecidos con base en estudios anteriores, de los que debemos mucho a Alvar (1996), Díaz Moreno *et al.* (en prensa), Menéndez Pidal (1999), Penny (2002), Sánchez-Prieto (1998 y 2012), Sánchez-Prieto *et al.* (2012) y Zamora Vicente (1967).

Se ha de notar que, salvo con los parámetros gráficos, no hacemos distinción entre las variantes gráficas, por ejemplo, entre *tobe* y *tove* en contraste con *tuve* respecto a la vocal del tema pretérito, de modo que en el parámetro *tove* se encuentren tales formas como *tobe*, *touiere<n>*, *toujemos*, *toujeron*, *toujero<n>*, *touo*, *tove*, *tovieron*, ecétera. La Figura 4 muestra la variación cronológica entre *tove* y *tuve*. Se observa la emergencia y paulatina sustitución de *tuve* por la variante antigua.

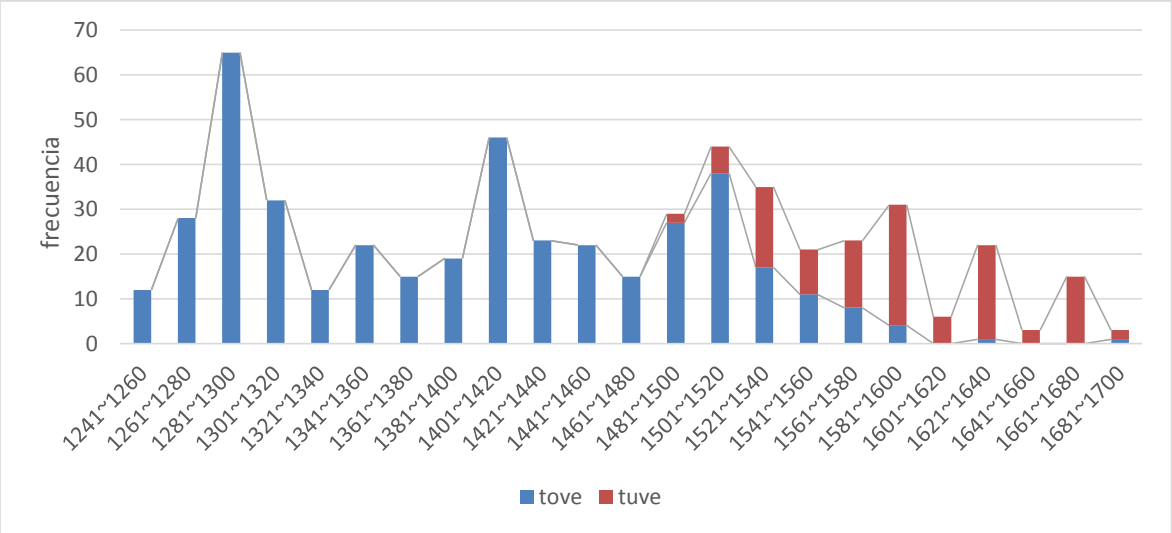

Figura 4. Variación cronológica entre *tove* y *tuve*

Nótese también que en algunos casos hemos ignorado formas desarrolladas, contando únicamente las plenas. A modo de ejemplo, respecto a la vocal penúltima, claros testimonios del diptongo o monoptongo son *castiella* o *castilla*, respectivamente, pero no *cast<ie>lla*, *cast<i>lla*, formas estas reconstruidas de acuerdo con las prácticas escriturarias corrientes en la época dada. A continuación enumeramos los parámetros establecidos.

3.1. Gráfico

El uso de *-pu-* por *-u-* (*escripuir*); la variante gráfica latinizante *subçesor* y *subçeder*; *-yn-* para palatal nasal (*ayno*); *-ny-* para palatal nasal (*anyo*); *-yll-* para palatal lateral (*aqueylla*); *-lly-* para palatal lateral (*fillyo*); *-quo-* para /ko/ (*ecclesiastiquo*) y *quoa-* para /k^{w}a/ (*quoal*); *v-* o *u-* inicial (*viere~uiere*); *h-* antietimológica en *huno* por *uno*; *ermano* sin *h-* inicial; *-d~-t* final (*merced~mercet*, *verdad~verdat*, *salud~salut*, *heredad~heredat*) (Kawasaki 2013a); el uso antietimológico de *-nt~-nd* (*algund~algunt*, *segund~segunt*); el uso de *z-* inicial

(*ziudad*) por *ç-*; *-ka-* y *-ke-* para /ka/ y /ke/, respectivamente (*kasa*, *akell*); *-th-* por *-t-* (*thener*); *-m* final por *-n* (*donaçiom*).

3.2. Abreviaturas

Como apunta Ueda (2013), las abreviaturas también presentan variación espacio-temporal, la explotación de la cual se ve propiciada por el uso de la versión paleográfica (*tiempo~tie<m>po~t<iem>po*; *quien~q<ui>en*; *escribano~esc<ri>bano*; *tierra~t<ie>rra*, etcétera). Podrán incluirse más en las futuras investigaciones.

3.3. Gráfico-fonética

3.3.1. Vocálica

-AU-, -AL-, -UO- > *-ou-* (*cousa*, *outro*, *dous*); -AI-, -A(V)I- > *-ei-* (*meyrino*, *eredey*); el mantenimiento de *-e* final (*abade*, *uoluntade*); la no diptongación de Ŏ y Ĕ latinas (*bon*, *logo*, *sempre*); la diptongación de Ĕ latina ante yod (*viengo*, *tiengo*); la caída de *-y-* intervocálica (*maor*, *maordomo*); la reducción de dos vocales idénticas (*seer~ser* junto a *seyer*, *seellar~sellar*); *re* por *rey*; *cuemo~como*; *decho~dicho*; *nenguno~ninguno*; *mesmo~mismo*; *monasterio~monesterio*; *lugar~logar*; la reducción de *-iello* a *-illo* (*castiella*, *castilla*); *-u* final (*otru*, *todu*); fonética sintáctica (*cona reyna*, *eno camjno*), etcétera.

3.3.2. Consonántica

El trueque de lateral por vibrante (*pubrico*, *conprir*); la solución en *-m-*, *-mn-*, *-mb-* del grupo consonántico románico -M'N- (*nome~nomne~nombre*); la llamada *-l-* leonesa (*dulda*, *selmana*, *julgar*); -LY- > *-y-* (*fiyo*, *meyor*); -LY- > *-ll-* (*fillo*, *muller*); CL- > *x-* (*xamar*); conservación de CL- y PL- inicial latina (*clamar*, *plegar*); grupo consonántico final *-rt, -nt* (*cort*, *vint*); grupo consonántico final *-ls*, *-rs*, *-ns* (*tals*, *labors*, *bens*); -CT- > *-it-*, -ULT- > *-uit-* (*feito*, *dito*, *muyto*); el mantenimiento del grupo -CT- (*sobredicto*, *maldicto*); *-tz* final (*totz*, *tengatz*); *-ç* final (*voç*, *fiç*); *autoridad~actoridad~auctoridad*; *propio~proprio*; *-s-* implosiva o *-z-* implosiva (*conosco~conozco*); *-d-* implosiva o *-z-* implosiva (*judgar~juzgar*, *-adgo~-azgo*); BADALLOCIO > *badalloz~badajoz*; la vocalización de la velar implosiva (*regno~reyno*); *f-~ff-~h-~ø-* inicial (*fazer*, *fijo*, *guadalfajara*); *gelo~selo* de ILLĪ ILLUM; *-bd-* o *-ud-* (*cibdad~ciudad*, *debda~deuda*); la representación gráfica de la palatal fricativa (*mugier~muger~mujer*); el mantenimiento o pérdida de la consonante palatal inicial *janero~enero*; el mantenimiento de la *-m* implosiva (*comde*); *cossa* por *cosa*; el ensordecimiento y velarización de la palatal sonora (*xamas*, *relixion*); la palatalización o caída de la implosiva velar (*sinal*), etcétera.

3.4. Morfo-fonético, morfosintáctico

pois < POST; *mais* < MAGIS; *ata* o *fata* por *fasta*; *despues~depues*; *eu*, *you* < EGO; *vinte*, *trinta* < VĪGINTĪ, TRĪGINTĀ por *veinte*, *treinta*; el pronombre femenino *sua*; *dizer* < DICĔRE por *dezir*; la forma analógica *saban* por *sepan*; la síncopa del futuro de subjuntivo en las personas

primera y segunda de plural (*dierdes*, *diermos*); *furon~foron* por *fueron*; *foi~foe* por *fui*; *fu* por *fui* o *fue*; la no diptongación en los paradigmas relacionados con el tema de perfecto (*quer*, *obere*); *tove~tuve*, *ove~uve*; *avía~avié*; *do~doy*, *so~soy*; *-iemos~-imos*, *-ieste~-iste*; la pérdida de la *-d-* en las formas llanas de segunda persona de plural (*avedes~aveis*); el sufijo adverbial (*-mientre~-miente~-mente*); *adelantre* por *adelante*; *convosco*, *connosco*; *pora~para*; *asi~asin~ansi*; *agora~ahora~aora*; *non~no* (Moreno Bernal y Horcajada 1997); *vala~valga*, *traya~traiga*; *traxo~truxo~troxo*; *complir~cumplir*; *hemos~avemos*; *nosso* por *nuestro*; *viren*, *oyren* por *vieron*, *oyeron*; la diptongación de Ĕ en *ser* (*ye*, *yera*); el posesivo *lur*; *fer~far* por *fazer*; el demostrativo *ço*; el pronombre indefinido *altre*; el demostrativo *aqueste* (Enrique-Arias 2012: 101-103); el mantenimiento de la sorda intervocálica (*toto*); el mantenimiento de la *-d-* intervocálica latina (*possedir*, *odran*); *dius* < DEORSUM; *sines*, *siene* < SĪNE; la metátesis o epéntesis consonántica en el futuro (*terne~tenre~tendre*); participio pasado en *-udo* (*conoçudo*, *tenudo*); subjuntivo de *ser* (*sia* frente a *sea*); *-as* > *-es* (*totes les*); *esti~este*, *aqueste~aquesti*, etcétera.

3.5. Léxico

Sustantivos abstractos (*guisa~manera*; *vegada~vez*); sustantivos concretos (*treudo*; *frau*; *alfayate~sastre*); *pedaço de tierra~pieza de tierra~quiñon* (Sánchez-Prieto 2012: 31-32); verbos (*trobar*, *lexar*); adverbios y preposición (*apres*, *ensemble*, *enton*, *ultra*); conjunción concesiva (*maguer~aunque*); adjetivo (*infrascripto*, *auandito~deuandicho*, *suso nombrado~pernominado* (Sánchez-Prieto 2008: 250)).

3.6. Frases formularias

Si bien no se trata de rasgos lingüísticos propiamente dichos, consideramos de utilidad la variación estilística de las frases formularias, latinas o romances, el empleo de las cuales se muestra vinculado a determinados territorios, épocas y tipología documental (Kawasaki 2013b). Hay variación en el encabezamiento tales como «Notu<m> sit om<n>jb<us> hom<in>jb<us> tam p<re>sentjb<us> q<ua>m fut<ur>ijs ...», «Con<n>oscida cosa sea a q<ua>ntos esta Carta uieren Como ...», «Sea conocida cosa a todos hom<e>s como ... », «Manifiesto sea atodos hom<ne>s Como ...», «Sepan q<ua>ntos esta Carta uieren & oyeren. Cuemo ...», «Sepan todos qua<n>tos aq<ue>sta publiq<ua> carta vera<n> q<ue>...», etcétera. Asimismo comprobamos la variación de las siguientes fórmulas, «anno domini»~«a nativitate»~«in era»~«sub era»; «in dei nomine»~«in xpisti nomine»~«in nomine domini»~«en el nombre de dios»; «dei gratia»~«por la gracia de dios»; «por la divina clemencia», etcétera.

4. DATACIÓN CRONOLÓGICA

4.1. Procedimiento

Una vez establecidos los parámetros, detectamos en cada documento la presencia o ausencia de los mismos a fin de componer la matriz de datos en la que ‘1’ y ‘0’ significa presencia y ausencia del parámetro dado respectivamente (Figura 5), proceso llevado a

cabo automáticamente por medio del programa de nuestra elaboración.

ID	Año	Provincia	Región	Tipología	SEER	SEYER	DEPUES	DESPUES	REGNO	REYNO	SEELAR	SELLAR	GUADALFAJARA	GUADALJARA	GELO	SELO	CASTIELLA	CASTILLA	PORA	PARA	CIUDAT	CIUDAD	LOGAR	LUGAR	TIEMPO	TIE<M>PO	T<IEN>PO	MUGIER
CODEA 1	1251	Sevilla	AN	C	1	0	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0
CODEA 2	1260	Córdoba	AN	C	0	0	1	1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1
CODEA 3	1262	Sevilla	AN	C	1	0	0	1	1	0	1	1	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1
CODEA 4	1277	Burgos	CV	C	0	0	0	1	1	0	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1
CODEA 5	1278	Segovia	CV	C	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
CODEA 6	1285	Burgos	CV	C	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	1
CODEA 7	1295	Valladolid	CV	C	0	1	0	0	1	0	1	0	1	0	1	0	1	0	0	1	0	0	0	1	1	1	0	0
CODEA 8	1383	Valladolid	CV	C	0	0	0	1	1	0	0	1	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0
CODEA 9	1387	Madrid	CN	J	0	1	0	0	1	0	1	1	1	0	1	0	1	0	0	1	0	1	1	0	0	0	1	0
CODEA 10	1392	Portugal	PT	C	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CODEA 11	1399	Guadalajara	CN	M	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
CODEA 12	1436	Guadalajara	CN	M	0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	0	0	1	0
CODEA 13	1458	Valladolid	CV	C	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0
CODEA 14	1460	Guadalajara	CN	C	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0
CODEA 15	1462	Valladolid	CV	C	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0
CODEA 16	1464	Guadalajara	CN	C	0	0	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	0	1	0	1	1	0
CODEA 17	1466	Madrid	CN	M	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
CODEA 18	1466	Guadalajara	CN	M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
CODEA 19	1467	Guadalajara	CN	M	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0
CODEA 20	1457	Palencia	CV	P	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0

Figura 5. Muestra de la matriz de presencia o ausencia de parámetros

A continuación, de esta matriz elaboramos otra del coeficiente de correlación llamado *phi* modificado (alias *Ochiai* o *Cosine*) que mide la afinidad entre documentos[8]. El coeficiente *phi* modificado, cuyo valor se encuentra entre cero y uno, se calcula bajo la siguiente fórmula:

$$0 \leq phi \text{ modificado } = \frac{a}{\sqrt{a+b}\sqrt{a+c}} \leq 1$$

donde *a* representa el número de parámetros compartidos por dos documentos A y B, *b* el número de parámetros presentes en el documento A pero no en B, y *c* el número de parámetros presentes en el documento B pero no en A. Cuanto mayor es el valor del coeficiente, o al acercarse a uno, mayor afinidad existe entre los dos documentos, y cuanto menor, o al acercarse a cero, menor afinidad entre los dos. El valor elevado del coeficiente se consigue cuando hay múltiples rasgos compartidos (*a*) y pocos rasgos no compartidos (*b* y *c*). Se presupone que dos textos de análoga procedencia espacio-temporal presenten un valor mayor que cuando son de distinta procedencia crono-geográfica, premisa esta que creemos más verosímil que la contraria.

A modo de ejemplo, si deseamos saber cuál de los dos documentos B o C se asemeja más a A, compárese el valor del coeficiente de correlación entre A y B (A:B) con el de correlación entre A y C (A:C). Nótese que es el patrón global de coincidencia de rasgos lingüísticos lo que incide en el mayor o menor grado de afinidad entre dos documentos.

[8] En Kawasaki (en prensa c) hemos demostrado el mayor grado de rentabilidad del coeficiente *phi* modificado sobre los demás.

		B	
		Sí	No
A	Sí	5(=a)	4(=b)
	No	3(=c)	

		C	
		Sí	No
A	Sí	5(=a)	6(=b)
	No	5(=c)	

$$phi \text{ modificado (A: B)} = \frac{a}{\sqrt{a+b}\sqrt{a+c}} = \frac{5}{\sqrt{5+4}\sqrt{5+3}} = \frac{5}{\sqrt{9}\sqrt{8}} \cong 0.589$$

$$phi \text{ modificado (A: C)} = \frac{a}{\sqrt{a+b}\sqrt{a+c}} = \frac{5}{\sqrt{5+6}\sqrt{5+5}} = \frac{5}{\sqrt{11}\sqrt{10}} \cong 0.477$$

Como el valor del coeficiente de correlación entre A y B (0.589) es superior al que se da entre A y C (0.477), juzgaremos que el documento A es más afín a B que a C.

Calculando el coeficiente de correlación entre todas las combinaciones, obtenemos la matriz del coeficiente de correlación (Figura 6)[9].

ID	Año	Provincia	Región	Tipología	CODEA1	CODEA2	CODEA3	CODEA4	CODEA5	CODEA6	CODEA7	CODEA8	CODEA9	CODEA10
CODEA 1	1251	Sevilla	AN	C	1.000	0.350	0.482	0.375	0.444	0.381	0.350	0.460	0.329	0.296
CODEA 2	1260	Córdoba	AN	C	0.350	1.000	0.559	0.735	0.429	0.501	0.375	0.426	0.355	0.343
CODEA 3	1262	Sevilla	AN	C	0.482	0.559	1.000	0.627	0.383	0.527	0.419	0.501	0.439	0.307
CODEA 4	1277	Burgos	CV	C	0.375	0.735	0.627	1.000	0.458	0.567	0.334	0.456	0.437	0.367
CODEA 5	1278	Segovia	CV	C	0.444	0.429	0.383	0.458	1.000	0.364	0.386	0.475	0.412	0.471
CODEA 6	1285	Burgos	CV	C	0.381	0.501	0.527	0.567	0.364	1.000	0.560	0.477	0.463	0.324
CODEA 7	1295	Valladolid	CV	C	0.350	0.375	0.419	0.334	0.386	0.560	1.000	0.453	0.409	0.343
CODEA 8	1383	Valladolid	CV	C	0.460	0.426	0.501	0.456	0.475	0.477	0.453	1.000	0.651	0.460
CODEA 9	1387	Madrid	CN	J	0.329	0.355	0.439	0.437	0.412	0.463	0.409	0.651	1.000	0.556
CODEA 10	1392	Portugal	PT	C	0.296	0.343	0.307	0.367	0.471	0.324	0.343	0.460	0.556	1.000

Figura 6. Muestra de la matriz del coeficiente de correlación

Luego se ordena en orden descendente la columna de cada documento para así extraer *k* (≥ 1) documentos que con este presentan la mayor afinidad. La fecha de composición se estima por el promedio ponderado de la *data chronica* de estos *k* documentos o *k-Nearest Neighbors*[10]. El valor de *k* no se determina *a priori* sino tras varias pruebas llevadas a cabo por distintos valores de *k*. La Figura 7 muestra la variabilidad respecto al promedio, media cuadrática y mediana del margen de error absoluto calculado|fecha estimada – fecha verdadera| en la datación realizada con distinto valor de *k* (= 1, 2, 3, …, 20). Si tomamos como término de comparación el promedio del margen de error absoluto (cuanto menor, mejor), el valor de *k* más apropiado será de 6. El promedio de 21 años quiere decir que los documentos se espera ser datados con un margen de error de ±21 años respecto a la *data chronica* verdadera.

[9] El cómputo ha sido realizado por el programa informático NUMEROS.xlsm para el análisis de datos numéricos, elaborado por el Prof. Hiroto Ueda de la Universidad de Tokio, que se puede obtener de forma gratuita en su página web http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/index.html.

[10] Respecto a la aplicación del método *k-NN* a la datación, véanse Feuerverger *et al.* (2005, 2008), Tilahun (2011) y Tilahun *et al.* (2012).

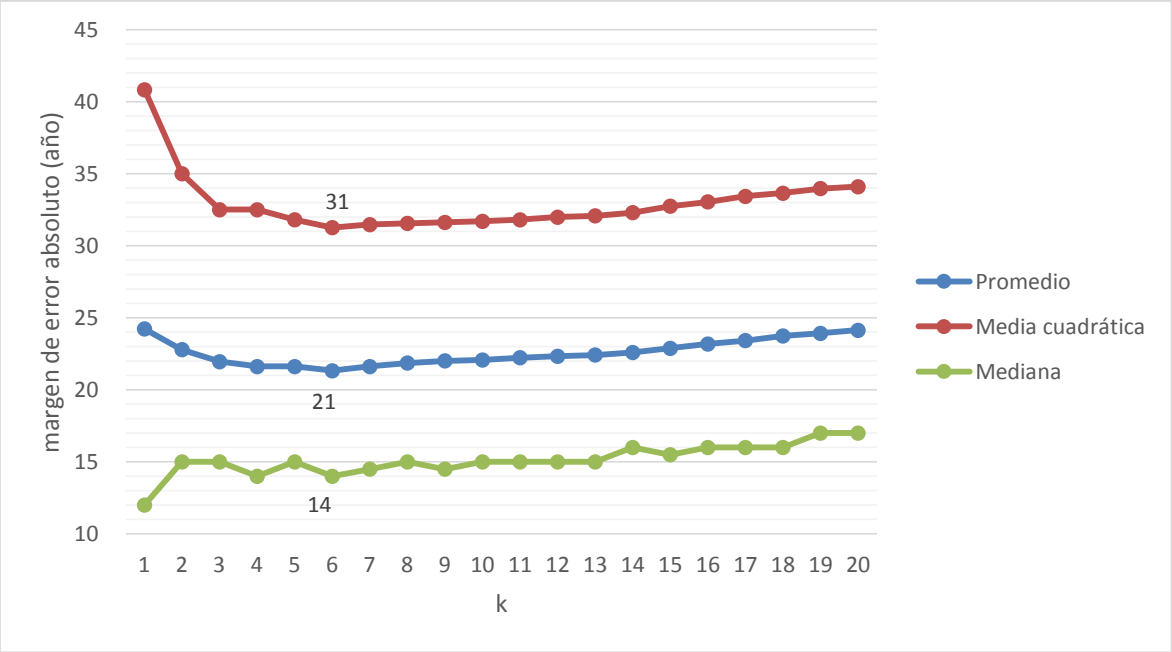

Figura 7. Promedio, media cuadrática y mediana del margen de error absoluto

A modo de ejemplo, realizamos la datación del documento CODEA1501 (Año 1515, Burgos) que por razones de espacio se reproduce parcialmente. En este texto encontramos los siguientes parámetros en orden de aparición: *çibdad* por *çiudad* con *-u-* o *çibdat* con *-t* final, *Castilla* por *Castiella*, *Reynos* por *Regnos*, *hija* por *fija*, *vall<adol>jd* por *vall<adol>jt* con *-t* final, *agora* por *ahora*, *lugar* por *logar*, *m<erce>d* por *m<erce>t* con *-t* final, *subçesor<e>s* por *suçesor<e>s*, *escriuano* por *esc<ri>uano* abreviado y *v<er>dad* por *v<er>dat* con *-t* final, los cuales van marcados en negrita:

> En la **çibdad** de burgos Cabeça de **Castilla** cam<ar>a dela Reyna n<uest>ra sen<n>ora lun<e>s honze dias del m<e>s de mayo an<n>o del nasçimj<ento> de n<uest>ro sen<n>or y salu<ado>r ih<es>u x<isto> de mjll e q<ui>nj<ento>s e q<ui>nze an<n>os estando en vna sala baxa delas casas del co<n>destable de **castilla** q<ue> son enla d<ic>ha **çibdad** donde posa el muy alto catholjco e muy poderoso p<ri>nçipe el Rey don f<ernan>d<o> n<uest>ro sen<n>or admjnjstrador e gov<er>nador destos **Reynos** de **castilla** de leon de granada etc por la muy alta e muy poderosa p<ri>nçesa la Reyna don<n>a juana n<uest>ra soberana sen<n>ora su**hija** y estando ay present<e>s el muy magnj<fico> y Reverendo sen<n>or don j<ua>n defonseca (…) porla villa de **vall<adol>jd** e antonjo de deça e el lj<cenciado> v<er>nal bazq<e>s de Ama<n> pr<ocuradore>s de cort<e>s porla **çibdad** de toro e j<ua>n debarrio nuevo e f<ernan>d<o> de morales pr<ocuradore>s de cort<e>s porla **çibdad** de sorja e don yn<n>jgo de arellano y el doctor françisco de medjna pr<ocuradore>s de cort<e>s[...][h. 3r] E alos v<ecino>s dellas conosçiesen d<e>sde **agora** los del co<n>sejo dela d<ic>ha Reyna don<n>a juana n<uest>ra sen<n>ora e admjnjstrasen justiçia alas d<ic>has **çibdad<e>s** e villas e **lugar<e>s** del d<ic>ho **Reyno** e (…) dixero<n> q<ue> en no<n>bre destos d<ic>hos **Reynos** de **castilla** e deleo<n> e de granada reçibieron la d<ic>ha **m<erce>d** q<ue> su alteza hazja ala reyna n<uest>ra sen<n>ora en sus **subçesor<e>s** e aestos d<ic>hos **Reynos** del d<ic>ho **Reyno** de navarra e (…) E yo El **dicho** bartolome Ruyz decastan<n>eda **escriuano** del consejo dela Reyna n<uest>ra sen<n>ora E **escriuano** delas d<ic>has cort<e>s p<re>sent<e> fuy alo q<ue> d<ic>ho es en vno conel d<ic>ho Secret<ario> pedro de

> quintana e co<n>los d<ic>hos pedro de cuaçola e luys delgadillo escri<vano>s delas d<ic>has cort<e>s E por ende fize aquj este mjo sig (signo)no en t<e>stim<on>j<o> de **v<er>dad** (bajo el signo) bartolome Ruyz (CODEA1501, Año 1515, Cancilleresco, Burgos, Castilla la Vieja).

Al calcular el coeficiente de correlación con los demás textos, obtenemos seis (*k*=6) documentos que presentan mayor coeficiente de correlación con este documento, que son CODEA292 (Año 1533), CODEA611 (Año 1517), CODEA1278 (Año 1508), CODEA1385 (Año 1485), CODEA25 (Año 1513) y CODEA1426 (Año 1521) exceptuando el CODEA1501, el propio documento a datar, con el que presenta, obviamente, el valor 1 (Figura 8).

ID	Año	Provincia	Región	Tipología	CODEA1501
CODEA1501	1515	Burgos	CV	C	1.000
CODEA292	1533	Toledo	CN	P	0.573
CODEA611	1517	La Rioja	CV	E	0.567
CODEA1278	1508	Burgos	CV	C	0.523
CODEA1385	1485	Valladolid	CV	C	0.502
CODEA25	1513	Valladolid	CV	C	0.499
CODEA1426	1521	Navarra	NA	P	0.485

Figura 8. Los seis documentos más afines al CODEA1501 (*k*=6)

Computando el promedio ponderado de 1533, 1517, 1508, 1485, 1513 y 1521 en proporción con el valor del coeficiente respecto a la suma total de los seis coeficientes que es de 3.149 (=0.573+0.567+0523+0.502+0.499+0.485), la fecha estimada resulta ser de *1513, apenas alejada de la verdadra 1515[11].

$$1533 \times \frac{0.573}{3.149} + 1517 \times \frac{0.567}{3.149} + 1508 \times \frac{0.523}{3.149} + 1485 \times \frac{0.502}{3.149} + 1513 \times \frac{0.499}{3.149} + 1521 \times \frac{0.485}{3.149} = 1513$$

De este modo la fecha estimada de un documento se obtiene calculando el promedio ponderado de la *data chronica* de *k* (≥ 1) documento(s) que con el documento a datar presenta(n) la mayor semejanza, cuya fórmula se expresa de la siguiente manera:

$$\sqrt[h]{\sum_{i=1}^{k}\left((\text{fecha}_i)^h \times \frac{\text{coeficiente}_i}{\sum_{i=1}^{k}\text{coeficiente}_i}\right)}$$

Hemos de hacer notar que este método requiere que el coeficiente tome valor positivo. En el presente estudio fijamos el exponente *h* (≥ 1) a 1.

[11] El asterisco * indica la fecha o lugar estimados.

Scriptum Digital
Vol. 3 (2014), pp. 29-63

4.2. Resultado

4.2.1. Sin distinción tipológica

El resultado de la datación realizada con distinto valor de *k* (= 1, 2, …, 20) con los 1026 documentos que tienen tanto *data chronica* como *data topica* aparece resumido en la Figura 9. Sin distinción tipológica, el promedio, desviación estándar y mediana de la *data chronica* de estos documentos resultan ser de 1431, 124 y 1421 años, respectivamente.

Margen de error	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
±5 años	356/1026 35%	292/1026 28%	247/1026 24%	260/1026 25%	257/1026 25%	260/1026 25%	247/1026 24%	223/1026 22%	228/1026 22%	224/1026 22%
±10 años	481/1026 47%	431/1026 42%	417/1026 41%	414/1026 40%	415/1026 40%	419/1026 41%	416/1026 41%	406/1026 40%	390/1026 38%	381/1026 37%
±20 años	644/1026 63%	619/1026 60%	628/1026 61%	640/1026 62%	638/1026 62%	642/1026 63%	629/1026 61%	631/1026 62%	637/1026 62%	634/1026 62%
±30 años	726/1026 71%	748/1026 73%	774/1026 75%	771/1026 75%	777/1026 76%	777/1026 76%	775/1026 76%	769/1026 75%	763/1026 74%	767/1026 75%
±40 años	805/1026 78%	844/1026 82%	864/1026 84%	864/1026 84%	866/1026 84%	871/1026 85%	868/1026 85%	873/1026 85%	870/1026 85%	876/1026 85%
±50 años	864/1026 84%	914/1026 89%	922/1026 90%	929/1026 91%	929/1026 91%	932/1026 91%	933/1026 91%	928/1026 90%	921/1026 90%	921/1026 90%
±100 años	996/1026 97%	1006/1026 98%	1009/1026 98%	1010/1026 98%	1012/1026 99%	1015/1026 99%	1013/1026 99%	1012/1026 99%	1012/1026 99%	1012/1026 99%
Promedio	24	23	22	22	22	21	22	22	22	22
Media cuadrática	41	35	33	33	32	31	31	32	32	32
Mediana	12	15	15	14	15	14	15	15	15	15
Máximo	388	221	200	196	194	195	193	191	190	188
Mínimo	0	0	0	0	0	0	0	0	0	0

Margen de error	k=11	k=12	k=13	k=14	k=15	k=16	k=17	k=18	k=19	k=20
±5 años	223/1026 22%	210/1026 20%	207/1026 20%	208/1026 20%	197/1026 19%	203/1026 20%	205/1026 20%	193/1026 19%	198/1026 19%	181/1026 18%
±10 años	389/1026 38%	389/1026 38%	384/1026 37%	384/1026 37%	379/1026 37%	362/1026 35%	357/1026 35%	351/1026 34%	350/1026 34%	350/1026 34%
±20 años	626/1026 61%	632/1026 62%	624/1026 61%	634/1026 62%	624/1026 61%	628/1026 61%	612/1026 60%	604/1026 59%	599/1026 58%	591/1026 58%
±30 años	764/1026 74%	771/1026 75%	774/1026 75%	759/1026 74%	755/1026 74%	760/1026 74%	762/1026 74%	747/1026 73%	746/1026 73%	752/1026 73%
±40 años	872/1026 85%	868/1026 85%	867/1026 85%	864/1026 84%	855/1026 83%	851/1026 83%	852/1026 83%	852/1026 83%	852/1026 83%	843/1026 82%
±50 años	915/1026 89%	918/1026 89%	917/1026 89%	916/1026 89%	914/1026 89%	910/1026 89%	907/1026 88%	911/1026 89%	904/1026 88%	907/1026 88%
±100 años	1012/1026 99%	1012/1026 99%	1013/1026 99%	1011/1026 99%	1012/1026 99%	1012/1026 99%	1012/1026 99%	1013/1026 99%	1013/1026 99%	1012/1026 99%
Promedio	22	22	22	23	23	23	23	24	24	24
Media cuadrática	32	32	32	32	33	33	33	34	34	34
Mediana	15	15	15	16	16	16	16	16	17	17
Máximo	186	196	196	204	212	208	205	199	207	213
Mínimo	0	0	0	0	0	0	0	0	0	0

Figura 9. Resultado de datación sin distinción tipológica con distinto valor de *k*

Si se toma como punto de referencia el promedio del margen de error absoluto, la mejor capacidad predictiva se obtiene cuando *k*=6, en cuyo caso un 60 % o 642 piezas quedaron datadas con un margen de error de ±20 años respecto a la fecha verdadera y cerca del 80 % con un margen de error de ±30 años[12]. Se aprecia que la predecibilidad no se ve gravemente afectada por el distinto valor de *k*.

Por otra parte, la Figura 10 ilustra gráficamente la correspondencia entre la fecha estimada y la fecha verdadera de los documentos datados. La mayoría se sitúa exitosamente sobre o poco alejada de la línea y = x[13].

[12] Mejorará la predecibilidad con la exclusión de documentos redactados en latín provenientes en su mayoría de antes de la primera mitad del siglo XIII, a los que los parámetros establecidos son poco aplicables.

[13] Aquí no nos ocuparemos de las piezas mal datadas, tarea no obstante importantísima que emprendemos en un futuro estudio. Algunos de los posibles factores desviadores en la datación serán la poca concentración documental en ciertos periodos, la brevedad del texto, la existencia en el corpus de traslados o copias en los que la fecha de elaboración final distara considerablemente de la de la composición original y/o la posible falsificación documental, entre otros.

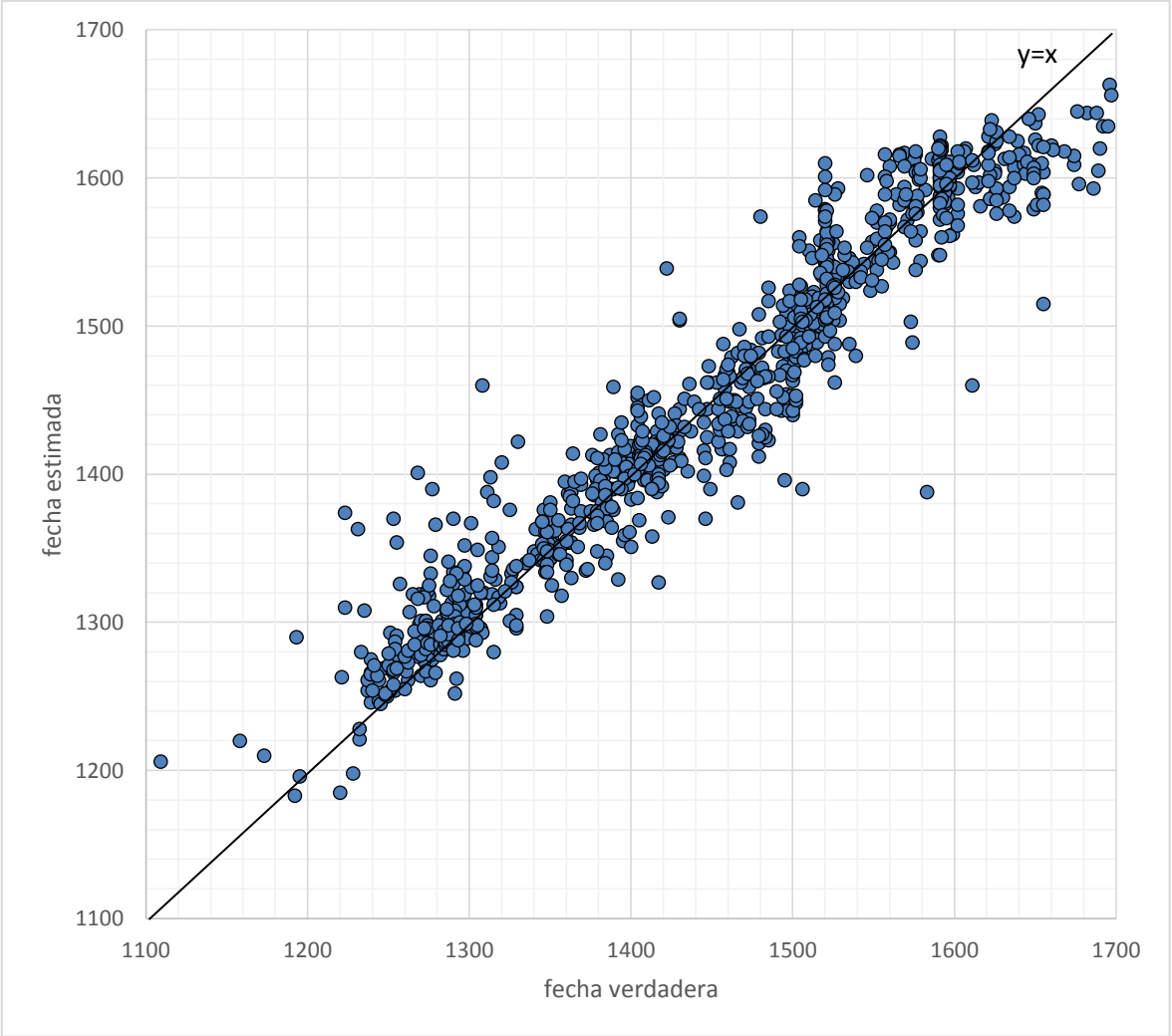

Figura 10. La fecha estimada contra la fecha verdadera (sin distinción tipológica, *k*=6)

Por último, señalaremos la correlación negativa (-0.64) que se da entre el número de piezas y el promedio del margen de error absoluto en el periodo dado (Figura 11).

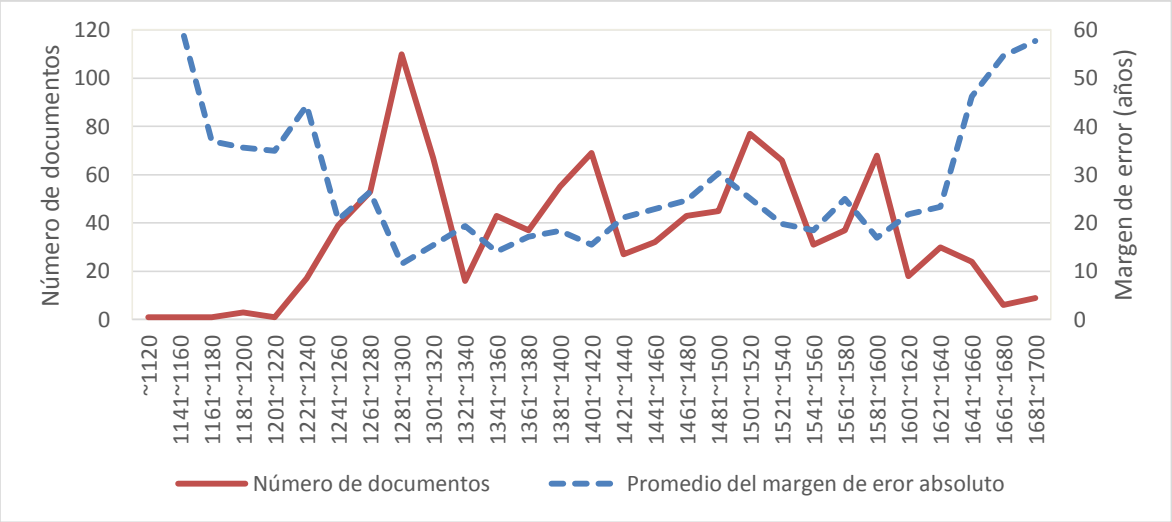

Figura 11. Relación entre el número de piezas en cada periodo y el promedio del margen de error absoluto

4.2.2. Con distinción tipológica

Si bien el control tipológico nos permite tratar la variación tipológica, sufrimos, en cambio, la merma en el número de piezas disponibles que puede afectar la predecibilidad, pues el método *k-NN* presenta un *trade-off* entre la restricción documental y el número de piezas disponibles[14]. A continuación, reproducimos, a modo de comparación, el resultado de la datación efectuada con el control tipológico (Cancilleresco, Eclesiástico, Judicial, Municipal y Particular). Como se puede apreciar en la Figura 12, la capacidad predictiva varía de una agrupación en otra. Los documentos cancillerescos presenta una funcionalidad más elevada con un margen de error medio de 16 años, valor por debajo del que se da sin distinción tipológica (21 años), mientras que baja la predecibilidad en los demás grupos. Recuérdese que el valor más apropiado de *k* está sujeto tanto al número de piezas disponibles *N* como a la distribución crono-geográfica de las mismas en cada tipología documental.

Margen de error	Cancilleresco (k=6, N=263)	Eclesiástico (k=4, N=329)	Judicial (k=4, N=91)	Municipal (k=3, N=74)	Particular (k=3, N=269)	Total (k=6, N=1026)
±5 años	86/263 33%	51/329 16%	14/91 15%	18/74 24%	74/269 28%	260/1026 25%
±10 años	134/263 51%	103/329 31%	21/91 23%	32/74 43%	111/269 41%	419/1026 41%
±20 años	185/263 70%	163/329 50%	35/91 38%	48/74 65%	162/269 60%	642/1026 63%
±30 años	218/263 83%	225/329 68%	53/91 58%	55/74 74%	202/269 75%	777/1026 76%
±40 años	238/263 90%	261/329 79%	64/91 70%	62/74 84%	218/269 81%	871/1026 85%
±50 años	248/263 94%	286/329 87%	73/91 80%	67/74 91%	233/269 87%	932/1026 91%
±100 años	263/263 100%	321/329 98%	88/91 97%	73/74 99%	264/269 98%	1015/1026 99%
Promedio	16	27	32	22	23	21
Media cuadrática	24	37	42	33	41	31
Mediana	10	21	25	14	14	14
Máximo	95	193	160	121	397	195
Mínimo	0	0	0	0	0	0

Figura 12. Resultado de datación con distinción tipológica

[14] Debemos a un revisor anónimo la sugerencia de hacer distinción tipológica.

4.2.2.1. Cancilleresco

Buena parte de los documentos cancillerescos provienen de Burgos, Madrid, Segovia, Sevilla, Toledo y Valladolid o de las dos Castillas históricas (Figura 13). La elevada predecibilidad se deberá a la escasa variación geográfica.

Cancilleresco	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Álava											1																1
Almería																				1							1
Ávila						1											1										2
Burgos			2	2	5	6	4	1		1	1	1		3			8	1			1						36
Cádiz					1																						1
Cantabria																1											1
Córdoba				1		2								1		1											5
Granada															1	2		2		1							6
Guadalajara				1		3	2					1		1	2												10
León	1			1		1																					3
Lugo			2																								2
Madrid						1		1	1		2	1				2	3	2	4	4	4		1	1	3		30
Navarra					2		1										1										4
País Vasco					1	2												1									4
Palencia				1		1																					2
Portugal											1										1						2
Salamanca							2								1		1										4
Segovia			1	1	1	1						1			6		1										12
Sevilla			1	9	1	3		1	3	3				1	1	1	1	2	1								28
Soria									1							1											2
Toledo					1	9		3	2	2			1	1	2		1	3	4								29
Valencia																										1	1
Valladolid				1	1	27	6		3	2	1	7		1	1	6	3	4	5	2							70
Vizcaya																	1										1
Zamora						3	1		1																		5
Zaragoza								1																1			2
Total	1		6	17	13	60	16	7	11	8	6	11	1	8	14	14	21	15	14	8	6		1	2	3	1	264

Figura 13. Distribución espacio-temporal de los documentos cancillerescos (número de piezas)

Scriptum Digital
Vol. 3 (2014), pp. 29-63

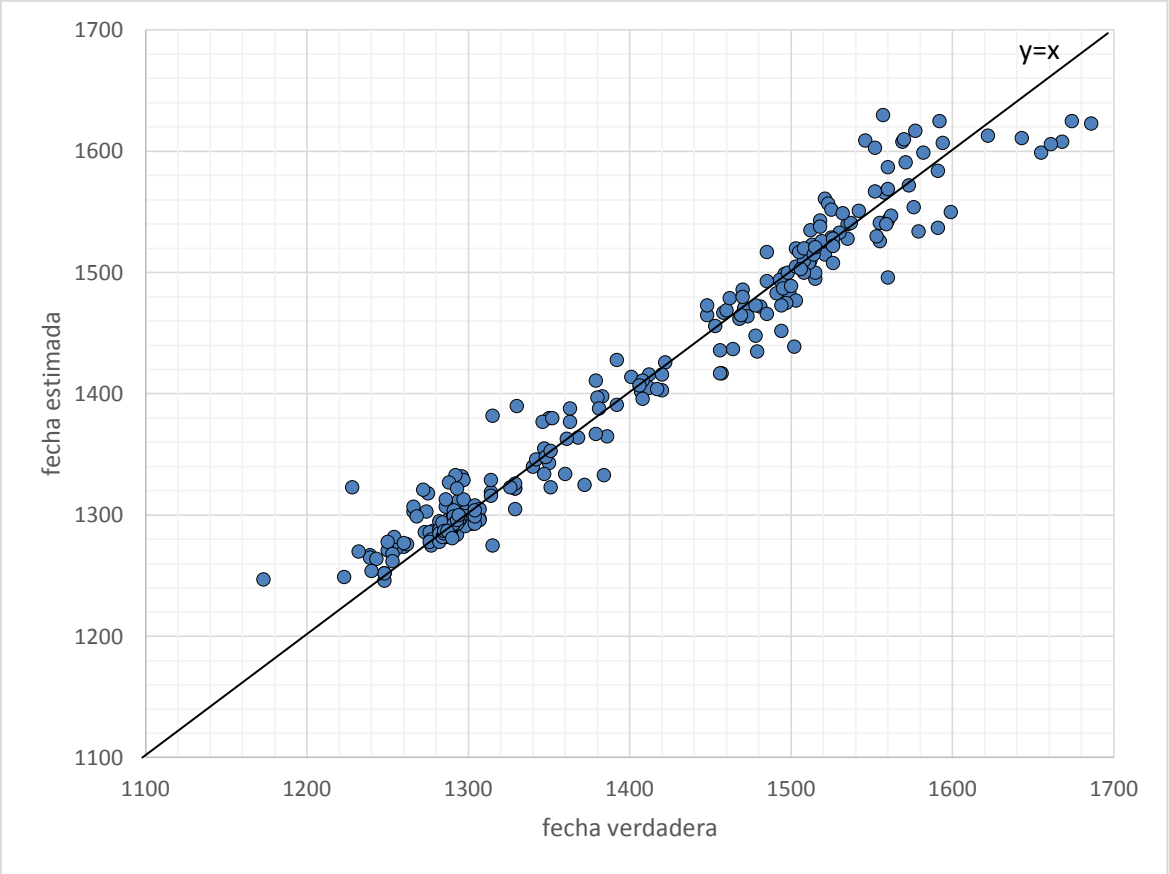

Figura 14. La fecha estimada contra la fecha verdadera (Cancilleresco, *k*=6)

Scriptum Digital
Vol. 3 (2014), pp. 29-63

4.2.2.2. Eclesiástico

Esta agrupación incluye gran cantidad de documentos procedentes de zonas leonesas y aragonesas (Figura 15). El número mayor de piezas (N=329) respecto al de los demás grupos que favorecería la datación parece ser contrarrestado con una variación geográfica grande.

Eclesiástico	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Asturias					1		2			1	1		1	1		1											8
Ávila				1	1	3	1																				6
Burgos	2		1	6	1	3	1	1								2		1									18
Cáceres									1	3	1	1															6
Cantabria		1			1	1	1			1	6	2			3	3	1	1									21
Cuenca			1																			2	1	1			5
Granada																							3	2			5
Guadalajara						1														1						1	3
Huesca	1				2					2	3					2											10
Italia																						1					1
Jaén								1							1	2											4
La Rioja			1	1	2	2	2	1		1		2	1	1			1										15
León			1	1	4		1				8	1	1	1	3	1	6					2	2	1	1		34
Madrid											1						2			1	3	2	3	3			15
Málaga																			1				1				2
Navarra					6	4	3	2	2					1			2										20
País Vasco						1																					1
Palencia				1	2	1						3		1	1		1					1					11
Salamanca			2	4				1	5	3	1	3	3		5	1	2				2			4			36
Segovia					1	3												1									5
Sevilla					1					1								2									4
Soria					1																						1
Teruel										2	4	6	2	2	2			1									19
Toledo	1				2						1		1		1	1			1								8
Valladolid						1				3	1	1				4	1					5	3				19
Zamora				1							1		4	2				1	1								10
Zaragoza			2	2	1	1	2	1	9	3	3	12	1	3			1								1		42
Total	4	1	8	17	26	21	13	7	17	20	31	31	14	12	16	17	17	7	3	2	5	13	13	11	2	1	329

Figura 15. Distribución espacio-temporal de los documentos eclesiásticos (número de piezas)

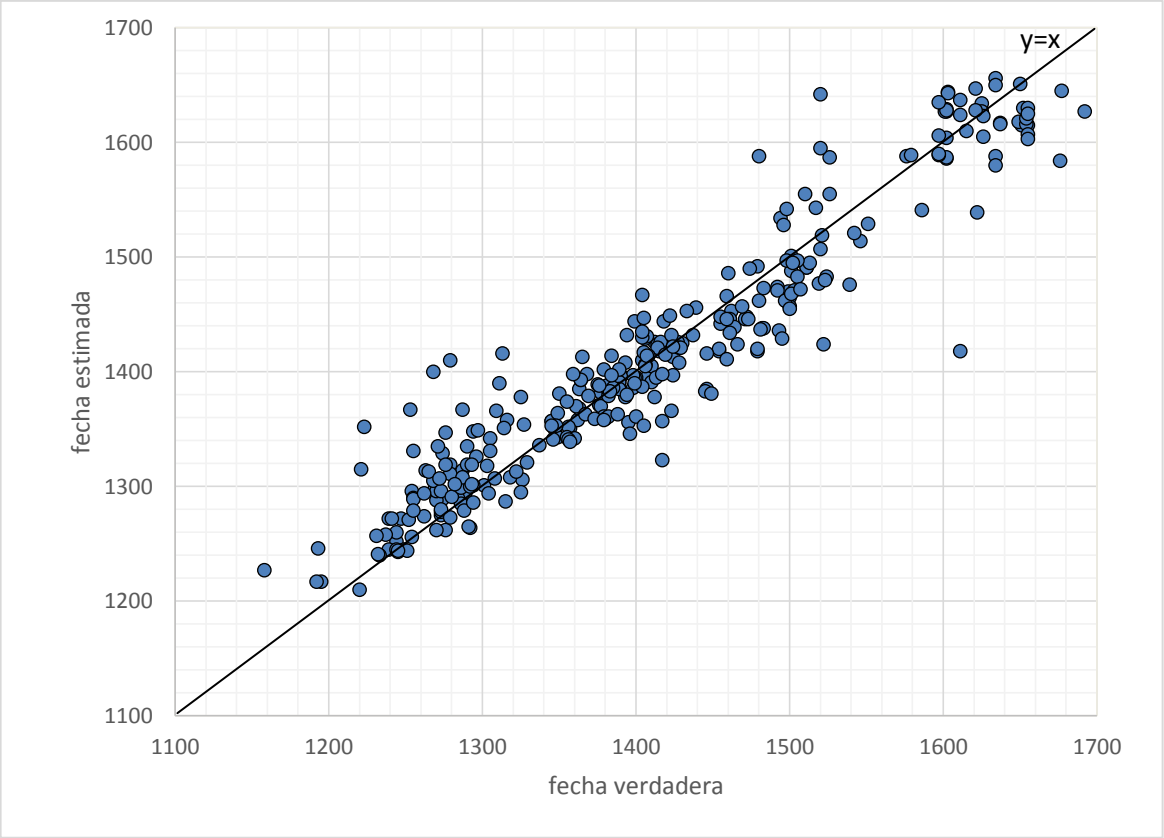

Figura 16. La fecha estimada contra la fecha verdadera (Eclesiástico, *k*=4)

4.2.2.3. Judicial

Como es de suponer, la escasez de documentos obstaculiza la datación dando como resultado un margen de error medio elevado de 32 años (Figura 17).

Judicial	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Asturias							1														1						2
Ávila																	1										1
Burgos			1															2									3
Cáceres									1									1									2
Guadalajara																1	3	2	1	4	2	4	4	5	1	4	31
Italia																					7						7
La Rioja						1																					1
León					1									1													2
Madrid											1					1		1		1	4					1	9
Murcia																					2						2
Navarra						1	1		1																		3
Palencia																	1										1
Salamanca														1			1										2
Segovia																	1										1
Sevilla							1												4								5
Teruel									1					1													2
Toledo						1									1			1					1				4
Valladolid																	3	5		1							9
Vizcaya																	1										1
Zamora						1														1							2
Zaragoza																		1									1
Total			1		1	4	3		3		1			3	1	2	11	13	5	7	16	4	5	5	1	5	91

Figura 17. Distribución espacio-temporal de los documentos judiciales (número de piezas)

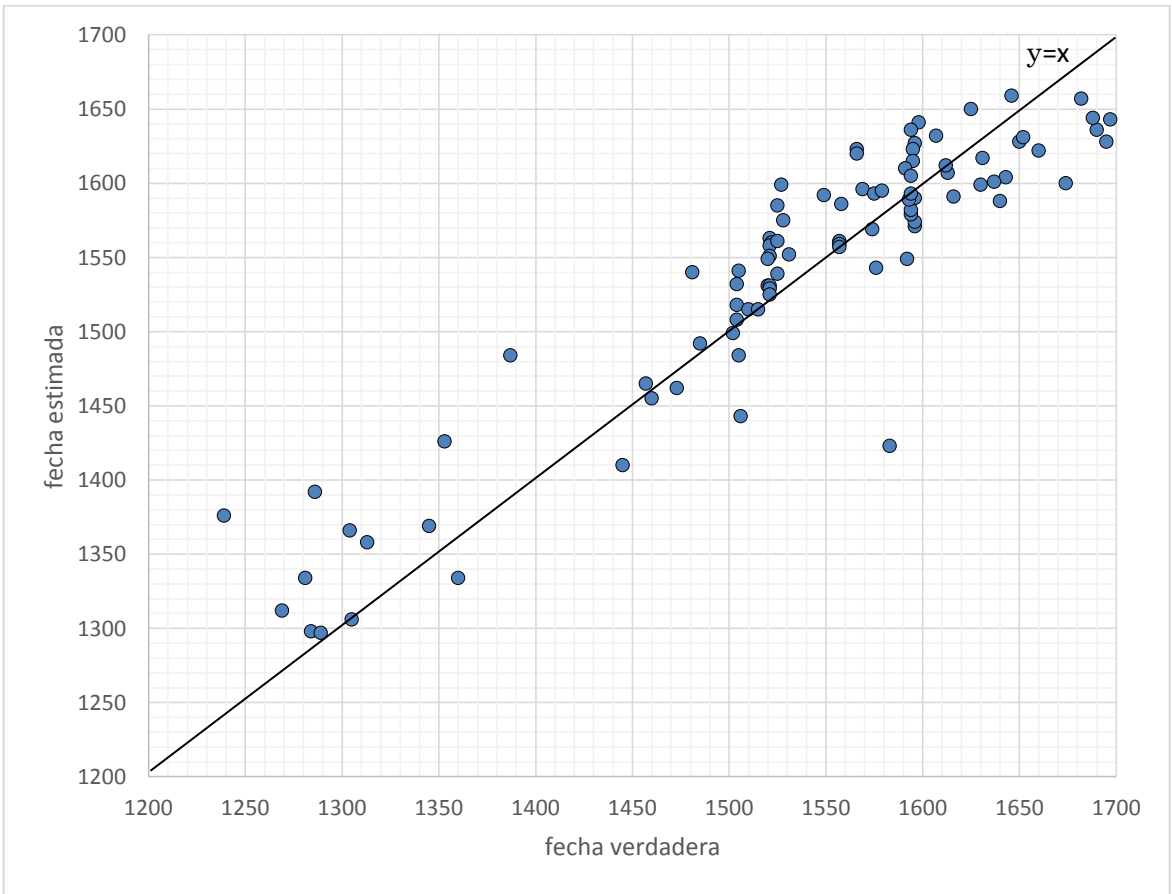

Figura 18. La fecha estimada contra la fecha verdadera (Judicial, *k*=4)

4.2.2.4. Municipal

La relativamente alta predecibilidad pese a la escasez de textos se deberá a la poca variación geográfica que presentan los documentos emitidos generalmente en los territorios ocupados por el reino de Castilla (Figura 18).

Municipal	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Ávila						1																					1
Badajoz																			1								1
Cádiz																	1										1
Córdoba																	1		1								2
Granada																		2					1				3
Guadalajara											1		1		2								1				5
Guipúzcoa												1								2							3
Jaén																							4				4
La Rioja					1																						1
León														1			1										2
Madrid															1	1	2	3	2	3	1						13
Murcia																		1									1
Navarra											1																1
Segovia																		1									1
Sevilla					1												1	4	2		2						10
Soria																				1							1
Toledo									1				3	1	2	1		3									11
Valladolid												1					2	2	1	2							8
Vizcaya																	1										1
Zaragoza				1					1			2															4
Total				1	2	1			2		2	4	4	2	5	2	9	16	7	8	3		6				74

Figura 19. Distribución espacio-temporal de los documentos municipales (número de piezas)

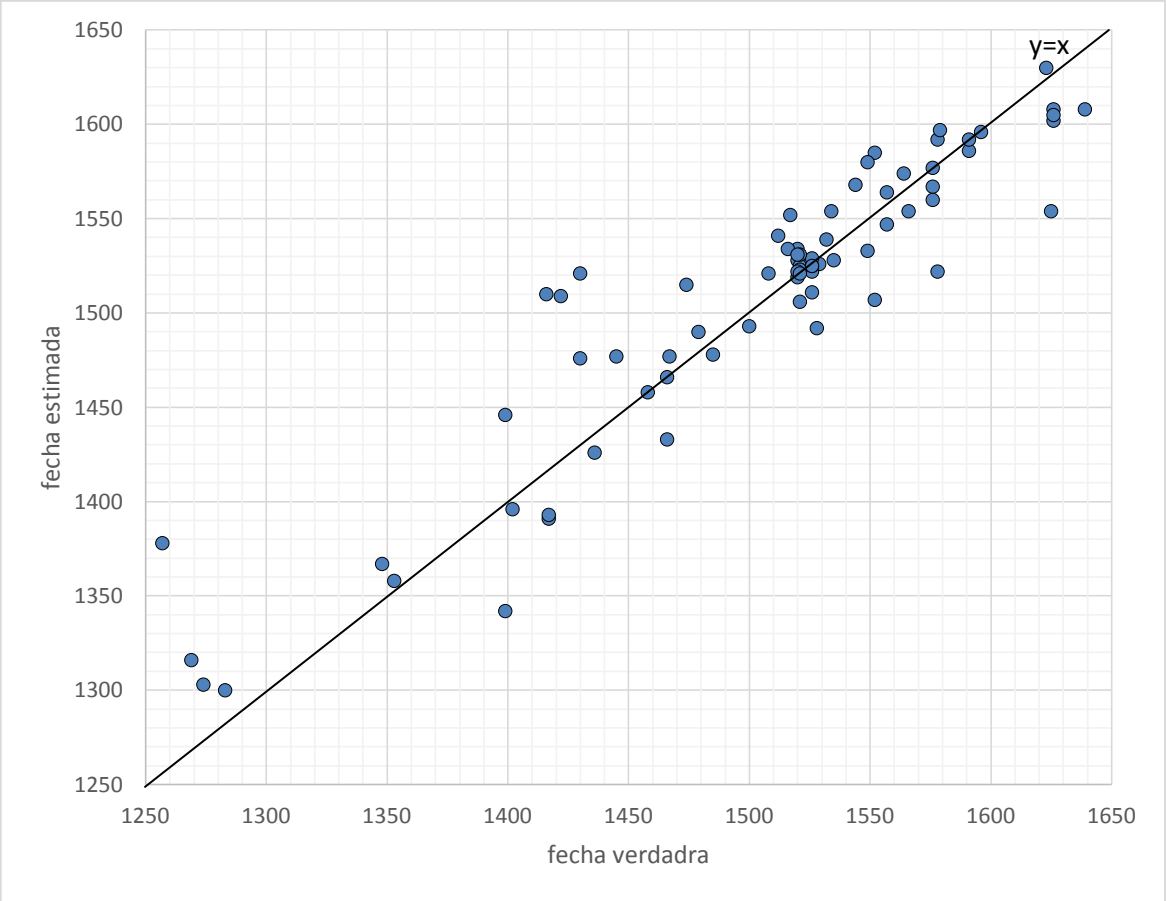

Figura 20. La fecha estimada contra la fecha verdadera (Municipal, *k*=3)

4.2.2.5. Particular

Pese a la gran variación geográfica, la predecibilidad no se ve gravemente afectada (Figura 21).

Particular	~1200	1201~1220	1221~1240	1241~1260	1261~1280	1281~1300	1301~1320	1321~1340	1341~1360	1361~1380	1381~1400	1401~1420	1421~1440	1441~1460	1461~1480	1481~1500	1501~1520	1521~1540	1541~1560	1561~1580	1581~1600	1601~1620	1621~1640	1641~1660	1661~1680	1681~1700	Total
Albacete																				3							3
Asturias					1																						1
Ávila						19	22				1																42
Badajoz									1												1						2
Burgos					1					1						1		1									4
Cáceres						1		1	3	1	1		1	1							1						10
Cádiz							1					2			1	1	1				7						13
Cantabria						1			1	1							4										7
Córdoba																	1										1
Cuenca																								3			3
Granada																				3							3
Guadalajara																1				4		1	3	1		1	11
Guipúzcoa																	1										1
Huelva																		1			2						3
Huesca					1	1	1				7		1	1													12
Islas Baleares																					1						1
Italia																					6						6
Jaén																1		1		1							3
La Rioja					1				1			1		1		1	1	2									8
León			1	2	1							1					3										8
Madrid												1									7		1			1	10
Málaga																			1		2						3
Murcia							1																				1
Navarra					3							2				1	1	1									8
País Vasco																		1									1
Palencia	1													1	2												4
Portugal																					1						1
Salamanca				1			8					1	1		3								1	1			16
Segovia					1	2						1					2										6
Sevilla																					10			1			11
Teruel							1		1	5	1	5	2	1	1			1									18
Toledo								1					1	2				5	1								10
Valladolid				1	1		1								1		2										6
Vizcaya																	1										1
Zamora			1						1	1	2	2	1														8
Zaragoza									2	1	3	7	1			4	2	2		1							23
Total	1		2	4	10	24	35	2	10	10	15	23	8	7	8	10	19	15	2	12	38	1	5	6		2	269

Figura 21. Distribución espacio-temporal de los documentos particulares (número de piezas)

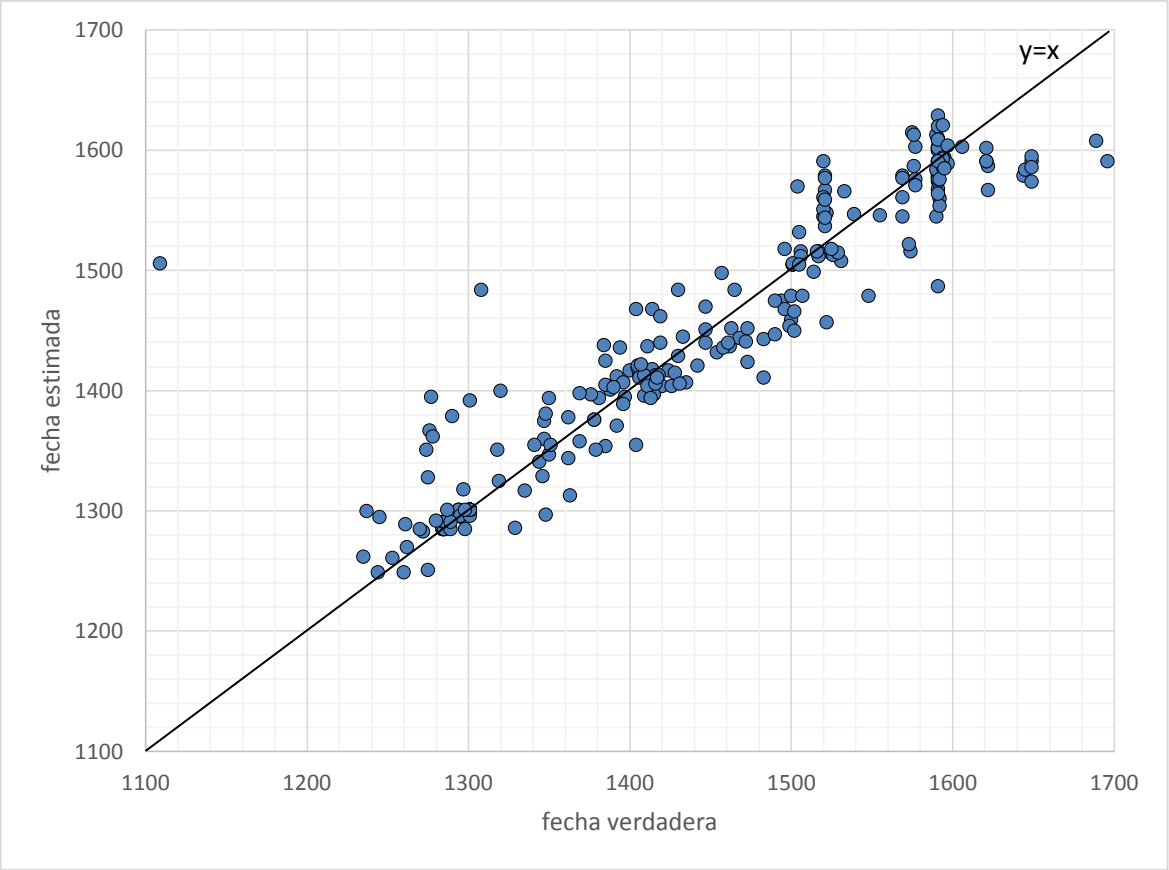

Figura 22. La fecha estimada contra la fecha verdadera (Particular, *k*=3)

4.2.3. Datación de documentos sin *data chronica*

Terminaremos este apartado datando unos documentos sin fecha explícita que, pese a ello, son datables a través del emisor constatado, dado que se trata de un personaje histórico. De las dos fechas estimadas, la primera ha sido computada sin distinción tipológica y la segunda entre corchetes con ella, que pueden ser consideradas acertadas (Figura 23).

ID	Provincia	Tipología	Emisor	Reinado	Fecha estimada
CODEA373	Sin lugar	Cancilleresco	Carlos I	1516-1556	*1565 (*1555)
CODEA854	La Rioja	Cancilleresco	Alfonso X	1252-1284	*1288 (*1275)
CODEA1439	Sin lugar	Cancilleresco	Enrique IV	1454-1474	*1442 (*1419)

Figura 23. Datación de documentos sin *data chronica*

5. ADSCRIPCIÓN GEOGRÁFICA

5.1. Procedimiento

Ahora bien, la matriz del coeficiente de correlación (Figura 6) también permite efectuar la adscripción geográfica de igual manera que la datación cronológica, de modo que el lugar de emisión estimado sea la localidad que posee la suma total del valor del coeficiente

mayor entre los *k* documentos más parecidos al documento en cuestión.

ID	Año	Provincia	Región	Tipología	CODEA1501
CODEA1501	1515	Burgos	CV	C	1.000
CODEA292	1533	Toledo	CN	P	0.573
CODEA611	1517	La Rioja	CV	E	0.567
CODEA1278	1508	Burgos	CV	C	0.523
CODEA1385	1485	Valladolid	CV	C	0.502
CODEA25	1513	Valladolid	CV	C	0.499
CODEA1426	1521	Navarra	NA	P	0.485

Figura 24. Los seis documentos más afines al CODEA1501

Volvamos a tomar como ejemplo el CODEA1501. El lugar de emisión de estos seis documentos es Toledo (CODEA292), La Rioja (CODEA611), Burgos (CODEA1278), Valladolid (CODEA1385 y CODEA25) y Navarra (CODEA1426), de ahí que la procedencia geográfica estimada del CODEA1501, al tomarse 'provincia' como término de referencia, sea *Valladolid con la suma todal del valor del coeficiente mayor 1.001 (=0.502+0.499), discrepando con la *data tópica* que es Burgos. Por otra parte, a nivel de 'región', la procedencia geográfica estimada es *CV (Castilla la Vieja) que posee un valor de 2.091 (=0.567+0.523+0.502+0.499) superior al de CN (Castilla la Nueva) 0.573 y de NA (Navarra) 0.485, coincidiendo con la *data topica*.

5.2. Resultado

5.2.1. Sin distinción tipológica

El resultado de la adscripción geográfica realizada con los 1026 documentos (*k*=6) a nivel de 'provincia' y 'región' está resumido en las Figuras Figura 25 y Figura 26 respectivamente, en las que tanto la fila como la columna van ordenadas en orden alfabético. En la primera fila aparecen los lugares de emisión documental, y en la primera columna los estimados. Se aprecia una concentración en el eje diagonal que representa los casos correctamente estimados. Utilizamos el mismo valor de *k* (=6) que en la datación cronológica[15], en cuyo caso la proporción de correspondencia a nivel de 'provincia' resulta ser del 48 % (491/1026), y a nivel de 'región' del 63 % (651/1026).

[15] Fijamos *k* al que proporciona mayor rentabilidad en la datación, a cuya precisión damos mayor importancia. Ni que decir tiene que la adscripción geográfica se ve afectada por el valor de *k* al igual que la datación cronológica.

	Álava	Albacete	Almería	Asturias	Ávila	Badajoz	Burgos	Cáceres	Cádiz	Cantabria	Córdoba	Cuenca	Granada	Guadalajara	Guipúzcoa	Huelva	Huesca	Islas Baleares	Italia	Jaén	La Rioja	León	Lugo	Madrid	Málaga	Murcia	Navarra	País Vasco	Palencia	Portugal	Salamanca	Segovia	Sevilla	Soria	Teruel	Toledo	Valencia	Valladolid	Vizcaya	Zamora	Zaragoza	Total
*Albacete																			1					1															1				3
*Asturias				7						1												1																				9
*Ávila					44		1	1	1	1											3						2		1			6	1	1		2		1				65
*Badajoz																				1																						1
*Burgos	1						20			3			1	1							2	3	1	1		1		1	2		4	2	4			7		11		1		66
*Cáceres					1	1	1	10	1													1		1							1		1					1				19
*Cádiz									3			1									1			1								1										7
*Cantabria										12																		2				1					3	1			19
*Córdoba																																				1		1				2
*Cuenca														1		1								1														1				4
*Granada							1						3	1					2		2	1		2							1		1			1	1		1		17
*Guadalajara							2	1	1			2	1	31	2				1	3		1		5	2				1		2	1	3					5				64
*Huelva													1																													1
*Huesca																	18										1								1						1	21
*Italia									1										5			1		6			1					1				1						16
*Jaén									1											1				2														1				5
*La Rioja							1														6		1									1										9
*León				1							1											20									1									2		25
*Lugo							1																																		1
*Madrid		2	1		1		1	1	2				4	4	1				4	1		1		37		1	2	1		2	4		5	2		4		10			1	92
*Málaga																											1															1
*Murcia						1																																1		1		3
*Navarra																	1										27															1	29
*País Vasco																																						1				1
*Palencia																						1		2					2									2				7
*Salamanca				1	1		2	5		10			1	7	1			1		2	3	7		1		1			6		40		1			4		2	1	5	1	103
*Segovia					1		2						2	1										1								5	1			7		1				21
*Sevilla							6		2		1	2	2	2					1	1	1	1		4	2	1	1	1	1			1	25	1		3		7		3		69
*Teruel																																		28							3	31
*Toledo		1					10		3		1		1	5			1			2	2	4		3	1				1			4	3			19		6				67
*Valladolid					4	1	13			1	5	2	1	6		2					4	2		9				3	2	1	5	5	10			11		56	1	4	2	150
*Vizcaya										1																																1
*Zamora				2																		5															2		1		9		19
*Zaragoza												1					2				1						1									10					63	78
Total	1	3	1	11	52	3	61	18	15	29	8	8	17	59	4	3	22	1	14	11	25	49	2	77	5	4	36	6	18	3	58	25	58	4	39	62	1	112	4	25	72	1026

Figura 25. Correspondencia a nivel de provincia (sin distinción tipológica, *k*=6)

	AN	AR	AS	CA	CN	CV	EX	GA	IT	LE	MU	NA	PT	PV	VA	Total
*AN	39			1	18	20			3	7	2	1		2	1	94
*AR		125			1							3				129
*AS			3													3
*CN	35	3			119	45	3		6	15	1	3	2	4		236
*CV	37	4	3		57	236	3	2		21		2	1	8		374
*EX	1				1	2	10				1					15
*IT	1				3				5	1		1				11
*LE	5		5		10	23	5			88				1		137
*NA		1										26				27
Total	118	133	11	1	209	326	21	2	14	132	4	36	3	15	1	1026

Figura 26. Correspondencia a nivel de región (sin distinción tipológica, *k*=6)

Por otra parte, la proporción de correspondencia por periodos aparece resumida en la Figura 27. A diferencia de en la datación, no se observa una fuerte correlación entre la cantidad de documentos y la precisión en el periodo dado, con el coeficiente de correlación *Pearson* de 0.22 y 0.25 a nivel de provincia y a nivel de región, respectivamente.

Periodo	Piezas	Correspondencia			
		Provincia		Región	
1100~1200	6	1	17%	3	50%
1201~1220	1	0	0%	1	100%
1221~1240	17	7	41%	10	59%
1241~1260	39	25	64%	28	72%
1261~1280	53	29	55%	38	72%
1281~1300	110	54	49%	82	75%
1301~1320	67	49	73%	55	82%
1321~1340	16	8	50%	10	63%
1341~1360	43	29	67%	32	74%
1361~1380	37	23	62%	30	81%
1381~1400	55	37	67%	43	78%
1401~1420	69	46	67%	55	80%
1421~1440	27	17	63%	19	70%
1441~1460	32	13	41%	20	63%
1461~1480	43	20	47%	28	65%
1481~1500	45	11	24%	21	47%
1501~1520	77	24	31%	36	47%
1521~1540	66	16	24%	29	44%
1541~1560	31	15	48%	19	61%
1561~1580	37	14	38%	22	59%
1581~1600	68	21	31%	28	41%
1601~1620	18	5	28%	6	33%
1621~1640	30	9	30%	12	40%
1641~1660	24	10	42%	13	54%
1661~1680	6	3	50%	3	50%
1681~1700	9	5	56%	8	89%
Total	1026	491	48%	651	63%

Figura 27. Precisión de adscripción geográfica por periodos

A nivel de 'provincia' la proporción de correspondencia es del 48 %, y a nivel de 'región' del 63 %. Pero ¿cómo evaluaremos este resultado? La respuesta depende del criterio en que uno se apoye respecto a la correspondencia entre la localidad y sus rasgos lingüísticos. A diferencia de la datación cronológica cuya funcionalidad se puede medir cuantitativamente, la adscripción geográfica, siendo cualitativa, es difícil de evaluar. ¿Daríamos por fallida la datación diatópica cuando, por ejemplo, el documento emitido en Andalucía durante la Edad Media fuese atribuido a la Castilla Vieja o a la Castilla Nueva o viceversa? y ¿es más verosímil suponer que la modalidad lingüística cancilleresca del siglo XIII de Sevilla fuera diferente a la de Burgos que lo contrario? El caso extremo lo constituyen los documentos emanados de Italia o Portugal incluidos en el corpus. ¿Deberían por fuerza ser adscritos a Italia y Portugal, respectivamente? Como por el momento no podemos dar una respuesta definitiva a esta cuestión, intentaremos

establecer el modo de evaluación apropiado en un futuro estudio.

5.2.2. Con distinción tipológica

A continuación ofrecemos el resultado obtenido de la adscripción geográfica realizada con el control tipológico (Figura 28). Para cada categoría utilizamos el mismo valor de *k* que en la datación cronológica.

	Adscripción geográfica		Datación cronológica		
	Correspondencia		Margen de error		
	Provincia	Región	Promedio	Desviación estándar	Mediana
Cancilleresco (k=6, N=263)	38%	55%	16	17	10
Eclesiástico (k=4, N=329)	52%	67%	27	26	21
Judicial (k=4, N=91)	47%	60%	32	28	25
Municipal (k=3, N=74)	35%	50%	22	24	14
Particular (k=3, N=269)	50%	65%	23	34	14
Total (k=6, N=1026)	48%	63%	21	23	14

Figura 28. Resultado de adscripción geográfica y datación cronológica

5.2.2.1. Cancilleresco

La proporción de correspondencia a nivel de 'provincia' es del 38 % (101/263), y a nivel de 'región' del 55 % (145/263).

	Álava	Almería	Ávila	Burgos	Cádiz	Cantabria	Córdoba	Granada	Guadalajara	León	Lugo	Madrid	Navarra	País Vasco	Palencia	Portugal	Salamanca	Segovia	Sevilla	Soria	Toledo	Valencia	Valladolid	Vizcaya	Zamora	Zaragoza	Total
*Burgos	1			11				2	1			1	1	1	1		1	1	2		1		7		1		32
*Cantabria																							1				1
*Córdoba																							1				1
*Granada																								1			1
*Lugo										1	1																2
*Madrid		1		2				1			1	17	1	1		1			3	1	3	1	4			1	38
*Navarra													2														2
*País Vasco												1											1				2
*Salamanca												1											1				2
*Segovia				2				2										4	1		3		1				13
*Sevilla				3			1					1			1			3	12	1	4		4				30
*Toledo				3	1				3	1		2					1	1	1		10		5				28
*Valladolid			2	15			4	1	5	1		7		2		1	2	3	9		8		44		4	1	109
*Vizcaya						1																					1
*Zamora																							1				1
Total	1	1	2	36	1	1	5	6	9	3	2	30	4	4	2	2	4	12	28	2	29	1	70	1	5	2	263

Figura 29. Correspondencia a nivel de provincia (Cancilleresco, *k*=6)

	AN	AR	CN	CV	GA	LE	NA	PT	PV	VA	Total
*AN	10			12							22
*CN	6	1	39	18	1	3	1	1	1	1	72
*CV	25	1	29	94	1	8	1	1	5		165
*GA						1					1
*LE				1							1
*NA							2				2
Total	41	2	68	125	2	12	4	2	6	1	263

Figura 30. Correspondencia a nivel de región (Cancilleresco, *k*=6)

5.2.2.2. Eclesiástico

La proporción de correspondencia a nivel de 'provincia' es del 52 % (170/329), y a nivel de 'región' del 67 % (221/263).

	Asturias	Ávila	Burgos	Cáceres	Cantabria	Cuenca	Granada	Guadalajara	Huesca	Italia	Jaén	La Rioja	León	Madrid	Málaga	Navarra	País Vasco	Palencia	Salamanca	Segovia	Sevilla	Soria	Teruel	Toledo	Valladolid	Zamora	Zaragoza	Total
*Asturias	2				1								1															4
*Ávila		2										1						1		1				1				6
*Burgos		1	11		2						1	2	1				1	2		1							1	23
*Cáceres		2	1	3																								6
*Cantabria	2				9								1					1							1			14
*Cuenca										1			2	1		1									2			7
*Granada														1		1												2
*Guadalajara														1							1							2
*Huesca									7														1				3	11
*Italia																		1	1									2
*Jaén											1																	1
*La Rioja			2		1	1			1			5	1			1								1				13
*León	2						1						17													1		21
*Madrid						1	2	1						6	1				3		1				3			18
*Málaga						1		1																				2
*Navarra							1									15												16
*Palencia																		1							1			2
*Salamanca		1	2	3	8		1	1			2	4	2	3	1			5	30	2		1		4	3	2		75
*Segovia			1																									1
*Sevilla														1														1
*Soria																				1								1
*Teruel																1							12				1	14
*Toledo			1									1	1	1		1			1						2			8
*Valladolid	1					1						1	2	1					1		1			1	7	1	1	18
*Zamora	1												6								1			1		6		15
*Zaragoza						1			2			1											6				36	46
Total	8	6	18	6	21	5	5	3	10	1	4	15	34	15	2	20	1	11	36	5	4	1	19	8	19	10	42	329

Figura 31. Correspondencia a nivel de provincia (Eclesiástico, *k*=4)

	AN	AR	AS	CN	CV	EX	IT	LE	NA	PV	Total
*AN	1			3					1		5
*AR		65		1	1				1		68
*CN	6	1		11	6		1	9			34
*CV	2	5	4	6	64			9	3	1	94
*EX					1	3					4
*IT					1						1
*LE	5		4	10	23	3		62			107
*NA	1								15		16
Total	15	71	8	31	96	6	1	80	20	1	329

Figura 32. Correspondencia a nivel de región (Eclesiástico, *k*=4)

5.2.2.3. Judicial

La proporción de correspondencia a nivel de 'provincia' es del 47 % (43/91), y a nivel de 'región' del 60 % (55/91).

	Asturias	Ávila	Burgos	Cáceres	Guadalajara	Italia	La Rioja	León	Madrid	Murcia	Navarra	Palencia	Salamanca	Segovia	Sevilla	Teruel	Toledo	Valladolid	Vizcaya	Zamora	Zaragoza	Total
*Asturias	2							1														3
*Burgos																	1					1
*Cáceres			1												1							2
*Guadalajara		1		1	28	1		1	5	2		1		1			2	4	1			48
*Italia						2			1													3
*León									1				1							1		3
*Madrid						3							1					1				5
*Murcia			1																			1
*Navarra											3					2						5
*Palencia					1																	1
*Segovia			1																			1
*Sevilla				1		1									4		1			1		8
*Toledo									1													1
*Valladolid					2				1									4			1	8
*Vizcaya							1															1
Total	2	1	3	2	31	7	1	2	9	2	3	1	2	1	5	2	4	9	1	2	1	91

Figura 33. Correspondencia a nivel de provincia (Judicial, *k*=4)

	AN	AR	AS	CN	CV	EX	IT	LE	MU	NA	PV	Total
*AN	4			1			1	1				7
*AS			1					1				2
*CN	1		1	38	9	1	4	2	2		1	59
*CV		1		4	5							10
*IT				1			2					3
*LE					1	1		2				4
*NA		2								3		5
*PV					1							1
Total	5	3	2	44	16	2	7	6	2	3	1	91

Figura 34. Correspondencia a nivel de región (Judicial, *k*=4)

5.2.2.4. Municipal

La proporción de correspondencia a nivel de 'provincia' es del 35 % (26/74), y a nivel de 'región' del 50 % (37/74).

	Ávila	Badajoz	Cádiz	Córdoba	Granada	Guadalajara	Guipúzcoa	Jaén	La Rioja	León	Madrid	Murcia	Navarra	Segovia	Sevilla	Soria	Toledo	Valladolid	Vizcaya	Zaragoza	Total
*Badajoz																	1				1
*Córdoba			1												2		1				4
*Granada						1		1			3				1			1			7
*Guadalajara						3	1														4
*Jaén								3			1										4
*Madrid					2						4					1		1			8
*Segovia															1						1
*Sevilla				1	1				1	1	2	1			3		3	1			14
*Toledo	1	1				1				1	2				2		6				14
*Valladolid				1			2				1			1	1			3	1		10
*Vizcaya																		2			2
*Zaragoza													1							4	5
Total	1	1	1	2	3	5	3	4	1	2	13	1	1	1	10	1	11	8	1	4	74

Figura 35. Correspondencia a nivel de provincia (Municipal, *k*=3)

	AN	AR	CN	CV	EX	LE	MU	NA	PV	Total
*AN	13		11	5			1			30
*AR		4						1		5
*CN	3		16	2	1	2			2	26
*CV	4		1	4					2	11
*EX			1							1
*PV				1						1
Total	20	4	29	12	1	2	1	1	4	74

Figura 36 Correspondencia a nivel de región (Municipal, *k*=3)

5.2.2.5. Particular

La proporción de correspondencia a nivel de 'provincia' es del 50 % (134/269), y a nivel de 'región' del 65 % (175/269).

	Albacete	Asturias	Ávila	Badajoz	Burgos	Cáceres	Cádiz	Cantabria	Córdoba	Cuenca	Granada	Guadalajara	Guipúzcoa	Huelva	Huesca	Islas Baleares	Italia	Jaén	La Rioja	León	Madrid	Málaga	Murcia	Navarra	País Vasco	Palencia	Portugal	Salamanca	Segovia	Sevilla	Teruel	Toledo	Valladolid	Vizcaya	Zamora	Zaragoza	Total
*Albacete						1						1		1																							3
*Ávila			41		1	2	1	2											1									3			1	3		1		56
*Burgos																																1					1
*Cáceres	1			1	1	6														1							1						1		2		14
*Cádiz							4			1						1			2	1	1	1						1	1		1	1					15
*Cantabria								4																										1			5
*Cuenca												1																									1
*Granada												1					2				2											1					6
*Guadalajara	2						1			1		4																		2		1	1			1	13
*Huelva																																1					1
*Huesca															10									1							2					3	16
*Italia				1							2						3				1								1								8
*Jaén																		1			1																2
*La Rioja							1											1						1		1											4
*León		1							1											6								1									9
*Madrid							1										1	1			3		1						2								9
*Málaga							1																														1
*Navarra													1		1				1					1													4
*País Vasco																			1																		1
*Salamanca						1		1											1							1		11				1			2		18
*Segovia					1																											2					3
*Sevilla							2			1		3		1							2	1		1	1			1		5							18
*Teruel															1															12							13
*Toledo							2				1	1		1					1			1				1		1	2			1					12
*Zamora			1		1														1							2		1							3		9
*Zaragoza																								4							4					19	27
Total	3	1	42	2	4	10	13	7	1	3	3	11	1	3	12	1	6	3	8	8	10	3	1	8	1	4	1	16	6	11	18	10	6	1	8	23	269

Figura 37. Correspondencia a nivel de provincia (Particular, *k*=3)

	AN	AR	AS	CA	CN	CV	EX	IT	LE	MU	NA	PT	PV	Total
*AN	15			1	15	4		2	2		1		1	41
*AR		52									5			57
*CN	15	1			15	5	1	1		1				39
*CV	3				4	58	2		1		1	1	1	71
*EX					1	3	6		4					14
*IT	3				1		1	3						8
*LE	1		1		1	5	2		25					35
*NA						1					1		1	3
*PV						1								1
Total	37	53	1	1	37	77	12	6	32	1	8	1	3	269

Figura 38. Correspondencia a nivel de región (Particular, *k*=3)

6. FUTURAS INVESTIGACIONES

Creemos haber demostrado el alcance del método *k-NN* en la datación crono-geográfica de documentos realizada en función de rasgos lingüísticos. Con miras a elevar aun más la predecibilidad tenemos previsto introducir algunas novedades en lo que atañe a la

selección de parámetros, operación matemática y evaluación de la fecha estimada, entre otras.

Repecto al primer punto, estamos trabajando en la detección de parámetros de índole extralingüística. Como lo hemos expuesto arriba, los parámetros utilizados en el presente estudio son todos de naturaleza puramente lingüística. Sin embargo, estas no son las únicas pistas a nuestra disposición encerradas en un documento. Nos referimos a datos extralingüísticos como nombres propios de personajes históricos, testigos, escribanos y otras personas o instituciones que aparecen mencionadas en el texto. Asimismo, es de interés aprovechar la variación paleográfica del documento (Torrens 1995), la codicológica, la sigilográfica, etcétera, que, una vez convertidas en datos cualitativos, pueden servir de parámetros. Ni que decir tiene que procuraremos establecer más rasgos lingüísticos relevantes.

En cuanto al tratamiento matemático, una posibilidad es multiplicar el peso de algunos parámetros que respecto a la discriminación de la diacronía posean rendimiento funcional mayor, a los cuales se les asignará más peso a través del cálculo de la matriz ponderado a partir del patrón de la distribución de los parámetros.

Por último, la evaluación de la fecha estimada consiste en cuantificar el grado de precisión o fiabilidad que le damos. Puesto que algunos textos resultan más difíciles de fechar a raíz de su poca extensión, un reducido número de parámetros presentes, poca concentración de documentos en ciertas épocas, la intervención de más de un escribano y/o la naturaleza de traslados o copias tardíos, entre otros motivos, mientras que otros pueden ser datados con relativamente mayor seguridad, nos vemos en la necesidad de indicar el grado de fiabilidad, lo mismo que la estadística inferencial hace acompañar el valor *p* junto al coeficiente computado de las variables[16]. Además, a fin de evaluar correctamente el margen de error, estamos tratando de medir el real, ya que en más de un caso el margen de error grande no es atribuible a la datación desacertada sino a la escasísima o nula concentración de documentos en ciertas épocas. Así cuando un documento provenga de una fecha muy alejada temporalmente de los demás, la datación está encaminada al fracaso rotundo. Recuérdese que el método *k-NN* se ve propiciado por la abundancia de documentos repartidos tiempo-espacialmente sin laguna, a cuyo respecto buena noticia es el aumento de los mismos que brindará el proyecto CODEA+ 2015 en marcha.

REFERENCIAS BIBLIOGRÁFICAS

ALVAR, Manuel (1996): *Manual de dialectología hispánica: el español de España*. Barcelona: Editorial Planeta.

AZOFRA, María Elena (2009): *Morfosintaxis histórica del español: de la teoría a la práctica*. Madrid: Universidad Nacional de Educación a Distancia.

CHARTA = SÁNCHEZ-PRIETO BORJA, Pedro (coord.) (2010-): *Corpus Hispánico y Americano en la Red:*

[16] Debemos al Prof. Hiroto Ueda de la Universidad de Tokio la valiosa sugerencia a este respecto. De hecho hemos comprobado cierta correlación negativa entre el valor promedio del coeficiente de correlación que frente al documento a datar presentan los *k* documentos y el margen de error absoluto respecto a la fecha verdadera, o sea que cuando se hallen textos bastante parecidos con el coeficiente de correlación elevado, digamos 0.8, la fecha estimada resulta ser mucha más fidedigna.

Textos Antiguos. http://www.bibliateca.es/charta/index.html [Consulta: 17/04/2014].

CODEA = SÁNCHEZ-PRIETO BORJA, Pedro (dir.) (2010-): *Corpus de Documentos Españoles Anteriores a 1700*. http://demos.bitext.com/codea [Consulta: 17/04/2014].

DEEDS = Documents of Early England Data Set. http://deeds.library.utoronto.ca/ [Consulta: 17/04/2014].

DÍAZ MORENO, Rocío, Rocío MARTÍNEZ SÁNCHEZ, José Luis RAMÍREZ LUENGO y Pedro SÁNCHEZ-PRIETO BORJA (en prensa): «Hacia una cronología evolutiva del español», en *Actas del IX Congreso Internacional de Historia de la Lengua Española (Cádiz, 10-14 de septiembre de 2012)*.

ENRIQUE-ARIAS, Andrés (2012): «Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad», *Scriptum Digital*, 1, pp. 85-106.

FEUERVERGER, Andrey, Peter HALL, Gelila TILAHUN y Michael GERVERS (2005): «Distance measures and smoothing methodology for imputing features of documents», *Journal of Computational and Graphical Statistics*, 14, 2, pp. 255-262.

FEUERVERGER, Andrey, Peter HALL, Gelila TILAHUN y Michael GERVERS (2008): «Using statistical smoothing to date medieval manuscripts», en Balakrishnan N., Pena E. y Silvapulle M. J. (eds.), *Beyond parametrics in interdisciplinary research: Festshrift in honor of professor Pranab K. Sen*, vol. 1, pp.321-331.

FIALLOS, Rodolfo (1997): «Procedure for dating undated documents using a relational database», en Brown J. y Stoneman W. P. (eds.), *A distinct voice: medieval studies in honor of Leonard E. Boyle*. Notre Dame, Indiana: University of Notre Dame Press, pp. 480-504.

FIALLOS, Rodolfo (2000): «An overview of the process of dating undated medieval charters: latest results and future developments», en Michael Gervers (ed.), *Dating undated medieval documents*. Suffolk: Boydell & Brewer, pp. 37-48.

GERVERS, Michael (1997): «The dating of medieval English private charters of the twelfth and thirteenth centuries», en Brown J. y Stoneman W. P. (eds.), *A distinct voice: medieval studies in honor of Leonard E. Boyle*. Notre Dame, Indiana: University of Notre Dame Press, pp. 455-480.

GERVERS, Michael (2000a): *Dating undated medieval charters*. Suffolk: Boydell & Brewer.

GERVERS, Michael (2000b): «The DEEDS project and the development of a computerised methodology for dating undated English private charters of the twelfth and thirteenth centuries», en Michael Gervers (ed.), *Dating undated medieval charters*. Suffolk: Boydell & Brewer, pp. 13-35.

JIN, Mingzhe (2009): *Tekisuto-deeta no toukei-kagaku nyuumon* (*Introducción al análisis estadístico de los textos*). Tokio: Iwanami shoten (en japonés).

KAWASAKI, Yoshifumi (2013a): «Apuntes para la datación de documentos dialectales: los casos de *-nt~-nd*, *-t~-d* final», *Hispánica* (Asociación Japonesa de Hispanistas), 57, pp. 111-117.

KAWASAKI, Yoshifumi (2013b): «Variación crono-geográfica de fórmulas en español medieval», Comunicación oral en el *LIX Congreso de la Asociación Japonesa de Hispanistas* (Tokio, 12-13 de octubre de 2013).

KAWASAKI, Yoshifumi (en prensa a): «Datación de documentos castellanos medievales» en *Actas del IX Congreso Internacional de Historia de la Lengua Española* (Cádiz, 10-14 de septiembre de 2012).

KAWASAKI, Yoshifumi (en prensa b): «Datación de documentos medievales castellanos según la 'coocurrencia' de parámetros», *Studia Romanica* (SOCIETAS JAPONICA STUDIORUM ROMANICORUM), 46.

KAWASAKI, Yoshifumi (en prensa c): «Datación por coeficientes de asociación», en *Actas del Congreso Internacional sobre el Español y la Cultura Hispánica en Japón* (ALFALito), (Instituto Cervantes, Tokio, 3 de octubre de 2014).

KAWASAKI, Yoshifumi (en prensa d): «La determinación cronológica de cambios gráfico-fonéticos y

la datación de documentos no fechados en el *CODEA*», en *Actas del II Congreso Internacional Tradición e Innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos* (Neuchâtel, Suiza, 7-9 de septiembre de 2011).

MENÉNDEZ PIDAL, Ramón (1999): *Manual de gramática histórica española* (vigésima tercera ed.). Madrid: Espasa-Calpe.

MORENO BERNAL, Jesús y Bautista HORCAJADA (1997): «Sobre *no* y *non* en español medieval», *Revista de Filología Románica*, 14, I, pp. 345-361.

MURAKAMI, Masakatsu (1994): *Shingan no kagaku: Keiryoo-bunkengaku nyuumon* (*Detectar falsificación: Introducción a la estilometría*). Tokio: Asakura shoten (en japonés).

MURAKAMI, Masakatsu (2002): *Bunka wo hakaru: bunka-keiryoogaku zyosetsu* (*Introducción a la culturametría*). Tokio: Asakura shoten (en japonés).

MURAKAMI, Masakatsu (2006): *Bunka-zyoohoogaku nyuumon* (*Culture and information science*). Tokio: Bensei shuppan (en japonés).

PENNY, Ralph (2002): *A History of the Spanish Languge* (Second Edition). Cambridge: Cambridge University Press.

SÁNCHEZ GONZÁLEZ DE HERRERO, María Nieves (dir.): *Documentación de cancillería castellana del siglo XIII.* http://campus.usal.es/~gedhytas/ [Consulta: 17/04/2014].

SÁNCHEZ-PRIETO BORJA, Pedro (1998): *Cómo editar los textos medievales: criterios para su presentación gráfica*. Madrid: Arco/Libros.

SÁNCHEZ-PRIETO BORJA, Pedro (2008): «La variación lingüística en los documentos de la catedral de Toledo (siglos XII y XIII)», en Javier Elvira, Inés Fernández-Ordóñez, Javier García y Ana Serradilla (eds.), *Lenguas, reinos y dialectos en la Edad Media ibérica*. Frankfurt/Madrid: Verfuert/Iberoamericana, pp. 233-256.

SÁNCHEZ-PRIETO BORJA, Pedro (2012): «Desarrollo y explotación del *Corpus de Documentos Españoles Anteriores a 1700* (CODEA)», *Scriptum Digital*, 1, pp. 5-35.

SÁNCHEZ-PRIETO BORJA, Pedro, Rocío DÍAZ MORENO, Rocío MARTÍNEZ SÁNCHEZ y Delfina VÁZQUEZ BALONGA (2012): «El CODEA, un corpus primario de fuentes documentales del español peninsular», en *Actas del XVI Congreso Internacional de la ALFAL*, pp. 2629-2638.

TILAHUN, Gelila (2011): *Statistical methods for dating collections of historical documents*. Tesis de doctorado, Universidad de Toronto. Inédita.

TILAHUN, Gelila, Andrey FEUERVERGER y Michael GERVERS (2012): «Dating medieval English charters», *The Annals of applied statistics*, 6, 4, pp. 1615-1640.

TORRENS ÁLVAREZ, María Jesús (1995): «La paleografía como instrumento de datación. La escritura denominada 'littera textualis'», *Cahiers de linguistique hispanique médiévale*, 20, pp. 345-380.

UEDA, Hiroto (2013): «Pautas y frecuencias grafotácticas de formas abreviadas: Su utilización para la datación de los documentos notariales del siglo XIII al XVII», Comunicación oral en el *III Congreso Internacional Tradición e Innovación: nuevas perspectivas para la edición, la investigación y el estudio de documentos antiguos* (5-7 de junio de 2013, Salamanca).

UEDA, Hiroto (en prensa): «Frecuencia contrastiva, frecuencia ponderada y método de concentración −Aplicación al estudio de las dos formas prepositivas del español medieval "por" y "para"−» en *Actas del IX Congreso Internacional de Historia de la Lengua Española* (Cádiz, 10-14 de septiembre de 2012).

UEDA, Hiroto NUMEROS.xlsm (programa informático para análisis de datos numéricos). http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/index.html [Consulta: 17/04/2014].

ZAMORA VICENTE, Alonso (1967): *Dialectología española* (segunda edición muy aumentada). Madrid: Gredos.