






Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

TESI DOCTORAL

LA CREACIÓ DE BASES DE DADES HISTORICODEMOGRÀFIQUES: PROCEDIMENTS I APLICACIONS AL CANVI SOCIAL (S. XVI-XIX)

MIQUEL VALLS FÍGOLS

Direcció:

Prof. ANNA CABRÉ PLA

Dra. JOANA MARIA PUJADAS MORA

Departament de Geografia / Centre d'Estudis Demogràfics
Universitat Autònoma de Barcelona

Doctorat en Demografia
Departament de Geografia / Centre d'Estudis Demogràfics
Universitat Autònoma de Barcelona

La creació de bases de dades historicodemogràfiques: Procediments i aplicacions al canvi social (S. XVI-XIX)

Tesi doctoral

Doctorand

Miquel Valls Fígols

Directores

Prof. Anna Cabré Pla

Dra. Joana Maria Pujadas-Mora

Tutora

Prof. Anna Cabré Pla

Any 2021

Contingut

1	Agraïments	7
2	Introducció.....	9
2.1	Objectius i estructura de la tesi doctoral.....	13
2.2	Estructura de la tesi	14
2.3	Fonts, dades i metodologia.....	17
3	Assaig 1: The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth – twentieth centuries).....	31
3.1	Introduction	31
3.2	Historical and Geographic context of the Baix Llobregat	32
3.3	Type of database	37
3.4	Transcription: Crowdsourcing platform, videogame, computer vision & volunteers.....	39
3.5	Quality checks	44
3.6	Harmonization and codification of data.....	45
3.7	Towards longitudinal data: The <i>Sant Feliu</i> life course database.....	47
3.8	Popularizing science: Onomastics browser & demographic visualizations.....	50
3.9	Concluding remarks	51
4	Assaig 2: Knowledge graph based methods for record linkage.	55
4.1	Introduction	55
4.2	Related work.....	56
4.3	The WERL method.....	59
4.4	Experiments.....	63
4.5	Discussion	67
4.6	Conclusions and future work.....	68
5	Assaig 3. A tals pares, tals fills: Transmissió intergeneracional a Sabadell, segles XVI - XVII.	69
5.1	Introducció.....	69
5.2	Els Llibres d’Esposalles de la Catedral de Barcelona.....	70
5.3	Metodologia.....	73
5.4	La transmissió intergeneracional en clau de grup social.....	76
5.5	Conclusions.	81
6	Assaig 4. What counted most for secularization: Politics, economy or family? Shifting seasonality of marriages in the Barcelona area, 1715-1880.....	83

6.1	Introduction.	83
6.2	Data & Methods.....	87
6.3	Results.	97
6.3.1	Measuring the likelihood of marrying in Lent in the Barcelona Area.	97
6.3.2	The determinants of the likelihood of marrying in lent in the Barcelona area	99
6.4	Was secularization also a family matter?.....	108
6.5	Conclusions	113
7	Conclusions.	117
8	Futures línies de treball.	125
9	Bibliografia.....	127

Índex de Taules

-	Taula 2.1: Bases de dades amb registres individuals (Regionals, Nacionals i Internacionals).....	11
-	Taula 2.2. Variables als llibres d'esposalles, 1451-1900.....	21
-	Taula 2.3. Variables en els padrons i censos de població, 1820-1959.....	24
-	Table 3.1. Number and time range of local censuses in BALL Demographic Database, by municipality.....	34
-	Table 3.2. Number of volunteer transcribers by municipality.....	43
-	Table 3.3. Individuals and households by census year in Sant Feliu de Llobregat 1828-1940.....	48
-	Table 4.1. Data Partition into train, test and validation sets.....	64
-	Table 4.2. Training set size of Evolution KG.....	64
-	Table 4.3. Record Linkage. Results for three data sets, presents Accuracy, Precision, Recall and F-score.....	66
-	Table 4.4. Additional Results for Santa Coloma de Cervelló and Castellví de Rosanes towns.....	67
-	Taula 5.1. Reconstrucció Genealògica de Sabadell (1545-1645).....	75
-	Taula 5.2. Mobilitat social intergeneracional a Sabadell (1545-1645).....	77
-	Taula 5.3. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus socio-ocupacional a Sabadell segons la posició social de pares i fills, 1545-1645.....	78
-	Taula 5.4. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus socio-ocupacional a Sabadell segons la posició social de pares i fills i per ordre de casament, 1545-1645.....	79
-	Taula 5.5. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus socio-ocupacional a Sabadell segons la ocupació de pares i fills menestrals, 1545-1645.....	81
-	Table 6.1. Codification and classification of the main occupations and social positions in the Barcelona area (1715-1880).....	89
-	Table 6.2. Number of families per number of Siblings.....	92
-	Tables 6.3. Descriptive statistics.....	95
-	Table 6.4. Logistic regressions on the likelihood of marrying in Lent by social groups, decades and political periods.....	101
-	Table 6.5. Logistic regression on the likelihood of marrying in Lent, interactions of social groups and categories of political sign.....	103
-	Table 6.6. Logistic regressions on the likelihood of marrying in Lent by social groups and sociodemographic characteristics.....	105
-	Table 6.7. Random-intercept logistic regressions on the likelihood of marrying in Lent by social groups and familial characteristics.....	109

- **Table 6.8.** Random intercept logistic models on the likelihood of marrying in Lent by social groups, familial characteristics and political sing categories.....111

Índex de Figures

- **Figura 2.1.** Organització territorial de la Diòcesi de Barcelona, segles XV-XX.....18
- **Figura 2.2.** Evolució de l'estructura dels llibres d'Esposalles.....19
- **Figura 2.3.** Imatges dels padrons municipals de Sant Feliu de Llobregat de 1828 i 1920.....23
- **Figura 2.4.** Fases de la creació de bases de dades.....25

- **Figure 3.1.** Map of Spain, Catalonia and the region of *Baix Llobregat*.33
- **Figure 3.2.** Map of BALL Demographic Database (*Baix Llobregat county*).....35
- **Figure 3.3.** Local Censuses previous to 1880.....36
- **Figure 3.4.** Local Censuses from 1880.37
- **Figure 3.5.** Social network view of census data.....38
- **Figure 3.6.** Data Model of the database.....38
- **Figure 3.7.** Crowdsourcing platform.....40
- **Figure 3.8.** Architecture of the proposed system in first scenario.....41
- **Figure 3.9.** Architecture of the proposed system in second scenario.....41
- **Figure 3.10:** Images of the two games.42
- **Figure 3.11.** Pictures of some groups of transcribers.....43
- **Figure 3.12.** Control during the transcription.....45
- **Figure 3.13.** Life course example. Bartomeu Carcereny Amigo (1874-1940).....49
- **Figure 3.14.** Onomastic browser.....50
- **Figure 3.15.** Wikipedia connection.,.....51

- **Figure 4.1.** WERL, overview.....59
- **Figure 4.2.** Example of the proposed Evolution KG representation.....61

- **Figura 5.1.** Oficialitat i deganats de la Diòcesi de Barcelona fins 1905.....71
- **Figura 5.2.** Matrimonis de la família Rossell-Bonvilar (1565-1611).....74

- **Figure 6.1.** The Barcelona area in the Spanish and Catalan context.....88
- **Figure 6.2.** Social group structure in the total of cases and by children's occupation in the genealogical sample (1750-1880).....91
- **Figure 6.3.** Index of marriage seasonality by months in the Barcelona area (1720-1880).....97
- **Figure 6.4.** Index of relative frequency of Lenten marriages in the Barcelona area (1715-1880).....98
- **Figure 6.5.** Index of Relative frequency of Lenten marriages in the Barcelona area (1715-1880) by some social groups.....99

- Figure 6.6. Probability of marrying in Lent by decades and social groups.....	102
- Figure 6.7. Probability of marrying in Lent by political periods and social groups.....	102
- Figure 6.8. Probability of marrying in Lent by social groups and political sign category.....	104
- Figure 6.9. Marginal effects on the likelihood of marrying in Lent (1715-1880).....	106
- Figure 6.10. Marginal effects on the likelihood of marrying in Lent (1807-1880).....	107
- Figure 6.11. Probability of marrying in Lent by social groups, interaction of civil status and political sign category.....	108
- Figure 6.12. Probability of marrying in Lent by the political sign period of children and political sign period of parents.....	110
- Figure 6.13. Probability of marrying in Lent by the political sign period of children and interaction of the political sign period of parents and if parents married in Lent.....	113

Índex Annex

- Annex 3.1. Variables of Local Censuses of Sant Feliu de Llobregat, 1828-1955..	53
---	----

1 Agraïments

Són moltes les persones a qui haig d'agraïr que aquesta tesi hagi arribat a bon port. En primer lloc a les meves dues directores. A la professora Anna Cabré, qui em va inocular la passió per la demografia i em va obrir les portes del Centre d'Estudis Demogràfics en aquell ja llunyà 1999. Gràcies per deixar-me per formar part de l'Àrea de Demografia Històrica del Centre d'Estudis Demogràfics, a aprendre a “pensar demogràficament” i poder gaudir del teu mestratge. Per altra banda, i molt especialment, a la Dra. Joana Maria Pujadas-Mora amb qui he après la passió per la recerca. La seva força de treball i generositat infinita l'han convertit en *l'alma mater* d'aquesta tesis doctoral. Sense ella, res de tot això hauria estat possible. No tinc suficients paraules per agrair-li tot el seu suport acadèmic, i l'amistat que ens hem forjat durant tots aquest anys. En aquestes primeres línies no puc deixar d'esmentar a la Dra. Àngels Torrents, qui em va transmetre la seva passió per la Demografia Històrica i de la seva mà vaig poder entrar dins del món dels arxius eclesiàstics amb l'elaboració de l'Inventari dels Arxius Sagramentals de Catalunya. Sé que des d'allà dalt, vetlla per tots nosaltres i per la demografia històrica.

També voldria agrair a les dues institucions que m'han allotjat durant tot aquest temps, per un costat el Centre d'Estudis Demogràfics i el seu actual director, el Dr. Albert Esteve; i el Departament de Geografia de la Universitat Autònoma de Barcelona.

A tots els membres de l'Àrea de Demografia Històrica del Centre d'Estudis Demogràfics amb els quals he compartit despatx, seminaris, congressos, conferències, dinars, sopars,... Joan Pau Jordà i Miquel Amengual (Mallorquins de soca-rel). Especialment al Dr. Gabriel Brea-Martínez (l'astre futbolístic brasiler) per la seva ajuda en els models estadístics i amb qui he compartit tant projecte de tesi com habitacions d'hotel. Així com també els doctorants Jaume Vendrell (Climentó il·lustre) i Joaquín Ruiz (Mazarronero d'orgien) amb qui he pogut gaudir de les vídeo-trucades setmanals en temps de pandèmia Covid-19 i que m'han fet costat en els moments finals d'aquesta tesis i dels quals he après moltíssim.

En un procés de tesis tan llarg com el meu, és impossible esmentar a totes les companyes i companys del CED sense l'enorme perill de deixar-se algú. A totes les persones que m'han ajudat en tot aquest temps, sabeu – veritablement – que us estic enormement agraït.

Als membres del Centre de Visió per Computador, el Dr. Josep Lladós, la Dra. Alícia Fornés, el Dr. Oriol Ramos, el Dr. Joan Mas, l'Alexandra Canet i en Jialuo Chen, per fer del seu centre de recerca la meva segona llar acadèmica, i compartir projectes d'investigació fascinants.

I en darrer lloc, però no menys important, als meus pares Josep i Teresa; i com no podia ser d'una altra manera a la meva família. A la Marta, per estar sempre al meu costat, sobretot en aquells moments que m'he enfonsat i ella molt pacientment ha sapigut tornar a fer flotar el vaixell. Al Roger, per la seva passió pel futbol i al Ferran per la seva actitud davant la vida. Tots ells m'han ajudat a continuar fent realitat els somnis. Us estimo molt. A totes i tots, moltes i moltes gràcies.

2 Introducció.

La construcció de les base de dades en la Demografia Històrica, no és un fenomen nou, així com tampoc ho és la integració de la ciutadania en la recollida de les dades, tal i com ho va demostrar l'experiència pionera del *Cambridge Group for the History of Population and Social Structure*, quan va realitzar la reconstrucció de la població d'Anglaterra amb voluntaris repartits per diferents parròquies (Wrigley et al. 1997:6). La Demografia Històrica ha estat sempre una disciplina que ha utilitzat grans quantitats de dades en els seus estudis, tenint sempre en compte la limitació tècnica de cada període (Billari i Zagheni, 2017). Així cal a dir que existeixen un gran nombre de bases de dades demogràfiques històriques en l'actualitat moltes de les quals s'iniciaren en els anys 70 i 80 del segle XX que han perdurat en ús fins a l'actualitat (Ruggles, 2012; Lee et al., 2010; Mandemakers, 2002; Edvinsson, 2000; Dillon et al., 2018; Bengtsson et al. 2012). D'aquesta manera, es disposen de milers de dades de registres individuals però la gran majoria fan referència a zones del nord d'Europa.

Si fem un repàs de les grans bases de dades demogràfiques existents (veure taula 1), podem diferenciar entre bases de dades d'àmbit regional o nacional i un altre d'àmbit internacional. Pel que respecta a Europa, s'observa una llarga tradició en els països nòrdics, Finlàndia, Noruega i Suècia, en la construcció d'àmplies bases de dades, així com també en els Països Baixos. Algunes d'aquestes bases són la *Scanian Economic Demographic Database* que recull les reconstruccions de famílies de 9 parròquies cobrint des de 1646 fins a l'actualitat. A més a més per 5 d'aquestes parròquies recull informació sobre el total de la població des de 1813. En global, la base de dades conté més de 150.000 registres individuals (Bengtsson et al., 2014). Un altre exemple a destacar és l'anomenat Norwegian Historical Population Register que inclou dades procedents dels censos de població i dels registres parroquials de Noruega des de 1800 fins a 1964, data en la qual ja tenen el registre continu de població. Aquesta base està en construcció en l'actualitat i es persegueix reconstruir el major nombre de persones que vivien a Noruega a partir del segle XIX (Thorvaldsen et al., 2015).

Amb informació sobre 660.000 persones i gairebé 5 milions de registres entre 1620 i 1900, la base de dades individuals històrica de Suècia (POPUM) és una de les més grans del món. Es recullen informacions procedents dels registres eclesiàstics, això com també

dels registres civils o registres de migracions. Combinada amb la POPLINK, el vincle a nivell individual entre dades històriques i modernes, abasta gairebé 400 anys (Edvinsson, 2000).

Als Països Baixos cal ressaltar la Historical Sample of the Netherlands que és una mostra de dades longitudinals que permet resseguir les trajectòries de vida d'unes 78.000 persones nascudes entre 1812 i 1922 en aquest país (Mandemakers, 2002). En el cas França i Bèlgica cal ressenyar la reconstrucció de les línies de vida de les persones que els seus cognoms comencen amb les lletres TRA per França entre 1800 i 1960 (Bourdieu et al., 2014), i per les persones del districte d'Anvers amb la combinació de les inicials COR des de 1846 fins 1920 (Matthijs i Moreels, 2010).

Pel que fa al sud d'Europa, tant a Espanya, Itàlia com Portugal, les bases de dades corresponen a recopilatoris de registres sacramentals de diverses parròquies, principalment rurals, que van des de 1550 fins 1910, però encara estan en processos d'integració per obtenir una base conjunta a cada territori. De tal manera que malgrat hi ha una gran quantitat de fons documentals per explotar, no hi han hagut fins ara projectes que hagin tingut per objectiu aglutinar i crear grans bases de dades històriques (Reher i Sanz Gimeno, 2007; Guimaraes, 2013; Breschi et al., 2020). Això demostra que encara hi ha molt camí per recorre sobretot en el cas català i espanyol, tot i que aquesta tesi és la mostra de que s'ha avançat en aquest sentit.

Si en fixem amb altres territoris no europeus, cal destacar les bases de dades historicodemogràfiques del Canadà, especialment les de la regió del Quebec. La BALSAC: Quebec population data base recull genealogies familiars, creades mitjançant vinculació nominal, des de 1621 fins a 1971 per a tota la regió (Jomphe, 2012). Actualment, el projecte BALSAC, s'ha integrat juntament amb el Registre de la Population du Québec Ancien (Desjardins, 1998) – que recull dades demogràfiques de la població catòlica de la vall del riu San Lorenzo – dins el programa de recerca interdisciplinari de demografia històrica del govern del Canadà. Les bases de dades històriques de la Xina, són molt rellevants, no tant sols pel volum de registres, sinó també perquè permeten l'estudi de processos socials i demogràfics longitudinals i multigeneracionals; ja que incorporen dades de les persones, les llars, les comunitats, la tinença de la terra,... La China Multigenerational Panel Database – Liaoning (Lee et al., 2010), esta formada per 1,5 milions d'observacions de més de 260.000 residents a 698 comunitats de la província xinesa de Liaoning, entre 1749 i 1909. Per altra banda, la China

Multigenerational Panel Database – Shuangcheng (Dong et al., 2015), recull informació longitudinal individual, familiar i comunitària amb 1,3 milions d'observacions i més de 100.000 individus únics que van viure en aquesta regió entre 1866 i 1913.

Taula 2.1: Bases de dades amb registres individuals (Regionals, Nacionals i Internacionals)

	Nom	Acrònim	Territori
Bases de dades Nacionals	Founders & Survivors: Tasmanian life courses in historical context	FAS	Austràlia
	Koori Health Research Database	KHRD	Austràlia
	Melbourne Lying-In Hospital Cohort	MLHC	Austràlia
	Female Demographic Biographies: Wald parish, 1880-1938	FDB-Wald	Austràlia
	Antwerp COR-database	COR	Bèlgica
	Historical Database of the Liège Region	HDLR	Bèlgica
	BALSAC: Quebec population database, 1621-1971	BALSAC	Canadà
	National Sample of the 1901 Census of Canada	CFP	Canadà
	Registre de la population du Québec ancien	RPQA	Canadà
	Canadian Century Research Infrastructure	CCRI	Canadà
	Odense database: Persons and buildings in Odense, 1741-1921	ODENSE	Dinamarca
	The Demography of Victorian Scotland	DVS	Escòcia
	Aranjuez Database: Individual and family trajectories	ADB	Espanya
	Texas Counties Database	TCD	Estats Units d'Amèrica
	Utah Population Database	UPDB	Estats Units d'Amèrica
	Karelian Database	KATIHA	Finlàndia
	Base TRA Patrimoine	TRA	França
	Italian Historical Population Database	IHPD	Itàlia
	Historical Sample of the Netherlands	HSN	Països Baixos
	Integral History Project Goningen	IHPG	Països Baixos
	Norwegian Historical Populaton Register, 1800-1964	HPR	Noruega
	Historical Sample Portuguese Social Mobility, 1850-1960	HSPM	Portugal
	Portuguese Genealogical Repository	PORGENER	Regne Unit
	Integrated Census Microdata (I-CeM) Project	I-CeM	Romania
	Historical population database of Transylvania, 1850-1914	HPDT	Romania
	Hungarian Historical Demographic Database	HHDD	Suècia
	POPLINK	DDB	Suècia
	POPUM	DDB POPUM	Suècia
	Scanian Economic Demographic Database	SEDD	Suècia
	Geneva Demographi Database	GDD	Suïssa
	The Roteman Database	TRD	Suècia
	China Multigenerational Panel Database-Shuangcheng	CMGPD-LN	Xina
China Multigenerational Panel Database-Liaoning	CMGPD-SC	Xina	
Bases de dades Internacional	North Atlantic Population Project	NAPP	Atlàntic Nord
	European Historical Population Samples Network	IDS/EHPS-Net	Europa
	Mosaic Project	MOSAIC	Europa

Font: Elaboració pròpia a partir de la European Historical Population Samples Network i el projecte IPUMS.

Finalment hi ha el cas australià, en que trobem tres exemples diversos de bases de dades que abasten des del segle XIX fins a inicis del XX, en les que es poden resseguir les

73.000 persones que es van ser traslladades a Tasmània per crear una nova societat al 1803, fins els 8561 registres sobre els naixements i les defuncions ocorregudes a l'hospital de Melbourne, o sobre les genealogies de 30.000 aborígens creades en el marc del projecte Founders&survivors.

Pel que fa als projectes d'àmbit internacional destaquen el North Atlantic Population Project (NAPP) que recull dades del censos del segle XIX d'Anglaterra, Canadà, Islàndia, Noruega, Suècia i els Estats Units (Ruggels et al. 2011). En aquest sentit, a Europa s'han constituït dos projectes, el primer és la European Historical Population Data Sample que aglutina un gran nombre de dades longitudinals, construïdes a partir dels registres d'esdeveniments vitals, com la Demographic Data Base i la Scanian Demographic Databas, de Suècia, la Balzac Population Register de Quebec, la Base TRA Patronime de França, i la Historical Sample of the Netherlands, dels Països Baixos (Alter & Mandemakers, 2014), entre d'altres i que s'han anomenat anteriorment. El segon projecte d'àmbit europeu és el Mosaic Project, que recull dades de censos històrics a nivell local de 28 països d'Europa coordinat pel Max Planck Institute for Demographic Research. Malgrat les mostres dels municipis siguin petites, aquestes tenen un format integrat i es poden realitzar comparacions històriques al llarg del temps i l'espai (Szoltysek & Gruber, 2015).

L'aparició de la informàtica va suposar un salt en la creació de les baes de dades i el seu posterior anàlisi i explotació. En l'actualitat, l'aplicació d'algoritmes d'Intel·ligència Artificial obre les portes a un nou canvi de paradigma dins la Història i les Ciències Socials (Ruggles, 2012; Matthijs et al. 2016). Ens trobem immersos dins l'era del Big Data i la Demografia Històrica no en pot quedar al marge. Disposem de les eines informàtiques per analitzar grans volums de dades i la utilització del Big Data¹ (Brands, 2014; Fox & Do, 2013) ens condueix a una nova forma d'entendre i concebre les bases de dades històriques.

Així a la construcció de bases de dades, la integració de la Intel·ligència Artificial a través de l'anomenada visió per computador del reconeixement de manuscrits, pot suposar reduir el temps de creació de les bases de dades. Els avenços en visió per computador

¹ Entenen Bigdata com: els conjunts de dades, els procediments i les aplicacions informàtiques, que, pel seu volum, la seva naturalesa diversa i la velocitat a què han de ser processades, ultrapassen la capacitat dels sistemes informàtics habituals. Aquest processament de dades massives s'utilitza per detectar-hi patrons, fent així prediccions vàlides per a la presa de decisions. Aquestes són les tres "V" que defineixen el Big Data (Volum, Varietat i Velocitat) (Ward & Barker, 2013).

proporcionen als investigadors algoritmes per al reconeixement d'imatges de text manuscrit. Actualment les tècniques d'OCR ja estan molt avançades en el reconeixement de caràcters i obren les portes a nous horitzons en les tècniques de reconeixement manuscrit (HTR) i de Keyword Spotting (KWS), que faciliten el buidatge massiu de fons documentals (Lladós et al., 2012).

No obstant això, l'estat actual d'aquestes metodologies encara necessiten de validació per part dels investigadors. Així encara es convenient mantenir el “*human in the loop*” com s'anomena en l'argot computacional. D'aquí la importància de la ciència ciutadana. De fet, la incorporació de la informàtica al dia a dia de les persones, i l'arribada massiva d'internet a les llars ha fet que en les darreres dècades la col·laboració de la ciutadania amb la recopilació de dades per a estudis científics hagi augmentat molt i s'hagi vist com una forma productiva per avançar en el coneixement (Bonney et al., 2009; Bonney, 2014; Newman et al., 2012, Vohland et al. 2021). A més a més, l'expansió educativa, l'augment del temps dedicat al lleure així com l'entrada a la jubilació d'un gran nombre de persones amb bona salut i amb inquietuds, fan que hi hagi una gran massa crítica susceptible a col·laborar amb projectes de ciència ciutadana (Haklay, 2015; Kummel, 2018). Per altra banda, des dels arxius s'està portant a terme un procés intensiu de digitalització dels seus fons documentals. Així, combinant Intel·ligència Artificial i ciència ciutadana, es cerca extreure de forma massiva la informació de la documentació original, en el nostre cas padrons d'habitants, censos de població, registre civil,... i tot allò referent a la població en el passat.

2.1 Objectius i estructura de la tesis doctoral.

El nostre propòsit és mostrar la importància de la creació de grans bancs de dades històrics que permetin entendre molt millor els comportaments i els mecanismes demogràfics i socials del passat en el marc del canvi social i el procés de modernització de la societat.

Objectius específics de la tesis doctoral.

- 1) Crear bases de dades demogràfiques.

- a) Recollida de dades a través de la digitalització de la documentació demogràfica històrica i aplicar processos de transcripció mitjançant *Crowdsourcing* i accions de Ciència Ciutadana.
 - b) Dissenyar aplicacions web per la recollida massiva de dades demogràfiques, mitjançant formularis de transcripció interactius.
 - c) Utilitzar les tècniques de visió per computador per escurçar el temps de construcció de bases de dades demogràfiques
 - d) Gamificar la validació automàtica de transcripcions de dades demogràfiques.
 - e) Normativitzar, harmonitzar i codificar les variables de les bases de dades per poder realitzar estudis i anàlisis comparables a nivell nacional i internacional.
 - f) Vincular nominalment els registres (Record Linkage), a través d'algoritmes per a la construcció de bases de dades longitudinals.
- 2) Mostrar les diferents potencialitats que aquests nous bancs de dades amb informació individual i longitudinal poden oferir per a la recerca de noves preguntes per conèixer el nostre passat així com la possible reinterpretació de conceptes i visions clàssiques sobre la societat.
- Dins aquestes potencialitats ens centrem en dos temes concrets:
- a) Estudiar el procés de transmissió intergeneracional de l'ocupació entre pares i fills als segles XVI i XVII.
 - b) Analitzar els processos de secularització a l'àrea de Barcelona mitjançant l'estudi de l'estacionalitat dels matrimonis, no tant sols observant el canvi de calendari, sinó afegint nous conceptes com la importància del sistema polític del moment o la transmissió del comportament familiar i la dedicació ocupacional.

2.2 Estructura de la tesi

La tesi doctoral esta concebuda a partir de la redacció de quatre assajos fruit de la tasca conjunta en el si de l'Àrea de Demografia Històrica del Centre d'Estudis Demogràfics dins el marc dels projectes de recerca "*Five Centuries of Marriages*" (5CofM) (FP7-ERC-2010-AdG-269796) dirigit per la Dra. Anna Cabré. I amb la col·laboració conjunta amb el Centre de Visió per Computador dels projectes: "*Eines i procediments per la*

informatització massiva de fonts històriques de població”, dirigit pel Dr. Josep Lladós i la Dra. Joana Maria Pujadas-Mora; i del projecte “*Xarxes. Tecnologia i innovació ciutadana en la construcció de xarxes socials històriques per la comprensió del llegat demogràfic*” liderat per la Dr. Joana Maria Pujadas-Mora i la Dra. Alícia Fornés, ambdós finançats pel programa RecerCaixa de la Fundació La Caixa i l’Associació Catalana d’Universitats Públiques (ACUP) en 2014 i 2017 respectivament. A més a més, a partir del 2019 estem treballant en el projecte *DemoDesigual: Condicionantes demográficos de la desigualdad económica, una aproximación histórica (Siglos XVIII-XX)*, encapçalat per la Dra. Joana Maria Pujadas-Mora i finançat pel Ministerio de Ciencia, Innovación y Universidades (RTI2018-095533-B-I00).

El primer assaig reflexiona sobre el procés de construcció de grans bases de dades històrico-demogràfiques.

- 1) *The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth – twentieth centuries)*. Aquest assaig ha estat publicat en format de capítol en el llibre “Nominative Data for Demographic Research in the East and the West” editat per a la Ural Federal University Press, pàgines 29-61 al 2019. També en són autors Joana Maria Pujadas-Mora, Alícia Fornés, Josep Lladó i Gabriel Brea-Martínez.

El segon és un treball sobre diferent tipus de metodologia de record linkage.

- 2) *Knowledge graph based methods for record linkage*. Aquest assaig ha estat publicat a la revista *Patterns Recognition Letters* 136:1-7 al 2020, , juntament amb Bhaskar Goutam, Oriol Ramos Terrades i Joana Maria Pujadas-Mora.

El tercer assaig analitza la transmissió intergeneracional de l’ocupació a Sabadell a època moderna.

- 3) *De tals pares, tals fills: transmissió intergeneracional a Sabadell, segles XVI-XVII*. Aquest assaig ha estat publicat a la revista *Arraona* 38:74-85 al 2019,. També en són autors Joana Maria Pujadas-Mora i Gabriel Brea-Martínez.

El quart i darrer assaig tracta sobre el procés de secularització a l’Àrea de Barcelona entre 1715 i 1880.

- 4) *What counted most for secularization: Politics, economy or family? Shifting seasonality of marriages in the Barcelona area, 1715-1880.* Aquesta recerca s'ha dut a terme dins de l'equip de recerca de l'Àrea de Demografia Històrica del Centre d'Estudis Demogràfics, juntament amb Joana Maria Pujadas-Mora, Gabriel Brea-Martínez i Anna Cabré.

A la part final hi ha un apartat dedicat a extreure les principals conclusions de la tesis així com preveure futures línies de treball.

2.3 Fonts, dades i metodologia

Fonts documentals

Al llarg d'aquesta tesis hem confeccionat dues grans bases de dades de naturalesa diferent, però amb la visió de construir estructures estables a partir de les quals puguem anar afegint noves dades i establir diccionaris d'harmonitzacions i codificacions que serveixen per futures bases de dades. La primera a partir de registres eclesiàstics de tipus fiscals sobre matrimonis a la Diòcesi de Barcelona i la segona a partir dels padrons i censos d'habitants realitzats per l'administració pública. Aquesta segona base de dades també s'ha complementat amb els padrons de les cèdules personals, que ens proporciona informació sobre renda, lloguer i contribució de les persones majors de 14 anys de cada municipi.

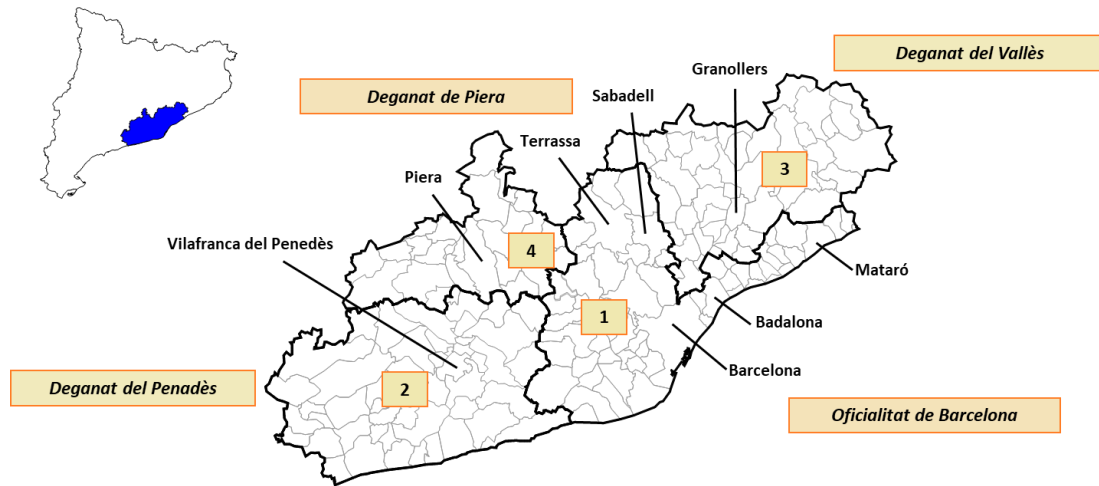
Barcelona Historical Marriage Database.

Els llibres d'esposalles estan formats per 291 volums, 244 dels quals corresponen a l'Oficialitat de Barcelona (el Deganat principal de la Diòcesi de Barcelona) i els 47 llibres restants corresponen al Deganat del Vallès (18 volums) pel període 1541-1803; al Deganat del Penedès (19 volums) pel període 1563-1802 i el Deganat de Piera (10 volums) pel període 1572-1831 (veure Figura 2). L'Oficialitat de Barcelona estava format per les actuals comarques del Barcelonès, gran part del Vallès Occidental, del Baix Llobregat i del Maresme. El Deganat del Penedès es delimitava entre les comarques del Alt i Baix Penedès i el Garraf. El Deganat del Vallès estava format pel Vallès Oriental i una part del Vallès Occidental i finalment el Deganat de Piera es localitzava a la comarca de l'Anoia.

La base de dades de matrimonis consta de 612.486 matrimonis en 285 parròquies distribuïts de la següent manera: Oficialitat de Barcelona 556.316 matrimonis en 106 parròquies diferents. El Deganat del Vallès amb 13.433 registres en 77 parròquies. El Deganat del Penedès amb 27.946 registres en 70 parròquies i el Deganat de Piera amb 14.791 registres i 32 parròquies. Tots aquests volums estan escrits en català, encara que en alguns hi podem trobar algunes paraules o textos en llatí, fins a 1860. A partir del 1 de gener de 1861, els llibres estan escrits en castellà. Entrant més en el detall d'aquesta font d'informació podem distingir dues sèries. La primera que correspon als llibres de l'Oficialitat de Barcelona (que és la més completa tant en informació com en període de

temps) i una segona que correspon als registres dels tres deganats (que és més incompleta i més reduïda en el temps).

Figura 2.1. Organització territorial de la Diòcesi de Barcelona, segles XV-XX.



Font: Elaboració pròpia.

Oficialitat de Barcelona.

El primer llibre conservat d'aquesta sèrie correspon al període 1451-1453. Els llibres no tenen una numeració i és molt probable que n'hi hagi d'anteriors que no han arribat fins als nostres dies, i fa suposar que era la continuació d'una llarga sèrie². Els llibres tenen una periodicitat bianual, amb la peculiaritat que s'inicien al primer de maig (seguint la costum administrativa de la Catedral de Barcelona i el seu Capítol per l'advocació a la Santa Creu que es celebra el dia 3 de maig (Baucells i Reig, 2002) fins el 30 d'abril de l'any següent. D'aquesta manera, cada volum inclou informació referent a 3 anys, però només un de sencer (el del mig). Aquesta periodicitat la trobem fins a 1860, els volums van passar a ser anuals, registrant les esposalles des de l'1 de gener fins el 31 de desembre, possiblement com a conseqüència de l'increment en el nombre de matrimonis degut a l'augment de la població.

Hem comentat que la sèrie de llibres va des de 1451 fins 1905, però malauradament hi ha alguns buits d'informació com a conseqüència de la desaparició dels llibres. Així la sèrie s'inicia amb dos volums que abasten des de 1451 fins 1455, i després hi ha una llacuna fins al 1481 en la qual s'estima que hi ha uns 13 volums desapareguts. Des de 1487 fins

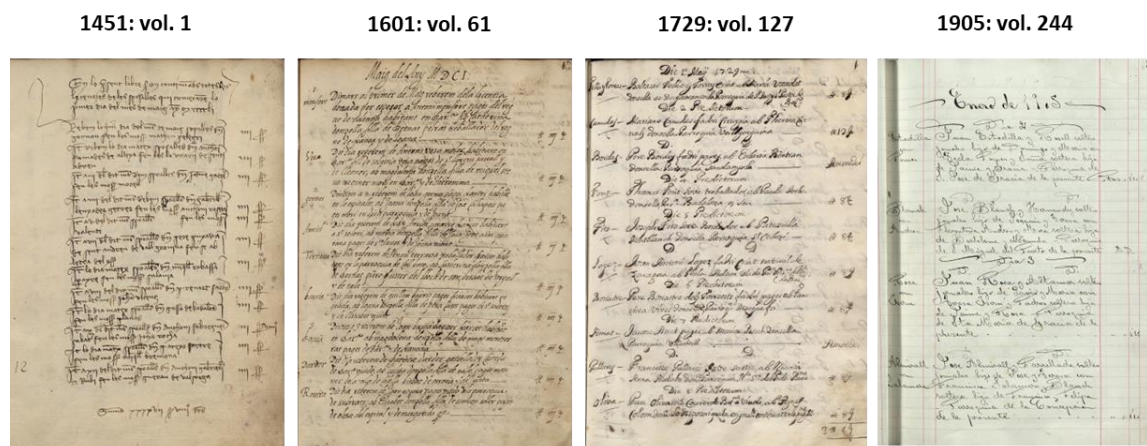
² Recordem que la concessió del impost per part de Benet XIII, es de 1409, o sigui 42 anys abans del primer volum conservat.

1491 es documenta una nova interrupció que suposarien 2 volums i a partir d'aquesta data fins al final tant sols manquen el volum 87 (1653-1655) i el volum 155 (1782-1784)³.

Estructura de la font.

Els llibres d'Esposalles tenen una estructura molt estable al llarg dels cinc segles d'informació que contenen. Hi trobem variacions en el seu contingut al llarg dels cinc segles, però no en l'estructura típica en forma de llibre de comptes que es manté en tres blocs diferenciats. A la part esquerra hi trobem el cognom del nuvi⁴, o d'ambdós contraents (aquesta segona versió en èpoques més recents, a partir de 1878). A la part central es registren els camps amb les variables que fan referència al matrimoni i finalment, una tercera columna a la dreta amb la quantitat a pagar per part dels contraents.

Figura 2.2. Evolució de l'estructura dels llibres d'Esposalles.



Font: Arxiu de la Catedral de Barcelona

En el cas dels llibres que registren la informació dels deganats, aquests tenen una enquadernació més petita pel que fa al format del llibre i a part dels registres individuals, també contenen anotacions col·lectives de parròquies senceres que liquiden tots els matrimonis d'un període en un sol import, fet que no trobem en els registres de l'Oficialitat de Barcelona. De tal manera, que en aquests casos tan sols disposem del total

³ En aquests dos volums (87 i 155) si que es conserven els índexs de matrimonis realitzats a posteriori per membres de l'arxiu.

⁴ Fins a 1539 no apareix aquesta primera columna amb el cognom del nuvi, i només hi havia dues columnes, la de la informació del matrimoni i la de l'import a pagar.

agregat. A tall d'exemple citem una liquidació del Deganat del Penedès referent a la parròquia de Vilafranca del Penedès.

“nosaltres bernat masdeu y pere roig obres menors de la taula de la obra de la seu de barcelona avem rebut del venerable mosen salvador vilella de vilafranca de penedes per mans del senyor pau moradell la suma de trenta set liures y quatre sou” (Deganat de Piera, Volumen 55, pàgina 41).

Els Llibres d'Esposalles estaven registrats per dos obrers menors⁵ i tenien una periodicitat de dos anys, amb la peculiaritat que els registres comencen a dia 1 de maig i finalitzen el 30 d'abril de l'any següent.

Les esposalles recullen informació sobre l'enllaç matrimonial. Hi ha unes variables que es mantenen durant els cinc segles i d'altres que van variant (veure Taula 1). La data del pagament, el nom i primer cognom del marit i la taxa a pagar, es recullen al llarg del temps. El nom de la muller, també s'anota en tots els períodes a excepció dels primers volums (1451-1499) en el que apareix el nom i cognoms del mossèn i el nom i primer cognom del pare de la muller. Pel que fa a la informació sobre els pares dels contraents, noms, cognoms, ocupació i residència, aquesta es recull sobretot durant el període 1550 fins 1650, que és el més ric en informació. Malauradament, durant el període 1650-1720 es produeix una reducció dràstica de la informació recollida en els llibres d'esposalles i només hi trobem informació sobre ambdós contraents. Per altra banda, a partir de 1720 es deixa de registra l'origen i residència, a excepció dels que són de fora de la diòcesi, per incorporar la parròquia de celebració de l'enllaç.

Pel que fa a la taxa de pagament, aquesta no està fixada fins a mitjans del segle XVI en que s'estableixen 7 categories: Alta noblesa, baixa noblesa, ciutadants honrats, mercaders i professions liberals, artistes, camperols i finalment un darrer grup exempt de pagament per les classes més baixes (Amore Dei). Aquestes taxes eren una unitat de compta valorada en sous, que es va mantenir estable fins a 1650 en el que el tots els valor d'aquestes taxes monetàries es dupliquen. En els darrers anys, a partir de 1892 s'utilitza la pesseta com a unitat de compte.

⁵ Els obrers menors eren els subordinats dels obrers majors, els sacerdots al capdavant de l'obra de la Catedral. El càrrec d'obrer menor era bianual i la seva tasca principal era gestionar i registrar els pagaments de les llicències d'esposalles (Carreras i Candi, 1913)

Taula 2.2. Variables als llibres d'esposalles, 1451-1900.

		Esposalles								
		1451-1499*	1500-1549	1550-1559	1600-1649	1650-1699**	1700-1749	1750-1799**	1800-1849	1850-1905
General	Nom i cognom del mossén	X								
	Data del pagament	X	X	X	X	X	X	X	X	X
	Parròquia de celebració						X	X	X	X
	Taxa	X	X	X	X	X	X	X	X	X
Marit	Nom	X	X	X	X	X	X	X	X	X
	Primer cognom	X	X	X	X	X	X	X	X	X
	Segon cognom									X
	Estat civil				X	X	X	X	X	X
	Ocupació	X	X	X	X	X	X	X	X	X
	Origen		X	X	X	X	X			
	Residència	X	X	X	X					
Muller	Nom		X	X	X	X	X	X	X	X
	Primer cognom					X	X	X	X	X
	Segon cognom									X
	Estat civil		X	X	X	X	X	X	X	X
	Ocupació									
	Residència									
Pare del marit	Nom			X	X			X	X	X
	Primer cognom			X	X			X		
	Segon cognom									
	Estat civil									
	Ocupació			X	X					
	Residència			X	X					
Mare del marit	Nom			X	X			X	X	X
	Primer cognom							X	X	X
	Segon cognom									
	Estat civil									
	Ocupació									
	Residència									
Pare de la muller	Nom	X	X	X	X			X	X	X
	Primer cognom	X	X	X	X			X		
	Segon cognom									
	Estat civil									
	Ocupació		X	X	X					
	Residència	X	X	X	X					
Mare de la muller	Nom			X	X			X	X	X
	Primer cognom					X		X	X	X
	Segon cognom									
	Estat civil									
	Ocupació									
	Residència									

* 1451-1455; 1481-1487; 1491-1499

** Excepte 1653-1655; 1782-1784.

Font: Elaboració pròpia

BALL Database: padrons i censos d'habitants

Pel que respecte a les fonts demogràfiques dins l'àmbit civil, cal destacar els registres de població de potestat municipal, els anomenats padrons d'habitants, i els d'àmbit estatal, els censos⁶, així com també el Registre Civil per a la inscripció dels naixements,

⁶ Reher, D. S. i Valero Lobo, Á. (1995): Fuentes de información demográfica en España, Madrid, Cuadernos Metodológicos, 13, Centro de Investigaciones Sociológicas; Cusidó i Vallverdú, T. A., Gil-Alonso (2012): Los censos en España: entre continuidad y cambio (1857 – 1970). Revista de Demografía Histórica, XXX, I, 29-67

matrimonis i defuncions. Uns instruments que a Espanya s'iniciaren al segle XIX com a conseqüència del procés d'enfortiment de l'Administració Pública en la instauració de l'Estat Liberal cercant la seva racionalització a partir del coneixement de la composició de la seva població o de les fonts de riquesa del seu territori⁷. Un procés que fou comú a l'Europa del segle XIX, encara que no simultani. Aquests registres mantenen la seva vigència en l'actualitat i habitualment es conserven els seus originals des dels inicis, malgrat que en el cas dels Censos de Població, les fulles individuals moltes vegades es destruïen i només es conservaven les dades agregades .

Els padrons de població s'institucionalitzaren a partir de 1823 amb la *Instrucción para el Gobierno Económico-Político de las Provincias, Islas y Posesiones Adyacentes*. No obstant això, amb anterioritat es poden trobar algunes iniciatives de caràcter local ja que la tradició censal, encara que no amb finalitat demogràfica sinó més de bé fiscal, havia començat segles enrere. A 1870 es decretà que la seva periodicitat havia de ser de cada cinc anys (en els anys acabats en 0 i 5), tema que es tornà insistir en una disposició de 1924⁸. Malgrat tot, com que l'aixecament dels padrons d'habitants era de potestat municipal, molts municipis conserven padrons d'habitants realitzats en altres anys (sobretot durant el segle XIX), fet que encara fa més rica aquesta font⁹.

És en aquest marc que el projecte: *XARXES: Tecnologia i innovació ciutadana en la construcció de xarxes socials històriques per a la comprensió del llegat demogràfic*, dirigit per la Dra. Joana Maria Pujadas-Mora i la Dra Alícia Fornés unissin esforços des de la demografia històrica i la visió per computador, respectivament, per desenvolupar instruments i procediments que facilitin la informatització massiva de les fonts demogràfiques del passat, orientada a la construcció de bancs de dades d'ús públic i a la millora de la consulta de documents dels arxius. I més concretament els padrons municipals d'habitants.

⁷Porter, T. (1995): Trust in numbers. The pursuit of objectivity in science and public life. Princeton: Princeton University Press i Woolf, S. (1989): Statistics and the modern state. Comparative Studies in Society and History, vol. 31 (3), 588-603.

⁸ Per a més informació sobre l'origen dels padrons veure: García Pérez, M. S. (2007): El padrón municipal de habitantes: origen, evolución y significado. Hispania Nova. Revista de Historia Contemporánea, núm. 7, <http://hispanianova.rediris.es/7/articulos/7a005.pdf>

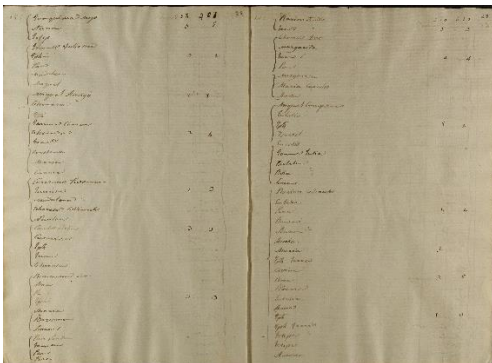
⁹ Les principals lleis que han regulat els padrons d'habitants són: Ley de Administración y organización municipal, de 4 de julio de 1856; Ley Municipal de 20 de agosto de 1870; Reglamento sobre población y términos municipales de 2 de julio de 1924; Llei 7/1985 de 2 d'abril, Reguladora de les Bases de Règim Local; Real Decret 2612/1996 de 20 de desembre; Llei 4/1996 de 10 de gener.

Estructura dels padrons d'habitants.

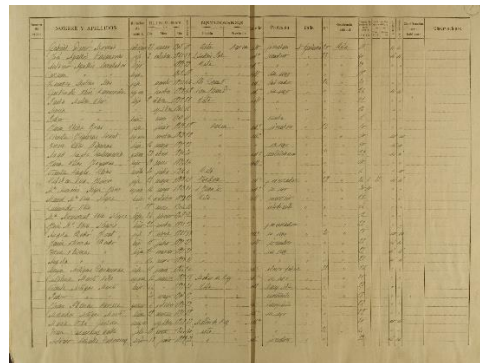
Els padrons són els recomptes municipals nominatius de població per excel·lència a Espanya i constitueixen una fotografia fixa de la població en un lloc i en un temps concret que mostra per cada un dels habitants les seves característiques sociodemogràfiques (veure Figura 2.3). D'aquesta manera, la unitat de recollida d'informació dels padrons és l'individu identificat en una llar amb una adreça concreta en el cas urbà o fora d'aquest. Per a cada individu es recollia el nom i cognoms, l'edat o data de naixement, l'estat civil i l'ocupació i molt sovint la relació familiar o laboral amb el cap de la llar. En alguns moments també s'hi pogueren afegir informacions sobre l'alfabetització o el nivell de rendes dels individus (veure taula 2.2). Els padrons abasten tot el territori espanyol ja que era i continua sent obligatori que cada municipi confeccionés els seus propis. Els padrons més antics no es realitzaren en un mateix moment en tota la geografia, com si es realitzaven els censos nacionals, considerats moderns a partir de 1857, però amb el temps tots els municipis acabaren realitzant-los en una mateixa data.

Figura 2.3. Imatges dels padrons municipals de Sant Feliu de Llobregat de 1828 i 1920.

Padró de 1828



Padró de 1920



Font: Arxiu Comarcal del Baix Llobregat.

Durant el segle XIX, l'administració estatal va intentar coordinar i uniformar les variables que es recollien en els padrons d'habitants. Al ser una funció de potestat municipal, cada ajuntament realitzava el recompte a la seva manera. Serà a partir de la realització del Cens de població de 1857, que des de l'administració central, comencen a establir quines eren les variables que havien de consignar-se en els padrons. No serà fins a la dècada de 1880 quan comencen a imprimir-se formularis estàndards per recollir la informació. Uns

formularis que no s'unificaran del tot fins als principis de segle XIX, doncs la seva impressió corria a càrrec de cada ajuntament

Taula 2.3. Variables en els padrons i censos de població, 1820-1959.

Variables	Padrons i Censos de Població												
	1820-1829	1830-1849	1850-1859	1860-1879	1880-1889	1890-1899	1900-1909	1910-1919	1920-1929	1930-1939	1940-1949	1950-1959	
Direcció postal		X	X	X	X	X	X	X	X	X	X	X	X
Nom	X	X	X	X	X	X	X	X	X	X	X	X	X
Primer cognom	X	X	X	X	X	X	X	X	X	X	X	X	X
Segon cognom			X	X	X	X	X	X	X	X	X	X	X
Sexe									X	X	X	X	X
Edat		X	X	X	X	X	X	X	X	X	X	X	X
Data de naixement					X	X	X	X	X	X	X	X	X
Estat civil		X	X		X	X	X	X	X	X	X	X	X
Ocupació			X		X	X	X	X	X	X	X	X	X
Relació amb el cap de la llar					X	X	X	X	X	X	X	X	X
Sap llegir / escriure					X	X	X	X	X	X	X	X	X
Municipi i província d'origen					X	X	X	X	X	X	X	X	X
Municipi i província de residència					X	X	X	X	X	X	X	X	X
Nacionalitat									X	X	X	X	X
Temps de residència al municipi					X	X	X	X	X	X	X	X	X
Contribució territorial / Industrial					X	X	X	X					
Ingrés o salari									X		X	X	X
Classificació com habitant										X	X	X	X

Font: Elaboració pròpia

Metodologia.

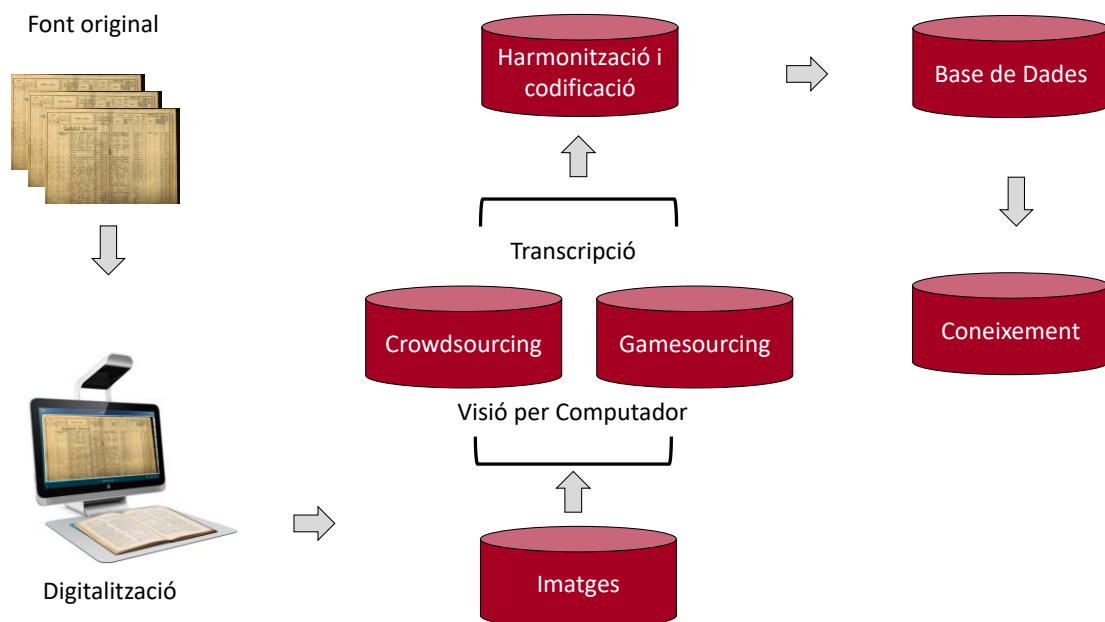
Aquest apartat està dividit en dos. En el primer s'especifiquen els mètodes i procediments per a la construcció de les bases de dades i en el segon s'exposen les metodologies estadístiques utilitzades per a l'anàlisi dels canvis socials.

Mètodes i procediments per la construcció de bases de dades

Els sistemes automàtics de reconeixement i extracció d'informació de documents històrics de població han experimentat progressos importants però la transcripció automàtica encara és un repte per assolir. Així doncs, les fases de creació de les bases de dades es poden estructurar segons l'esquema de la Figura 2.4. Una vegada localitzada la font original conservada a l'arxiu, s'inicia el procés de digitalització mitjançant la fotografia o escaneig, per passar del format paper al format digital. El següent pas és la creació d'un entorn web amigable per tal de portar a terme la transcripció. La confecció d'un *data entry tool* és indispensable pels posteriors processos de crowdsourcing i col·laboració ciutadana. En primer lloc cal un tractament digital de la imatge (color, tonalitat, grandària del fitxer per la seva posterior càrrega a l'aplicatiu web) i també la confecció del formulari *ad hoc*, per cadascuna de les fonts que es volen transcriure. Posteriorment hi ha un doble procés, per un costat hi ha la transcripció mitjançant *crowdsourcing* i un altre de gamificació. Ambdós processos estan recolzats per la visió

per computador i la intel·ligència artificial. Transcrites les dades per un o altre procediment i després de la seva correcció, es produeix la darrera fase abans d'obtenir la base de dades, que és la normativització, harmonització i codificació de totes les variables. D'aquesta manera s'obté una base de dades que respon als estàndards internacionals per a la seva comparació i utilització que permetrà obtenir nous avenços en el coneixement.

Figura 2.4. Fases de la creació de bases de dades.



Font: Elaboració pròpia.

Pel que fa a la transcripció automàtica de manuscrits, existeixen diferents procediments que ens poden ajudar a aconseguir-ho però que a dia d'avui encara requereixen de l'acció humana, sobretot en la seva fase final de validació. Podem dividir la transcripció automàtica en tres fases: a) imatge, b) anotació i c) semàntica. La primera fase correspon al tractament de la imatge. Malauradament, com a conseqüència dels estats de conservació de les fonts històriques o de la utilització de tintes i papers de poca qualitat, el primer pas a realitzar és una neteja de les imatges escanejades i deixar la imatge com més clara millor, alhora que també s'ha de realitzar un procés d'algoritmes per visualitzar l'estructura del text (Alvaro et al. 2015) i crear un esquelet de la imatge per separar

correctament cadascuna de les línies del text i els seus paràgrafs (Fernandéz-Mota et al. 2014).

A la segona fase s'apliquen les principals tècniques d'anàlisi d'extracció d'informació de documents manuscrits, que són: *Key Word Spotting* (KWS) i *Handwritten Text Recognition* (HTR). La KWS s'utilitza per buscar paraules o també per a la indexació dels continguts dels documents analitzats/tractats (Lladós et al. 2012). Consisteix bàsicament en retallar en imatges les parts de text que volem llegir i posteriorment cercar aquestes imatges o estructures (posició de píxels) en tot el document manuscrits, de tal manera que en poguem fer un indexació. La HTR és el tasca de processar imatges per obtenir una transcripció textual. La tècnica consisteix en anar identificant cadascuna de les paraules per després obtenir la transcripció de tota la frase sencera.

Finalment, la tercera fase és la del reconeixement semàntic. Aquests processos requereixen molt de temps d'entrenament i validació de dades per part dels algorismes. Un període de temps que s'anomena "ground truth" i consisteix en reconèixer les paraules i quina és la seva posició dins del text per comprendre el seu significat, que també s'utilitza en els procediments de KWS i HTR, que hem esmentat anteriorment. Actualment aquestes tecnologies es basen en els *Hidden Markov Models* (Rabiner, 1989) i els models de xarxes neuronals (Neural Networks) (Graves et al. 2009). D'aquesta manera, dins d'un text l'algoritme pot saber si està llegint el nom d'una persona, d'un carrer o una localitat. En aquesta tercera fase també hi ha els processos de *record linkage*, que permeten unir informacions de diferents imatges. D'aquesta manera un cop feta la transcripció poden reconèixer una mateixa persona en diferents documents i crear genealogies o línies de vida. En el cas dels padrons o censos de població, això ens permet poder seguir a un mateix individu al llarg del temps.

Pel que fa al procés de validació, la gamificació permet realitzar-ho a partir de resultats extrets de forma automàtica. De tal manera que els procés de validació és molt més àgil i entretingut de realitzar. L'interès i dedicació dels transcriptors sol disminuir al llarg del temps, d'aquí la idea de gamificar. Així, l'aplicació d'elements i principis de disseny de jocs en contextos que no són lúdics com en el cas del crowdsourcing, ha sorgit com una solució prometedora per mantenir l'interès dels usuaris alhora que ofereix una experiència atractiva (Morschheuser et al. 2016; Chen et al. 2018). D'aquesta manera, es pot ampliar molt més el públic col·laborador.

Processos de validació, correcció, normativització, harmonització i record linkage

Una base de dades nominativa només transcrita a format digital té poca utilitat per a la recerca. És necessari la realització de processos de correcció, normativització, harmonització i codificació de les dades, per tal d'obtenir una base de dades estandarditzada, que es puguin comparar al llarg del temps amb d'altres bases demogràfiques històriques. Així doncs, un cop obtinguda la transcripció literal– realitzada d'aquesta manera per tal de poder ensenyar als algoritmes a llegir documentació històrica mitjançant les tècniques de “Ground truth” – aquesta ha de passar per un procés de depuració de tal manera que es corregeixin aquelles errades en la introducció d'una informació en un camp que no tocava (per exemple que en el lloc de residència hi trobem transcrita una ocupació, o que l'estat civil ha estat transcrit en el camp corresponent a l'edat).

El següent pas és la normativització de la informació com a conseqüència de la inexistència d'una normativa lingüística de la llengua catalana fins a principis del segle XX, amb la publicació de la normativa ortogràfica i gramatical de Pompeu Fabra¹⁰, la disparitat en la forma d'escriure era molt elevada. Així, la gran varietat dialectal de la llengua catalana, juntament amb la influència d'altres llengües com el castellà o el francès i la pròpia transcripció fonètica per part dels escriptors (religiosos o civils) d'èpoques passades, fa que un mateix nom, cognom, ofici o població el trobem escrit de múltiples formes (Rubio Lizondo, 1997; Peytaví, 2010).

D'aquesta manera el procés d'harmonització de les diferents variables es pot agrupar en tres grans grups: harmonització nominal, codificació ocupacional i georeferenciació

Harmonització nominal: En les bases de dades nominatives, els noms i cognoms són bàsics a l'hora de poder enllaçar informacions de la mateixa persona al llarg del temps o en diferents bases de dades. D'aquesta manera, s'ha de realitzar un procés d'harmonització per no perdre aquells registres que contenen alguna variació ortogràfica i que complicarien el linkatge automàtic dels registres. En el cas dels noms, el santoral català i els diccionaris onomàstics ajuden a definir les variacions d'un mateix nom com per exemple: Josep: Josep, José, Pep. (Albaigès, 2004; Moreu-Rey, 1991). Pel que fa els

¹⁰ No serà fins la publicació de l'obra de Pompeu Fabra que es normativitzar la llengua catalana. Fabra, P. (1937 [1917]) (1983 [1932]) (1995 [1918]) (2013 [1913]).

cognoms cal recorre als treballs de Moll (1959), Alcover i Moll (1993), Coromines (1999) o Albaigès (2005) com en el cas de Ferrer: Farre, Ferre, Farrer,...

Codificació ocupacional: Un exemple molt clar de la necessitat de realitzar matrius de sistematització de la informació va sorgir en els primers intents de codificació d'ocupacions de fonts històriques (Treiman, 1977, 1980). Però no va ser fins 2002 quan apareix un sistema de codificació per classificar les ocupacions en el passat, l'anomenat HISCO (Historical Standard Classification of Occupations), creat pels historiadors M. Van Leeuwen, A. Miles i I. Maas, a partir dels catàlegs publicats per l'Oficina internacional del treball (ISCO'68)¹¹. Aquesta mateixa metodologia s'ha aplicat per la creació d'una classificació social HISCLASS (Historical International Social Class Scheme) (van Leeuwen & Maas, 2005). Es tracta d'una classificació que converteix els codis ocupacionals derivats de HISCO en classes socials, basant-se en elements com el grau d'especialització, la jerarquia ocupacional o si es tractaven de treballs manuals o no (Pujadas et al. 2014).

Georeferenciació: Per tal de tenir una localització geogràfica i així poder mapificar i observar les dades espacials, es necessària l'harmonització geogràfica. Realitzarem aquesta codificació segons les bases cartogràfiques que disposem, generalment de municipis. Com més petita sigui l'escala territorial que obtinguem millor, ja que d'aquesta forma sempre podem agregar unitats més grans.

Totes aquestes normativitzacions i harmonitzacions són importants no tant sols per mantenir una bona consolidació de la base de dades, sinó que són crucials per la construcció de bases de dades longitudinals o per vincular un mateix individu que apareix en diferents fonts mitjançant el *record linkage*. Aquesta informació longitudinal proporciona un ampli ventall de possibilitats d'investigació i obertura de noves línies d'investigació als historiadors i als demògrafs històrics.

Metodologia estadística per l'anàlisi dels canvis socials

Per a l'elaboració d'aquesta tesis doctoral hem utilitzat dos tipus de metodologia. Una primera centrada en la creació de bases de dades longitudinals, i una segona basada en l'anàlisi descriptiu i multivariable. Malgrat que dins de cadascun dels quatre assajos de la

¹¹ Per més informació veure Van Leeuwen et al. (2002), i pel cas català, Pujadas-Mora, et al. 2014.

tesí hi ha un apartat dedicat a la metodologia específica utilitzada en cada moment, voldríem fer esment d'alguns dels aspectes metodològics més novedosos.

Per a la creació de les bases de dades longitudinals s'ha utilitzat un software propi anomenat "Buscadescendències", dissenyat dins del projecte Five Centuries of Marriage, el qual mitjançant dos algoritmes (string distances) vincula els individus en dos registres diferents. El software proporciona un valor de coincidència de les lletres del nom o cognom dels individus que va de 0 a 1. A més a més, aquest software compte amb tot un paquet lingüístic per esmenar possibles similituds o diferenciacions escrites d'un mateix cognom, per tal d'afavorir o penalitzar combinacions de lletres o de noms sencers. Un exemple és la similitud entre els cognoms Ferrer i Farre, en el que el software ens donaria un valor més favorable que el que tindria només pel fet de comparar lletres o bé la diferenciació entre els cognoms Pons i Pous, que malgrat tenir lletres molt semblants, corresponen a cognoms diferents. D'aquesta manera el software del Buscadescendència té en compte les variacions més comunes dels cognoms catalans com ara la supressió de la lletra "r" al final de paraula, (Bover o Bove), així com també els canvis de la lletra "a" per una "e" quan són àtones (Cases o Casas). O bé la confusió de lletres entre una "b" i una "v" (Rebull o Revull); o entre una "y" i una "i" (Isern o Ysern); l'ús de la lletra "h" a principi o a final de paraula (Homs o Oms; Bosch o Bosc), entre d'altres. Aquestes vinculacions nominatives també estan subjectes a limitacions temporals i a un procés de validació probabilístic per corregir falsos positius, mitjançant la comparativa amb d'altres variables com ara municipi d'origen o ocupació. Així, si una persona té els mateixos noms i cognoms, però sabem que els seu municipi d'origen es diferent en un dels dos registres, el vincle no es realitza. De la mateix manera, si entre els dos registres es produeix un canvi d'ocupació poc probable, com ara passar de pastor a advocat, aquest vincle tampoc es realitza. A més a més, en l'assaig número 2 es proposa un nou mètode de record linkage basat en el Knowledge Graph, que utilitza matrius de ponderació per tal d'optimitzar la vinculació de registres. De tal manera, que amb l'ajuda de la Intel·ligència Artificial, el software tingui la capacitat d'aprendre i ser cada vegada més precís en els seus resultats.

Pel que fa a l'anàlisi quantitatiu per a l'explotació de resultats de les bases de dades creades, s'ha combinat l'anàlisi descriptiu amb una perspectiva multivariable. D'aquesta manera en el tercer assaig "*A tals pares, tals fills: Transmissió intergeneracional a Sabadell, segles XVI – XVII*" s'han fet servir tècniques estadístiques bivariants per quantificar els nivells de mobilitat entre generacions, utilitzant taules de doble entrada per

comparar l'ocupació dels pares amb els fills. I després s'han estimat els residus estandarditzats d'aquestes distribucions per analitzar si la coincidència entre els grups de pares i fills està sobrerrepresentat o si pel contrari és poc freqüent en relació a l'estructura social de la població analitzada.

Pel que fa al quart assaig: “*What counted most for secularization: Politics, economy or family? Shifting seasonality of marriages in the Barcelona area, 1715-1880*”, s’ha tingut la voluntat d'utilitzar anàlisis estadístics més innovadors en demografia històrica. Tal és així, que s’ha calculat per un costat l’índex clàssic d’estacionalitat dels matrimonis (Henry, 1980) per obtenir una dada comparable durant tots els mesos d’un any, però alhora també s’ha realitzat el càlcul d’un segon índex d’estacionalitat dels matrimonis pels dies concrets que dura la Quaresma. Tenint en compte les variacions anuals d’aquesta litúrgia de 40 dies, que es basa en el calendari de la lluna, i que s’inicia com a molt d’hora el dia 4 de febrer i no pot finalitzar mes tard del dia 10 de març. D’aquesta manera, a partir d’algoritmes que calculen el dia exacte d’inici i fi de Quaresma per cada any, s’ha obtingut el nombre exacte de matrimonis celebrats durant aquest període litúrgic que ens ha permès precisar millor el nostre anàlisis sobre el procés de secularització. En aquest assaig també s’han realitzat anàlisis multivariables a partir de models de regressió logística per tal d’observar la influència del període polític, el grup social i l’estat civil sobre el canvi del calendari nupcial com a indicadors del procés de secularització de la població de l’àrea de Barcelona. Amb aquests models s’ha determinat si les probabilitats de casar-se durant la Quaresma augmenten, així com també les diferents interaccions entre elles.

3 Assaig 1: The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth – twentieth centuries).

3.1 Introduction

Historical studies always need empirical evidence, and the enhancement of historical research in the «Big Data Revolution» (Ruggles, 2014), inspired additional need for large quantities of high-quality data. Historical Demography has always consistently used considerable amounts of data, depending on the technical developments of the moment (Billari and Zagheni, 2017). Many historical demographic databases have a long tradition, such as the Historical Samples of the Netherlands (HSN), the China Multigenerational Panel Dataset (CMGPD-LN); the North Atlantic Population Project (NAPP), the Demographic Database at Umeå University (CEDAR), the Research Program in Historical Demography at Montreal University (PRDH), the Scanian Economic Demographic Database at Lund University (SEDD), among others (Cf Ruggles et al., 2011; Lee and Campbell, 2010; Mandemakers, 2002; Edvinsson, 2000; Dillon et al., 2018; Bengtsson et al. 2012.).¹² These kind of data is particularly important to understand the demographic past (fertility, family formation, health, social stratification, social inequality, etc.) and can help to understand the present and to forecast the future. However, the digitization of historical sources is still time-consuming with high staff costs. Therefore, the recent conjunction of Historical Demography and the

¹² For more information see of each of those databases: The Historical Samples of the Netherlands (HSN) (<https://socialhistory.org/en/hsn/index>); China Multigenerational Panel Dataset (CMGPD-LN) (<https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/27063>); North Atlantic Population Project (NAPP) (<https://www.nappdata.org/napp/>); The Demographic Database at Umeå University (DDBPOPUM) (<http://www.cedar.umu.se/english/ddb/databases/popum/>); The research Program in Historical Demography at Montreal University (PRDH) (<https://www.prdh-igd.com/>); The Scanian Economic Demographic Database at Lund University (SEDD) (<https://www.ed.lu.se/databases/sedd>). This list can be extended with some novel impressive databases as the Historical Population Register (HPR) from the Norwegian Historical Data Centre (<http://www.rhd.uit.no/nhdc/hpr.html>) or the ones which are created for the International Demographic Unit at the Ural Federal University (Russia) (<https://urfu.ru/en/research/international-researchcollaboration/international-research-laboratories/international-demographic-unit/>). For more information on historical longitudinal databases please see: <https://www.ehps-net.eu/databases>.

Computer Sciences promises to shorten the construction time for historical individual level databases and allows the building of bigger and more informative databases (Pujadas-Mora et al., 2016; Hall et al. 2000).

The *Baix Llobregat (BALL) Demographic Database* is an ongoing database project containing individual census data from the Catalan region of *Baix Llobregat* (Spain) during the nineteenth and twentieth centuries. The *BALL* Database is built within the project '*NETWORKS: Technology and citizen innovation for building historical social networks to understand the demographic past*' directed by Alicia Fornés from the Center for Computer Vision and Joana Maria Pujadas-Mora from the Center for Demographic Studies, both at the Universitat Autònoma de Barcelona, funded by the Recercaixa program (2017-2019). Its webpage is <http://dag.cvc.uab.es/xarxes/>. The aim of the project is to develop technologies facilitating massive digitalization of demographic sources, and more specifically the *padrones* (local censuses), in order to reconstruct historical 'social' networks employing computer vision technology. Such virtual networks can be created thanks to the linkage of nominative records compiled in the local censuses across time and space. Thus, digitized versions of individual and family lifespans are established, and individuals and families can be located spatially.

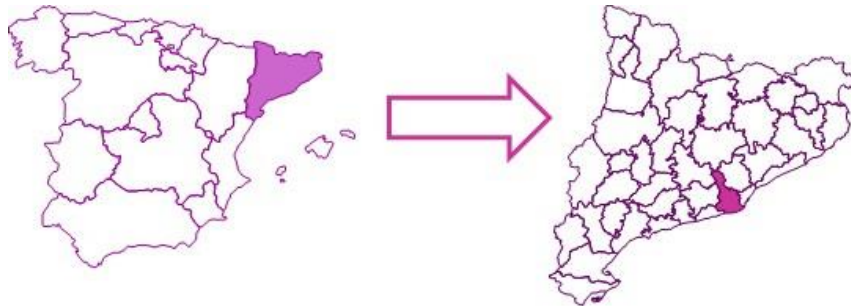
3.2 Historical and Geographic context of the Baix Llobregat

Catalonia was one of the first places in Southern Europe to industrialize (Brea-Martínez and Pujadas-Mora, 2018; Martínez-Galarraga and Prat, 2016). The Baix Llobregat region played an important role because since the second half of the nineteenth century (Figure 3.1), the flourishing Barcelonese cotton industry moved towards the lower parts of the Llobregat River and its delta in search of water for the demanding production (Nadal, 1992). Then, the region changed from an agricultural to a wider occupational and social structure.

The industrial growth first featured textile production and afterwards metallurgy (Carbonell i Porro, 1995). However, in the last decades of the nineteenth century the agricultural crisis in Europe affected prices and incomes. In Catalonia the *phylloxera* vineyard infection worsened the crisis (Garrabou et al., 1991). Although the agrarian crisis also disturbed the economy of the *Llobregat* riverside, its effect was limited, because the region disposed of highly fertile soils and had faced an incipient commercialization of its agricultural production since the eighteenth century. During the

nineteenth century, the city of Barcelona used to be provided with fruits and other orchard products from the *Llobregat* and since the start of the twentieth century they exported agricultural commodities abroad (Tribó, 1989).

Figure 3.1. Map of Spain, Catalonia and the region of *Baix Llobregat*.



Source: Author's own elaboration (BALL Demographic Database).

Thus, the *Baix Llobregat* region represents a highly interesting case study. On the one hand, the region played an important territory in Catalan industrialization so that the BALL Database includes observations of a society facing all stages in the industrial revolution: the end of the preindustrial era and the take-off, implementation, expansion and consolidation of industrialization. On the other hand, the successful agrarian commercialization in the *Baix Llobregat* since the eighteenth century adds extra elements to the classical model of modern economic and industrial growth, which implies an increasing secondary sector and a declining primary sector (Kuznets and Murphy, 1966; Nadal, 1975; Carreras, 1990; Martínez-Galarraga and Prat, 2016; Brea-Martínez and Pujadas-Mora, 2018).

The BALL Database, therefore, offers analytical opportunities regarding the relation between industrialization and demography. For instance, a classic theory argues that declining fertility was caused by modernization and industrialization and the emergence of nuclear families (Franck and Galor, 2015; Freedman, 1979; Cherlin, 2001). However, our area of study experienced the earlier fertility decline within Catalonia - together with France the earliest fertility declines in Western Europe (Cabré, 1999; Weir, 1993; Coale and Watkins, 1986). Moreover, the *Baix Llobregat* region was traditionally featured by the strong presence of stem families. Nowadays, this area is still shaped by strong family ties (Reher, 2004; Fauve-Chamoux, 2009; Borderías and Ferrer, 2017; Esping-Andersen, 1999). Thus, the BALL Database provides an interesting and complete socioeconomic

and socialhistorical “lab” for studying the demographic, familial and individual responses to a transforming world.

Sources: The *Padrones* (Local censuses)

Local censuses (*Padrones* in Spanish) were taken regularly in Spain since the nineteenth century. They were a result of administrative centralization and the efforts of the liberal state to increase population and wealth (Porter, 1995; Wolf, 1989). Institutionalization of state statistics was a response to a desire for quantifiable material within a framework of epistemological development of scientific objectiveness and impersonal knowledge (Porter, 1995). This was a common although not simultaneous process in nineteenth century Europe. Spanish local censuses were compulsory after 1823 (García Pérez, 2007; Reher and Valero-Lobo, 1995). Thus, they were carried out before the first modern national census (1857) or the definitive implementation of the civil register (1871).

Table 3.1. Number and time range of local censuses in BALL Demographic Database, by municipality.

Municipality	Number of Local Censuses	Period	Population 1857	Population 1950	Individual records
Begues	10	1854 / 1928	799	968	9.658
Castellví de Rosanes	7	1857 / 1950	323	268	2.100
Collbató	32	1852 / 1950	865	416	12.836
Corbera de Llobregat	18	1857 / 1950	885	1.397	15.124
El Papiol	14	1875 / 1950	1.100	1.159	14.256
Molins de Rei	14	1852 / 1955	3.002	8.024	55.566
Sant Feliu de Llobregat	20	1828 / 1955	2.484	7.327	83.528
Santa Coloma de Cervelló	7	1900 / 1950	211	1.227	7.537
Torrelles de Llobregat	30	1842 / 1950	496	732	21.062
Total	152	1828 / 1950	10.165	21.518	221.667

Source: Authors' own elaboration (BALL Demographic Database, version July 2018).

The state progressively implemented local censuses throughout the country. Several decrees informed municipal officials of the obligation to take them. Thus, the Royal Decree of March 14th 1857 demanded all the registers to be nominative and simultaneous. Only after the enactment of the Municipal Law of August 20th 1870 was the taking of local censuses fixed to 5-year intervals. Finally, in the Municipal Statute of March 8th 1924, the formats of local censuses were standardized with respect to the type

of variables that should be recorded (García Pérez, 2007; García Ruipérez, 2012). In this way, local censuses showed the sociodemographic features of each inhabitant in a particular household in an urban center or in the countryside. For each person, the register included first names and surnames, age or birth date, civil status and occupation and the family or working relationship with the household head. In some periods, the information is also available about the individuals' literacy and income. Most importantly, the local censuses contain the only data preserved at the individual level in Spain since national census manuscripts used to be destroyed once the population number was estimated and the main variables aggregated.

The BALL Demographic Database until now contains nine municipalities (Figure 3.2): *Begues, Castellví de Rosanes, Collbató, Corbera de Llobregat, El Papiol, Molins de Rei, Sant Feliu de Llobregat, Santa Coloma de Cervelló* and *Torrelles de Llobregat*, altogether representing a total number of 152 local censuses comprising the period 1828 to 1950 (Table 1). The medium and small size of the communities registered in the BALL local censuses is far from an inconvenience for the study of networks. It is in reality a strong feature, making it easier to assess the mechanisms that individuals and households used at the micro level in order to adapt to and interact with socioeconomic changes at the macro level (Coleman, 1986).

Figure 3.2. Map of BALL Demographic Database (*Baix Llobregat* county).



The pre-1880 local censuses were taken before the standardization of census formats. They are basically nominative lists of the inhabitants gathered in the census (Figure 3.3). For this reason, the number of variables registered usually includes only first names, surnames, ages and civil statuses.

Figure 3.3. Local Censuses previous to 1880.



Collbató, 1842

Sant Feliu de Llobregat, 1857

Castellví de Rosanes, 1866

Source: Arxiu Comarcal del Baix Llobregat (ACBLL).

From 1881 the number of variables registered in local censuses increased to around 15 in each census (Figure 3.4). In this way, the recorded variables were first names, two surnames, occupations, civil status, age and/or birthdate, address (street and house number), birthplace as well as time of residence. From 1906 onwards, other variables are the relationship to the household head, literacy (reading and/or writing), and fiscal contributions (see Annex 1). Nevertheless, in spite of the rich detail in these sources, the local censuses masculinize the data, leaving out many women's occupations. Female work was usually underrecorded or registered with labels such as *su sexo* (tasks of her sex/gender) or *sus labores* (her own tasks) (Borderias, 2012). The only exception to this female invisibility in the local censuses is found in those from 1936 during the Spanish II Republic (1931 - 1939), which mostly registered women's occupations as well as which factories or enterprises employed the wage earners.

Figure 3.4. Local Censuses from 1880.

The figure displays three historical census tables. The first table, from 1887, is a dense grid of handwritten entries. The second table, from 1899, is titled 'PADRON DE HABITANTES' and is organized into columns for 'Calle N.' and 'Sexo'. The third table, from 1910, is also a grid of handwritten entries, similar in format to the 1887 table.

Torrelles de Llobregat, 1887

Molins de Rei, 1899

Sant Feliu de Llobregat, 1910

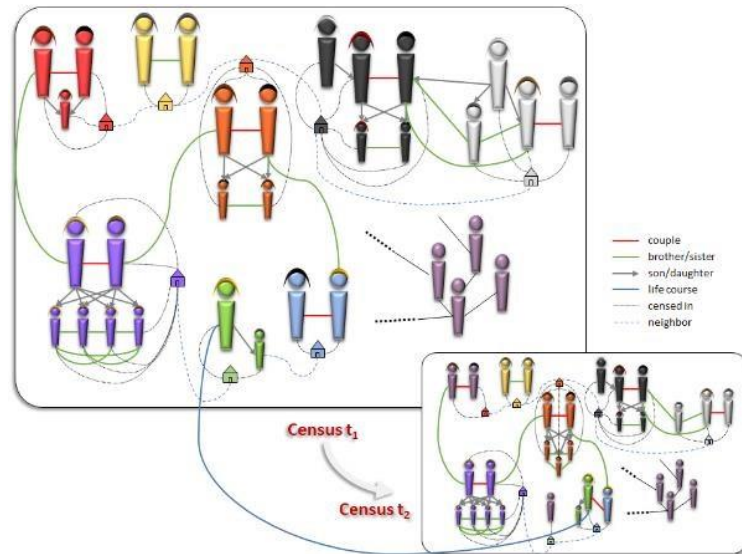
Source: Arxiu Comarcal del Baix Llobregat (ACBLL).

3.3 Type of database

As mentioned above, the aim of the project is to create historical ‘social’ networks using census materia. Therefore, we use a social network model to transcribe the nominative census records. This data model also allows using the powerful techniques of graph based data analytics for querying the database. This representation is scalable, allowing not only to integrate and link data from other censuses but also other demographic sources like birth, marriage, death records. Figure 5 illustrates how the census data is structured according to a social network or graph representation, with the time dimension as a stack of linked graphs. Each individual graph is a static representation of the population at time t (a particular census) where the nodes are observations of people. The corresponding households where they were registered and the graph edges represent relations (genealogic or other relevant affinities deduced from the source documents like occupations, household neighbours etc). In dynamic, time varying graphs, the life courses of individuals are constructed by linking the corresponding record of their observation at time t_n with the observation at time t_{n+1} .

Technically, we implemented this in a relational database according to the data model in figure 3.5. Current database functions allow us to implement graph structures in a native format, while providing query methods to analyse the graphs (community detection, record linkage, centrality-based node detection etc). As future work, we plan to migrate to such a format. The proposed design is inspired by the *Intermediate Data Structure* (IDS), a standard proposed by the EHPS Network (Alter & Mandemakers, 2014). This should ensure the interoperability among data sets, and be the basis for the implementation of data analytics.

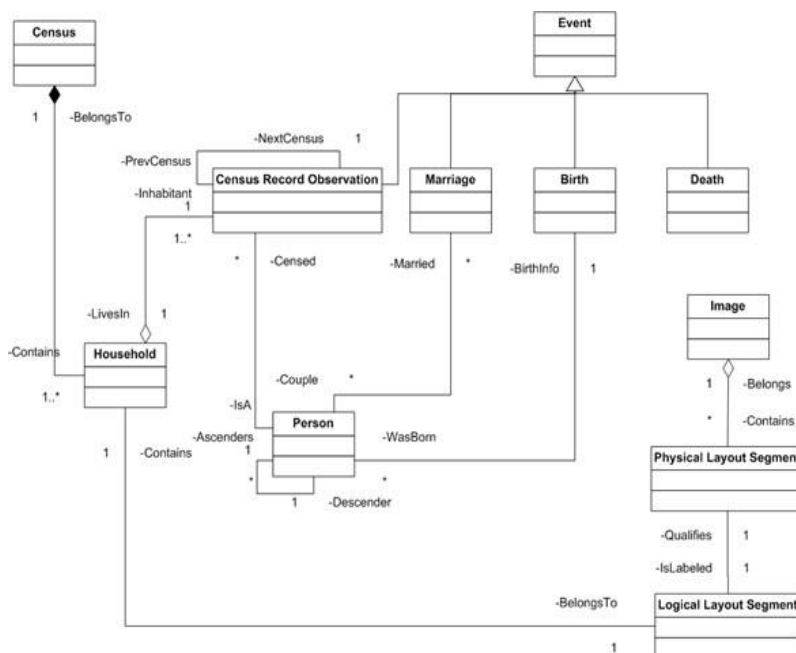
Figure 3.5. Social network view of census data.



Source: Authors' own elaboration

The design proposed in Figure 3.6 has the central entity *Person*, containing the information that is permanent over time. Each person has different *Observations*, belonging to the corresponding *censuses* that have been analyzed.

Figure 3.6. Data Model of the database.



Source: Authors' own elaboration (in unified modeling language)

Like in the census, observations are grouped in *Households*. A census record is considered an *Event*, which is a class with different subclasses (according to the different event types: census record, marriage record, birth record or death record). This structure allows future users to reconstruct the life course of a person based on the extraction of his/her events.

This design is flexible enough to incorporate other event types, not necessarily related to demographic data, but relevant in the life course. Finally, the database architecture contains entities for storing the images from which the information is extracted and the image segments that contain the relevant information as cropped by the computer vision algorithms.

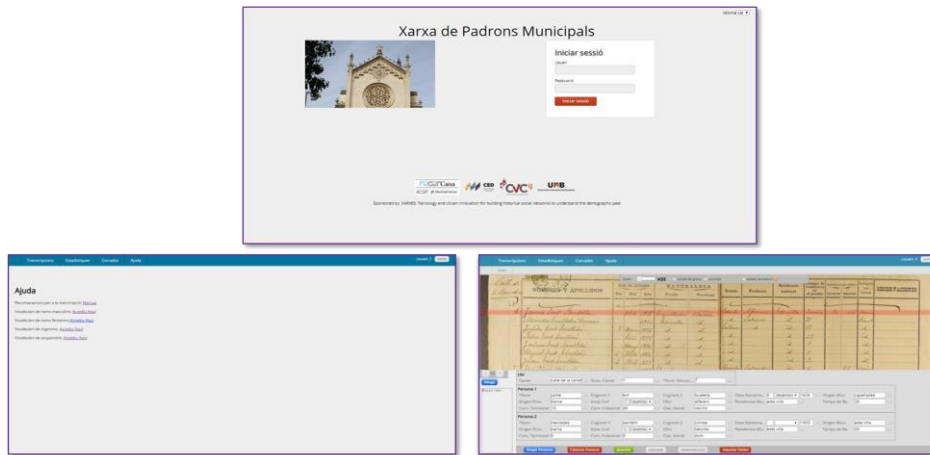
The transcribed data are transferred into a digital format compatible with different formats (Excel, SPSS, STATA, R, etc), downloadable from a crowdsourcing platform used also for collaborative transcription of the original sources.

3.4 Transcription: Crowdsourcing platform, videogame, computer vision & volunteers

The transcription of the manuscripts can be performed either manually or through computer vision techniques. Given that paper degradation and the high variability in handwriting styles in historical manuscripts impose many difficulties, the existing handwriting recognition techniques still need more development before we shall trust automatic transcription completely. For this reason, the local censuses have been transcribed by a community using a web-based crowdsourcing platform, and validated through a gamesourcing mobile application, both assisted with computer vision techniques.

Since manual transcription is tedious and time-consuming, we have developed a crowdsourcing platform for data entry (Figure 3.7). The idea of crowdsourcing is to split the work into many micro-tasks and ask contributions from a large group of people, especially from the online community. Thus, the task is shared among many users and finished in less time. The graphical user interface of our web-based platform offers a data entry tool with a user friendly environment, integrating both the original source and the data entry form in the same view.

Figure 3.7. Crowdsourcing platform: <http://dagapp.cvc.uab.es/PadronsXarxes/>



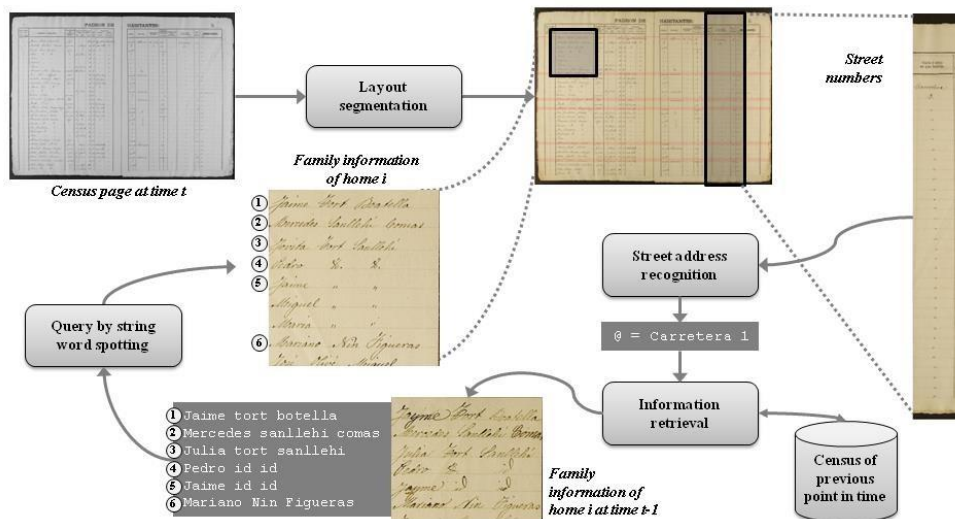
Source: Author's own elaboration

We also use the transcribed data to teach the computer vision algorithms how to interpret further manuscripts. Concretely, our handwriting recognizer, based on deep learning, uses document images with their corresponding transcriptions to improve the transcription system and to adapt itself to every new handwriting style. As a result, these algorithms can be used to speed up the transcription in two different scenarios: first, to assist the transcription via information transfer, and second to perform an automatic transcription with manual validation.

In the first scenario, we use the redundancy in censuses to automatically transfer repeated information from one census to the next. As was stated before, local censuses in Spain were recorded in intervals of a few years and the nominative information about individuals recorded within households, was quite stable. This redundancy is used to assist the transcriber and speed up the transcription. Once one census has been manually transcribed, the redundant information (names, surnames and addresses) is transferred to the next one, so it is only necessary to update the changes manually: adding new members or deleting those who left or died in each household. For this purpose, household records from consecutive censuses are automatically aligned using the street address. Then, the individuals are automatically located in the picture of the census manuscript using visual word search, «word spotting». Concretely, the individual names and surnames from a census are searched in the corresponding home's records in the next census (figure 3.8). Since the process is based on a focused search, the accuracy is high. This redundancy has

been used to assist the transcription of the 1886 census, once the 1881 census was transcribed (Mas et al., 2016), which led to a 70% reduction in the transcription time.

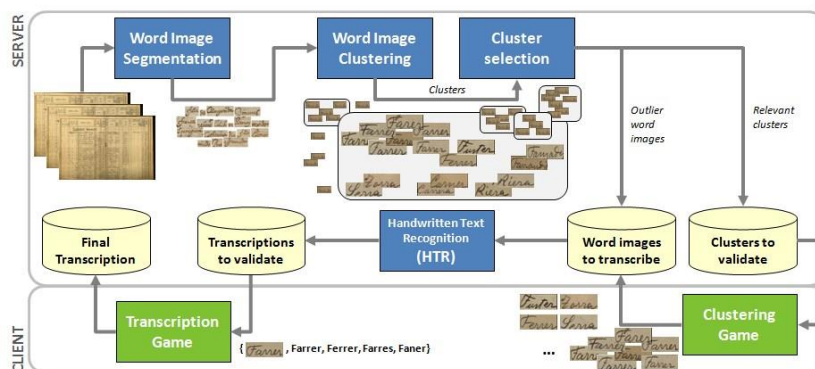
Figure 3.8. Architecture of the proposed system in first scenario.



Source: Mas et al., 2016

In the second scenario, the transcription is performed automatically, followed by manual validation. First, snippets corresponding to word images are segmented from the manuscripts. Then, these word images are clustered to find high frequency words that can be jointly transcribed using a small percentage of representative instances. Afterwards, the software sends the words in each cluster to the deep learning based transcription system, which provides the most plausible transcriptions for each word image (Figure 3.9).

Figure 3.9. Architecture of the proposed system in second scenario.



Source: Chen et al., 2018

Finally, clusters that contain words with the same transcription can be transcribed at once. In this way, we avoid the validation of every single word, speeding up the transcription while maintaining high quality performance.

The manual validation has been done through gamesourcing, understood as crowdsourcing via gamification which consists in the application of game-design elements and principles in a non-game context. Concretely, two simple Android games have been developed (Chen et al., 2018). The first game is designed to validate the word clustering algorithm. It shows some instances of a given cluster, and asks the user to confirm that those words are the same, thus belonging to the cluster. The second game validates the output of the transcription algorithm. When the player selects one word, the system shows the most probable transcriptions according to the algorithm, and the user selects the correct answer among these possibilities (Figure 3.10). Experiments demonstrate that the transcription effort can be significantly reduced, and that user engagement is higher than with the traditional crowdsourcing web-based application.

Figure 3.10: Images of the two games.



Source: Chen et al., 2018

Thanks to the coordination and collaboration with the Regional Archive of the *Baix Llobregat*, local study groups, town halls and local archives, it was possible to gather groups of transcribers, volunteering in each municipality (figure 3.11). Simpler collaboration of volunteers in gathering data was already used in the “Cambridge Group for the History of Population and Social Structure” (Wrigley & Schofield, 1989). Until July 2018, 119 transcribers participated (58 females and 61 males), whose mean age is

around 60 (Table 3.2). The transcribers had diverse cultural backgrounds, but enthusiasm about local history and genealogy in common.

Figure 3.11. Pictures of some groups of transcribers.



Torrelles de Llobregat



Corbera de Llobregat

Source: Authors' own elaboration.

In order to enhance transcription quality and continuously engage the inhabitants, we have performed several informative sessions for presenting the NETWORKS project and how the online application of transcription works as well as raising people's awareness about the local censuses and their importance. These informative sessions succeed to engage and train the volunteers, who afterwards received authorization for conducting online transcription.

Table 3.2. Number of volunteer transcribers by municipality.

Municipality	Men	Women	Total
Begues	3	4	7
Castellví de Rosanes	3	6	9
Collbató	3	2	5
Corbera de Llobregat	4	1	5
El Papiol	4	4	8
Molins de Rei	4	4	8
Sant Feliu de Llobregat	27	31	58
Santa Coloma de Cervelló	8	3	11
Torrelles de Llobregat	5	3	8
Total	61	58	119

Source: Authors' own elaboration.

In summary, these technological developments allow citizens and archivists to participate in the extraction of demographic information through web-based crowdsourcing

platforms and gamesourcing applications, which incorporate handwriting recognition algorithms to assist the transcribers. This collaboration can be understood as a way of popularizing science for the public and helps to instill critical thinking. We shall develop new user experiences in the near future, like geoprojections on interactive maps . Thus, the project facilitates the consumption and dissemination of the historical knowledge in an illustrative and pedagogical way.

3.5 Quality checks

Having online applications for the purpose of transcription enhances supervision and quality control in real time, as well as interplay with the transcribers. In addition, functions for correction are also implemented, as are periodical controls in order to ensure better transcription:

Pre-transcription:

Recommendations for transcription were delivered to all transcribers in face-to-face training sessions, in order to ensure a common and systematic approach to transcription.

Control during the transcription:

The online application marks compulsory fields with an asterisk for the most important variables, blocking the submission of an incomplete transcription (figure 3.12). Additionally, in order to avoid leaving out any household member, the total number of household members has to be declared when beginning transcription of every new household. The application includes a “Help” menu with different tools such as dictionaries, lists of names as well as geographic and occupational guides to help the transcribers when in doubt. Moreover, all the transcribers can continuously address their queries via email or social networks to enhance the communication between researchers, technicians and transcribers. Every transcriber had a reviewer in charge of verifying transcription quality and solving doubtful cases. Finally, we released statistics about the transcription and database progress to the community of transcribers in order to dynamise the transcription process.

Figure 3.12. Control during the transcription.



Source: Authors' own elaboration.

Post-transcription control:

Once the transcription of a given local census was finished, several analyses are conducted regarding the frequencies of different variable values for identifying likely inconsistencies.

3.6 Harmonization and codification of data

One central characteristic of the BALL Demographic Database is that instead of transcribing data in a harmonized way, the volunteers have been told to transcribe all records literally in order to avoid inconsistent interpretations of the nominative data. In this way, once the sources have been transcribed it is necessary to apply linguistic harmonization. The same individual appears with different names and surnames, or apparent variants because they were recorded with different spelling, or abbreviations. One combination of a common first name and surname might refer to several individuals (Goiser and Christen 2006; Herzog et al., 2007, Schürer, 2007). These questions need particular attention in the case of the Catalan language, which was not standardized until 1913. Also in occupation titles, locations, relationships with the household head or even civil status there is vast variability. Apart from reducing such variability, harmonization is important to make variables and outcomes more accessible for analysis and comparisons at the national or international level. For this reason, we recode the variables in each local census in order to remove phonetic variations of names and other variables

caused by the different dialect and influences from other languages which favored many written variations (Peytaví 2010; Rubió and Lizondo 1997). For instance, we found the surname *Ferrer* (Smith) written as *Ferré*, *Farré*, *Farrer*, and so on. By contrast, Spanish sources do not present the same orthographical variability because standardization dates back to the 15th century (Herzog, 2007). Foreign surnames were adapted to the Catalan or Spanish languages.

Linguistic harmonization helps overcome these obstacles and clusters anthroponyms in order to compile dictionaries of names and surnames (Bloothoof, 1998; Christen, 2012). Thus, nominative data have been harmonized according to language criteria to facilitate the record linkage that identifies the same individual in different censuses (Jordà, 2016; Jordà et al., 2013). In addition, places have been geolocated, and occupations have been coded using the Historical International Classification of Occupations (HISCO) (van Leeuwen et al., 2002).

Once all occupations were codified, we ranked and classified them according to sociooccupational position by means of HISCLASS (van Leeuwen and Maas, 2011) and HISCAM (Lambert et al., 2013), respectively. HISCLASS differentiates individuals in consonance with the social group to which they belonged according to dimensions like manual/non-manual division, skill level, degree of supervision and economic sector, which gives 12 different classes going from unskilled rural workers at the bottom to higher managers and professionals at the top. HISCAM is a different occupational stratification scale based on the Cambridge Social Interaction and Stratification scheme, using marriage data from Belgium, Britain, Canada, France, Germany, the Netherlands, and Sweden (Prandy, 2000). The main idea behind this scaling is that individuals who interact more (in terms of occupational and social relations) are closer in terms of social position, assuming that these interactions represent the occupational stratification structure. The result is a ranking of occupations (theoretically from 0 to 99), showing not only similar social standing but also differences between occupations. Finally, we have grouped each occupation into economic sectors, following an adaptation of HISCO for the Catalan historical labor market (Pujadas-Mora et al., 2014).

The BALL Database, in July 2018, is composed of 9 municipalities (figure 2) and 152 local censuses that correspond to a total of 221,667 individual observations (see table 1). The harmonization of its names, occupations and places uses the reference tables created during the harmonization of the Barcelona Historical Marriage Database (Jordà, 2016).

In this way, we rely on a reference table for the names (female and male names) with 27,434 unique records, the surnames with 101,238 unique records, the occupations with 24,399 unique records and the place names with 47,556 unique records, which facilitate the work of linguistic normalization and harmonization.

3.7 Towards longitudinal data: The *Sant Feliu* life course database

Among potential derivations from the BALL Demographic Database is the possibility to create longitudinal databases of individuals and households by linking local censuses. For this, we have used one of the most populated towns in *Baix Llobregat*, one with a large time coverage, namely *Sant Feliu de Llobregat*, to create a sub-database. Thus, the longitudinal database of *Sant Feliu* is based on reconstruction of individual life-courses using local censuses. The town of *Sant Feliu de Llobregat* was one of the most important in the region, in economic and administrative terms, being the judicial district capital, with the arrival of new economic activities such as textile and metallurgical industries since the second half of the nineteenth century and the railway station in 1855. The database contains the information in all the 15 censuses recorded in *Sant Feliu* from 1828 to 1940. This information has benefited from computer-assisted data transcription through crowdsourcing in which 58 volunteers have collaborated (27 men and 31 women) for a period of 2 years. As explained above, the local censuses in Spain (including *Sant Feliu*) were taken at short intervals.

The Longitudinal Database of *Sant Feliu* contains 59,084 observations of individuals, which increased chronologically, mainly from the 1920 onwards when *Sant Feliu de Llobregat* started having important immigration. The individuals are distributed in 12,748 households across the entire period with a mean number of 4.6 individuals per household, levels that decreased over time from 5.2 persons per household in 1828 to 4 in 1940 (table 3.3). The nominative data has been harmonized, addresses have been geolocated, and occupations have been encoded.

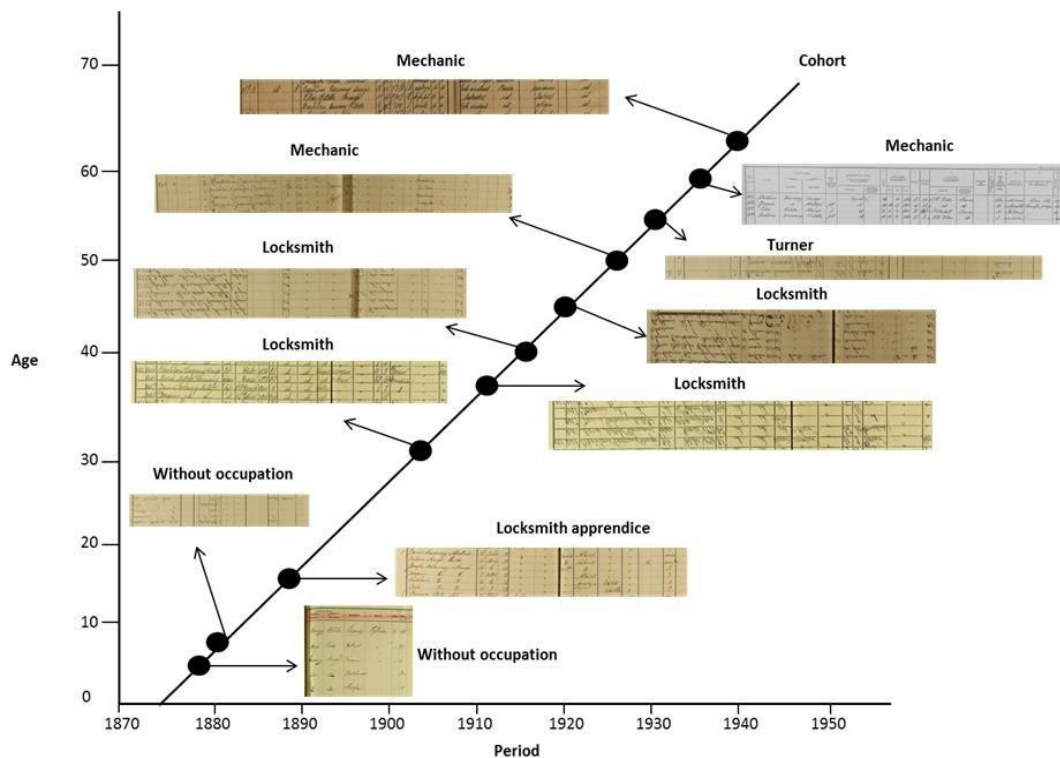
Table 3.3. Individuals and households by census year in Sant Feliu de Llobregat 1828-1940

Years	Individuals	Households	Individuals per Household
1828	2.209	426	5,2
1833	1.470	313	4,7
1839	1.946	377	5,2
1857	2.472	533	4,6
1878	2.762	610	4,5
1881	3.005	598	5,0
1889	3.118	644	4,8
1906	3.606	805	4,5
1910	3.809	866	4,4
1915	4.330	936	4,6
1920	4.353	918	4,7
1924	5.575	1081	5,2
1930	6.392	1459	4,4
1936	7.023	1458	4,8
1940	6.727	1675	4,0
Total	59.084	12.748	4,6

Source: Authors' own elaboration.

Harmonization and codification significantly reduced the number of spelling variations, erroneous or misleading data. From the initial 1,976 given names transcribed literally, the harmonization dropped the number to 1,033 variants. Since Spanish individuals usually receive two surnames (paternal and maternal), the surname harmonization was multiplied by two. The 4,269 different paternal and 5,641 maternal surnames, became 2,767 and 3,312 after harmonization, respectively. Finally at the geographic level, the 3,743 different addresses observed were harmonized to 1,552 different geographic points. Nevertheless, as the database includes local censuses spread over a wide period, not all contain the same information. For instance, local censuses in *Sant Feliu* prior to 1906 did not have information on the specificity of relationship with the household head, i.e. who were spouses, children and/or other relatives. We have used other variables such as surnames, civil status and age in order to reconstruct family relationships within the same household. Finally, the procedure of harmonizing and coding the occupations resulted in 1,611 different occupational titles that were simplified into 292 HISCO codes.

Figure 3.13. Life course example. Bartomeu Carcereny Amigo (1874-1940).



Source: Authors' own elaboration.

After standardization, nominative record linkage was performed. The procedure used probabilistic criteria with algorithms of two string distances, the Bag distance and the Levenshtein distance, together with a language model that favoured similarities in pronunciation and penalized typographic errors (Villavicencio, Jordà, and Pujadas-Mora, 2015). This allowed us to follow 10,405 individuals at least in two different censuses. In the best cases, we reconstructed up to 12 different observations for individuals, which signifies assessing the entire life-cycle of a person. For instance, consider the lifecourse of Bartomeu Carcereny Amigó (figure 3.13). This individual was a native of *Sant Feliu* and born on 1874. From then on, the longitudinal database allowed us to follow him in 11 different enumerations across his life until the last observation found in 1940, when he was 66 years old.

As can be seen from this example, one possibility of analysis with longitudinal data, is the work career of individuals. The first two observations of Bartomeu were without occupation. Around 14 years old, we found Bartomeu's first occupational observation as locksmith apprentice, which he continued until the first decade of the twentieth century. However, from the 1920s until the 1940s the four last observations of Bartomeu informs

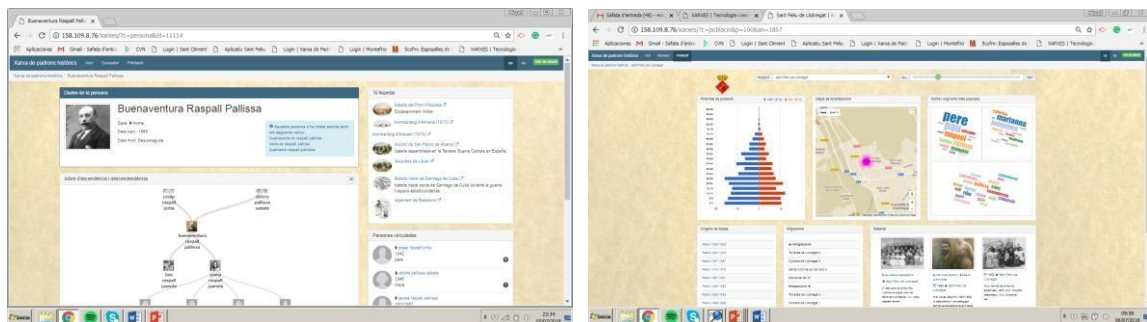
us of two different intermittent occupations, mechanic and turner. Thus, as can be observed through a single case, the longitudinal database of Sant Feliu informs us not only about changes at the individual micro level, but also how this individual reacted to macro level transformations such as industrialization, given he turned from an artisan position to a job that emerged with a new mode of production.

3.8 Popularizing science: Onomastics browser & demographic visualizations

The NETWORKS project has an important commitment to citizen science that goes beyond their participation in census transcription. The project also aimed at reaching the public by contributing to society through the creation of a dedicated browser constructed to display onomastics and interactive demography (figure 3.14). In May 2017 the onomastics application of the Sant Feliu database was presented: <http://158.109.8.76/xarxes/>. This also shows the genealogical tree (with a varying number of generations) for each individual in the town.

Moreover, it includes interactive visualizations of population pyramids and the town's most common names and surnames during the nineteenth and twentieth centuries. Additionally, we linked the web site with Wikipedia to provide the historical context of the database and biographical information about individuals recorded in the censuses. Wikipedia articles are also presented through an app (figure 3.15).

Figure 3.14. Onomastic browser.



Source: Authors' own elaboration.

Figure 3.15. Wikipedia connection.



Source: Díaz, 2018.

Another mobile app is developed to enrich the census information with pictures, kept in historical archives or family archives.

3.9 Concluding remarks

Citizen science, crowdsourcing, historical demography, computer vision and gamification are the cornerstones of the *Ball demographic database*. Formally, the Ball database is a relational database containing census data at the individual level from the Catalan county of *Baix Llobregat* in Spain for the nineteenth and twentieth centuries with more than 221,667 individual observations built by interdisciplinary projects in Digital Humanities. These projects are a joint venture of the Center for Demographic Studies and the Computer Vision Center, both at the Universitat Autònoma de Barcelona. Hence, this database is a proof that activating volunteer citizens through an *ad hoc* online crowdsourcing platform as a data entry tool and games for touch-screen devices, and Handwritten Text Recognition (HTR) techniques into data collection of primary sources can meaningfully reduce the time for building individual-level historical demographic databases. This integration has been possible thanks to recent advances in Handwriting Recognition, the expansion of information technologies, the popularization of handheld devices, the assimilation of internet in everyday life and the massive digitalization of historical documents. However, the current state-of-the-art of Handwriting Recognition still requires human intervention.

Thus, two scenarios have been designed where computer vision algorithms, based on deep learning techniques have been applied in the BALL database to speed up transcription. First, assisted transcription by information transfer from one document to another and second, automatic transcription with manual validation. These scenarios have proved to reduce progressively the human effort. In the first scenario, the system benefits from the redundant information (names, surnames and addresses) shared by consecutive censuses only needing the manual transcription (through a crowdsourcing platform) of one census, whose information is transferred to the next census. Using visual word search, so-called word spotting, each transcribed household in the first census is linked with the image of the same household a few years later, assisting the manual updating of information from one census to the next one. Moreover, the use of gamesourcing experiences for the validation of automatic transcription carried out by handwriting recognition algorithms is a twist in the integration of computer vision methods and keeping the human the intervention. At the same time, the engagement of citizens through gamification appears to be higher, as can be expected, which has a beneficial effect on the total number of transcribed words. Overall, we have set the stage for effective semi-automatic processing of large document collections in order to create databases in a faster and more effective way as part of the Big Data revolution.

Additionally, the BALL database is more than a database for demographic and computer vision research. The possibility of browsing this database through names and surnames using an open access and user friendly webpage boost greatly their use, mainly for the public at large, who is usually not familiar with this kind of resources. Queries can be also done through genealogical trees, that have been generated using record linkage techniques, facilitating extremely the search task for many users, who are looking for their ancestors. At the same time, visualizations of the population's evolution and onomastics through graphics are offered as tools to popularize demography among the general audience. In this way, citizens who are a fundamental part in the building of the database can also take advantage of the final result through the above-mentioned webpage, which is an important asset in terms of knowledge transfer to society. Besides, in spite of pursuing scientific objectives, these citizen-centered projects have an important potential social impact in terms of literacy or shortening the technological gap in terms of social inclusion and cohesion.

Annex 3.1. Variables of Local Censuses of Sant Feliu de Llobregat, 1828-1955.

Variables	Local Censuses																	Total					
	1828	1833	1857	1857	1878	1881	1886	1889	1890	1896	1900	1906	1910	1915	1920	1924	1930		1940	1945	1950	1955	
Name	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	21
Surname 1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	21
Surname 2			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	19
Occupation			X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	18
Sex																X	X	X	X	X	X	X	6
Number of Men in the household	X																						1
Number of Women in the household	X																						1
Number of household																X	X	X	X	X	X	X	6
Relation to the head of household							X			X	X	X	X	X	X	X	X	X	X	X	X	X	13
Civil Status		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	19
Age		X	X	X	X	X	X		X	X	X	X	X	X	X	X							14
Street Name		X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	18
Number of the household	X				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	18
Citizenship																X	X	X	X			X	5
Municipality of Birth						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Province of Birth						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Person Number			X	X		X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Date of Birth					X	X	X	X	X	X	X	X	X		X				X	X	X	X	13
Month of Birth					X	X	X	X	X	X	X	X						X	X	X	X	X	13
Year of Birth					X	X	X	X	X	X	X	X		X			X	X	X	X	X	X	14
Municipality of usual residence					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Province of usual residence																X	X	X	X	X	X	X	6
Years living in the municipality						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Months living in the municipality						X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Territorial contribution							X	X	X	X	X	X	X										7
Industrial contribution							X	X	X	X	X	X	X										7
Classification of inhabitant							X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	13
Literacy						X			X	X	X	X	X	X	X	X	X	X	X	X	X	X	14
Residential Status																X	X	X	X	X	X	X	5
Wages and Salary																		X				X	2
Total	4	6	7	7	6	17	18	18	21	22	22	22	22	19	19	21	22	25	24	23	25	370	

Source: Authors' own elaboration.

4 Assaig 2: Knowledge graph based methods for record linkage.

4.1 Introduction

Nowadays, it is common in historical demography the use of individual-level data as a consequence of a predominant lifecourse approach for the understanding of the demographic behaviour, family transition, mobility, etc. The building of individual lifecourses often involves the link of the birth certificate (baptism), the death certificate of the same individual and, for those who got married, their marriage certificate. These certificates were compiled nominally and manually in the so called *Quinque Libri* (parish registers) from the Council of Trento in the 16th century onwards throughout Europe (Willigan, 2013). Moreover, and as a consequence of the establishment of the Liberal State in Europe, civil registers and censuses spread out from the 19th century, although in some countries spread out even before [García Pérez, 2007; García Ruiperez, 2012; Valero, 1986]. Consequently, individual lifecourses can also be enriched by sequential observations of the same individual over time and space using census material. Indeed, record linkage is the task of finding records in a data set that refer to the same entity across different data sources in which individuals do not have a unique identifier [26]. In addition, other studies in the discipline apply an intergenerational perspective to determine the transmission of demographic behaviour, social outcomes or diseases, which is key in genetic studies and necessarily requires the connection across generations (Fu et al., 2014), which also implies the linkage between different generations of a same family. However, poor data quality can turn record linkage into a challenging task. The conservation status of original documents, the scanning process, the large number of similar values in names, ages or addresses are just some of the multiple factors that can affect data quality. More importantly, the relationship between household members and their head can significantly change between two censuses. For example, people were born and died, got married, changed occupation, or moved home. These changes on individual's records make record linkage in the historical census, i.e. those collected in the 19th and early 20th century and where only limited information about individuals were available, even more challenging. As a result, record linkage methods as those based on string distances are not reliable enough, and many false or duplicate matches are often

generated. This is also a common problem in other record linkage applications, such as author disambiguation (Elmagarmid et al., 2007). Knowledge graphs (KG) allow organizing semantics in a structured manner. They represent a collection of interlinked descriptions of real-world objects, events, situations or abstract concepts, which are called entities, in a formal structure. Popular KG include Freebase, DBpedia, YAGO, Satori, etc. and they have millions of entities and billions of entity links. Many researchers have developed various link prediction methods, also known as entity alignment (EA) methods, using KG embedding techniques since these large KG are still incomplete with some missing links (Bordes et al., 2013; Lin et al., 2015; Wang et al., 2014).

Therefore, there is the need of record linkage methods able to deal with heterogeneous data sources. Existing methods are tailored to a particular set of attributes and this makes more difficult record linkage across heterogeneous sources. Furthermore, these methods must be able to cope with data variability, due to individual lifecourse changes and robust to acquisition errors and unverified transcriptions as well. KGs provide a conceptual framework in which record linkage prediction and cross-linkage, between different sources, are naturally defined. Indeed, the proposed method, named WERL, is a step forward in record linkage methods since it overcomes some of the main difficulties discussed earlier. The main contributions of this paper are:

- Evolution knowledge graph. We use KGs for census data and we enrich them to take into account changes on attribute values of linked records.
- A two step learning process. Thereby, we separately learn attribute embeddings and their weight in record linkage tasks. This allows higher flexibility on embeddings learning and faster training.

In summary, we propose an EA method able to deal with heterogeneous data and that can be applied to other types of KGs in which entities are described in terms of attributes.

4.2 Related work

In this section we briefly review the main methods related to the record linkage task. First, we summarize the main methods which are devoted to this task. Then, we review those KG based methods which are closely related to the proposed method.

Record linkage methods

The record linkage task has been explored by a number of disciplines, including databases, statistics and artificial intelligence being key for disciplines like historical demography. Each discipline has formulated the problem slightly different and consequently different techniques have been proposed (Ravikumar i Cohen, 2004).

Early development in computing devices like the Hollerith Tabulator was a result of increasing census data size in the late 17th century and the lack of resources to tabulate and to analyze this information (Kistermann, 2005). Since then, and mostly from the 19th century onwards, many states gathered population censuses with the purpose of rationalizing the state population and wealth (Porter, 1995; Raphael, 2008; Wolf, 1989; Yule i Knedall, 1919). State statistics institutionalisation was a response to the wish of quantifiable material within an epistemological development framework of scientific objectiveness and an impersonal knowledge of issues and phenomena (Porter, 1995). It is worth noting how research centers from different countries have made their historical census data publicly available (see IPUMS NAPP¹³ 1 and Mosaic project¹⁴).

In the database community, this task is also known as deduplication. Deduplication aims at eliminating repeated data or multiple copies and thus compressing the database. To this end, it has been proposed various string edit-distance based methods for record matching like a general-purpose scheme (Monge i Elkan, 1997) or a knowledgeintensive approach (Raman i Hellerstein, 2001). In statistics, a long line of research has been conducted in probabilistic record linkage, largely based on the seminal paper by Fellegi (1969) formulated entity matching as a classification problem, where the basic goal is to classify entity pairs as matching or non-matching. They proposed using unsupervised methods, based on a feature-based representation of pairs which are manually designed and are, to some extent, problem-specific.

Although this can be a major problem when linking data from different sources, these proposals have been, by and large, adopted by subsequent researchers, often with elaborations of the underlying statistical model. The Jaro-Winkler distance has been used commonly for record linkage purposes (Winkler, 1999; Jaro, 1995). The AI community has focused on applying supervised learning to the record-linkage task for parameter learning of stringedit distance metrics (Ristad i Yianilos, 1998) and combining the results

¹³ <https://www.Nappdata.org/napp/>

¹⁴ <https://censusmosaic.demog.berkeley.edu/>

of different distance functions (Tejada et al., 2001). Unsupervised methods include similarity-based ones (Nilolov et al., 2012), probabilistic ones (Suchanek et al., 2011), hierarchical graph model-based ones (Ristad i Yianilos, 1998) and self-learning and embedding based entity alignment (Guan et al., 2017).

Knowledge-based representations

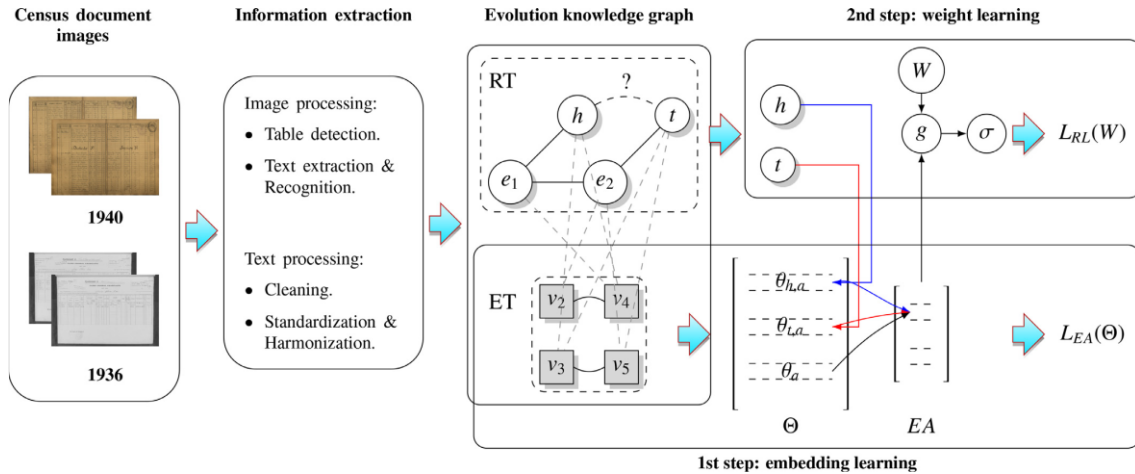
The most basic knowledge representation is the entity-relation (ER) graph, whose nodes are entities and the adjacency matrix represents the entity relations. The Translations in the embedding space (TransE) method generates knowledge graph embeddings of entities and relations, so that two related entities must have close entity embeddings in the embedding space (Bordes et al., 2013). A major drawback of this model is the poor accuracy while modelling one-to-many and many-to-many relations.

In order to overcome this problem, Wang et al. (2014) proposed the Translating on hyperplanes (TransH) method. In that work, an entity is first projected onto a hyperplane given for each relation. Thus, the method projects the entity according to the relation it is involved in. This allows the method to perform well even with one-to-many and many-to-many relations. Moreover, Lin et al. (2015) observed that the optimal dimension used to embed entities and relations could be different since they are completely different objects. In order to address this issue they proposed the TransR method, which learns a matrix to project embedded entities into the relation embedding space. This method performs well with one-to-many and many-to-many relations since it uses a different projection matrix for each relation.

RESCAL is a compositional method, which generates relational embeddings using tensor factorization (Nickel et al., 2011). Each embedding dimension represents a single entity feature, thus if two entities have the same feature value, both should have similar values in the same embedding dimension. Similarly, the Holographic Embedding (HOLE) method combines the expressive power of the tensor product with the efficiency and simplicity of TransE by using the circular correlation of vectors to represent pairs of entities (Nickel et al., 2016). In addition, Path-based TransE (PTransE) is an extension of the TransE method, which considers multi-step relation paths along with direct links (Lin et al., 2015). Finally, the Bootstrapping Entity Alignment (BootEA) method aims to update entity embedding using bootstrap process which adds likely alignments to the knowledge graph (Sun et al., 2018). In recent times, there have been an abundance of

temporal knowledge graphs which have a timestamp attached to each triplet. Such graphs are useful to represent dynamic data where order of events matter (Trivedi et al., 2017). attempted to predict future events by evolving entity embedding with time (Li et al., 2017). presented Event-Evolutionary principles for recognising order of various events.

Figure 4.1. WERL, overview.



Document image analysis techniques and information extraction methods are first applied. Then, a KG representation with evolution is built. Learning is a two steps process: first, attribute and attribute values are learned. Second consist of attribute weights learning.

The KR-EAR method relies on a entity-attribute-relation (EAR) knowledge representation (Lin et al., 2015). In that work, the authors identified two kinds of elements in their knowledge graph: entities, which are related between them by means of relations, and attributes, which are a sort of entity features. Consequently, they defined the EA task as a product of two posterior probabilities. One modeling entities and relations and the other modeling attributes and entities. Moreover, self-learning and active learning techniques are used for embedding learning when not enough annotated data is available. The Self-Learning and Embedding based entity alignment (SEEA) method is an extension of the KR-EAR method (Guan et al., 2017). There, the authors applied reinforcement learning techniques to improve the performance of their method.

4.3 The WERL method

The Weighted embedding based record linkage (WERL) method relies on an extension of the EAR graph and its parameters are learned in a two step process. The full training process is depicted in Fig. 4.1. First of all, there are the usual document image analysis processes consisting of layout analysis, text line detection and text recognition. Then, the

extracted text has to be processed, cleaned, standardized and harmonized before building the KG. All these processes are by themselves complex enough and far from being error free therefore the historical data used in this paper has been checked and validated by social science experts to avoid being conditioned by those transcription errors in our experiments.

The next step is to build the KG. We modified the EAR representation by adding relations between attribute values, we named it Evolution knowledge graph (EKG). These relations have been used to learn attribute embeddings and attribute value embeddings in a first step and no other embeddings have been learned. Finally, we have learned the attribute weights to be used later during the record linkage process.

At record linkage time, we have considered pairs of candidate records, which can be from different census years, as for instance entities h and t in Fig. 4.1 . For all the attributes in common between these two records, we have compared their values by using the same EA method used in training. Then, the values have been aggregated and weighted accordingly to provide a score between 0 and 1 that gives the matching probability

In what follows, we provide the details of the WERL method. We start by introducing the notation used to describe all the different steps to be done for training and records linkage prediction.

Evolution knowledge graph (EKG)

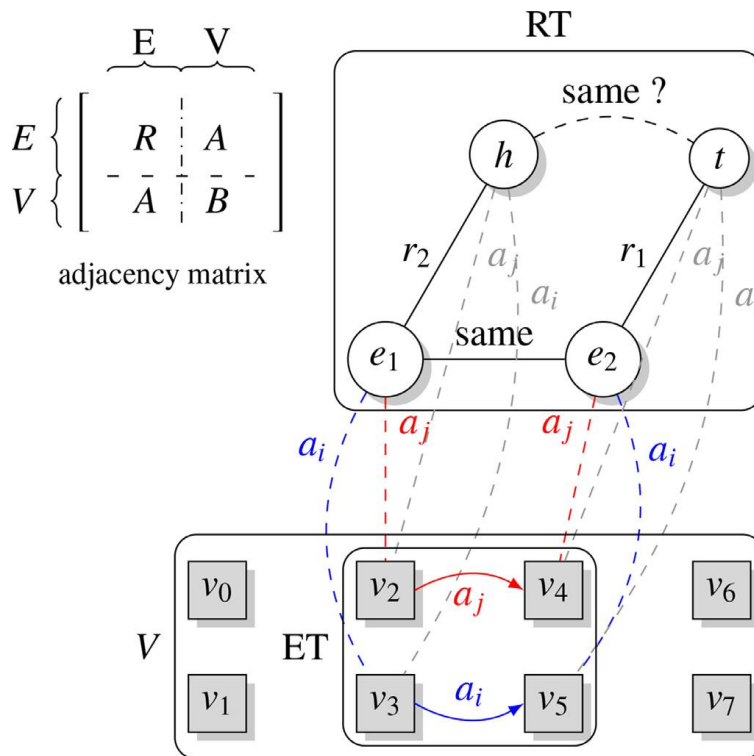
Let $G = (E, V, R, A, B)$ a KG. E and V are the graph nodes while R , A and B are the graph edges, see the top left matrix in Fig. 4.2 . E is the entity set and R the set of edges between entities. Given two entities h and t , respectively called head and tail, a relational triple (RT) is the triple: (h, t, r) .

Regarding KG attributes, V is the set of the attribute values while A , which is called attribute set, is the edge set linking entities and attribute values. We denote by AT the set of attribute triples: (h, v, a) and by V_a the attribute value subset, which is the domain of an attribute a . Similarly, V_e is the attribute value set corresponding to entity e . In other words and in the context of census records, V_e corresponds to the actual contents of a record e . Although the elements of V_e are functions of entity attributes $a : v = v(a)$, we will omit them when it be clear. Moreover, A_e is the attributes set of e and $A_{h,t} = A_h \cap A_t$ the set of attributes shared by both entities h and t .

Finally, and as being one of our contributions, we add the dependencies between attribute values in our KG representation. These dependencies model the possible changes on people records through time. To illustrate it, let us consider a kid born in 1867. In first records, he will appear as the son of the head of the household and, probably, his civil status, if reported, single. Later, some years later he would appear as being the head of his own household and probably married. Once he gets married, he would never become single again but could be widow and married. The rationale of adding these attribute values dependencies is to model these semantic dependencies.

Given 2 attribute values $v_i, v_j \in V_a$, that are related because they appear in two linked records, we define the evolution triple (ET) by the triple: (v_i, v_j, a) . These triples add arrows between attribute values into the KG, see Fig. 4.2 for graphical representation of our KG.

Figure 4.2. Example of the proposed Evolution KG representation.



RT is the set of relational triples: (e_1, h, r_2) , (e_1, e_2, same) and (e_2, t, r_1) . V is the set of attribute values: $\{v_0, \dots, v_7\}$. Triples: (e_1, v_3, a_i) , (e_1, v_2, a_j) , (e_2, v_5, a_i) , (e_2, v_4, a_j) and gray dashed triples belong to the AT set and $ET = \{(v_2, v_4, a_i), (v_3, v_5, a_j)\}$ the set of evolution triples. In this representation will be a missing value between h and t (dashed line). Top left: KG adjacency matrix.

Record linkage prediction

As we have explained at the beginning of this section, we perform record linkage by comparing the attribute values of two candidate records. Given two entities, h and t , we take their attribute value embedding vectors, $\theta_{h,a}$ and $\theta_{t,a}$ for all attributes $a \in A_{h,t}$ and we compare them by means of the EA method used to learn them. Then, we compute the weighted average given by g :

$$g(h, t; \Theta, W) = \sum_{a \in A_{h,t}} w_a S(v, u) EA(v, u, a; \Theta) \quad (1)$$

where $v \in V_h$ and $u \in V_t$, $W = (w_a)_a$ is a weight vector associated to attributes a , $S(u, v)$ is set to 0 if v equals u and 1, otherwise, and $(-)$ are the embedding vectors learned. To convert the score return by g to a value to be probability interpreted, we apply the sigmoid function, σ :

$$P(y = 1|h, t; \Theta, W) = \frac{1}{1 + \exp\{-g(h, t; \Theta, W)\}} \quad (2)$$

First step: embedding learning

The first step consists of learning the embeddings used to predict record links. To learn them, we can use any of the methods reviewed in Section 4.2 but applied only to the ET set. We do not need to learn embeddings to RT and AT sets since they are not used for record linkage.

Consequently, negative sampling must be done for ET sets. To do so, we generate them as it is done in similar approaches (Lin et al., 2014; 2015). For each attribute value $v_i \in V_a$, we sample attribute values from $V_a \setminus E_{v_i}$, where $E_{v_i} = \{v_j \in V_a \mid (v_i, v_j, a) \in E_T\}$. We denote by E^+ and E^- the positive and negative evolution triple samples, respectively. We use the same loss function for attribute and attribute values embeddings as the original KG-based embedding method. If we named that loss function L_{KG} , the actual loss function used to train the embedding is:

$$L_{EA}(\Theta) = \sum_{(v,u,a) \in E^+} L_{KG}(v, u, a; \Theta) + \sum_{(v,u,a) \in E^-} L_{KG}(v, u, a; \Theta) \quad (3)$$

Second step: weight learning

Once we have learned the embeddings, we must learn the attribute weights, which provide the impact of each attribute to record linkage prediction. Positive and negative RT, named respectively T^+ and T^- , come from the candidate pair set. Similarly to the ET set, T^+ is composed of all the linked records in the training set while T^- is composed of all the possible linked records that are not in T^+ . The loss function is the hinge loss evaluated on the predicted record linkage probability given by Eq. (2). This loss is evaluated in both positive and negative tuples as shown below:

$$L_{RL}(W) = \sum_{(h,t) \in T^+} \max\{0, \lambda_{RL} + P(y = 1|h, t; \Theta, W)\} + \sum_{(h,t) \in T^-} \max\{0, \lambda_{RL} - P(y = 1|h, t; \Theta, W)\} \quad (4)$$

4.4 Experiments

We carried out two experiments in order to evaluate the WERL method. In order to properly evaluate the contributions of this paper, we slightly modified the WERL method. We named MERL the modified method and it consists of not considering weights when predicting record linkage. The first experiment consists of comparing the WERL performance with some state of the art EA methods, described in Section 4.2, over three benchmark data sets detailed in the next section, namely BALL, Febri and Cora data sets. The second experiment aims at evaluating WERL, and MERL, robustness to data changes. To this end, we used a model trained on a census coming from one town, and we applied to census from neighbouring villages.

Data sets

The Baix Llobregat (BALL) Demographic Database provided the census records for Sant Feliu town of Barcelona collected in 16 different censuses from 1828 to 1940 (Pujadas-Mora et al., 2019). We note that the population of the town doubled two decades (1920–1940). The data set contains around 60K records of individuals with 30 attributes. Available attributes include individual’s full name, year of birth, civil status, occupation and relationship with the head of the family.

The Freely Extensible Bio-medical Record Linkage (Febri) is a synthetic data set about bio-medical records of patients (Christen, 2008). Each record provides information like full name, address, postal code and birth date. It contains 10K records split across two data sets of 5K records each. For each record, there is a duplicate record in the other data set. We attempt to link these duplicate records across these two data sets.

Cora data set contains bibliographical information about scientific papers in XML format (Draisbach i Naumann, 2010). It contains 1879 records of scientific citations with details like title, author(s), publisher, journal and date. The citations refer to 191 unique papers. We attempt to link all citations referring to the same scientific paper.

We split data into training, validation and test sets. We further partition each set in two data sets A and B to build candidate pairs (a, b). For FEBRL and CORA dataset, we split train-test sets randomly, while for the BALL dataset we create candidate pairs from two consecutive years. Consider C Y as the set of records for a census year Y from the BALL dataset. For training we use two pairs of dataset (A,B) as (C1889 , C1906) and (C1930 , C1936) . For validation we use (C1906 , C1910) and (C1936 , C1940) . For testing we selected (C1910 , C1924) and (C1924 , C1930) . To reduce the number of candidate pairs in the BALL data set we filter those pairs having the same second surname, see Table 1 for the counts of each of these sets. Then, we built the corresponding EKG for each of these data sets, see Table 2 for the complete counts.

Table 4.1. Data Partition into train, test and validation sets.

Data set	BALL			Febri			Cora		
	# Train	# Val.	# Test	# Train	# Val.	# Test	# Train	# Val.	# Test
Dataset A	9510	9384	10,629	3000	1500	500	470	314	156
Dataset B	10,629	10,722	10,536	3000	1500	500	470	313	156
Candidate Pairs	238,854	240,954	244,696	29,066	7674	959	220,900	98,282	24,336
True Pairs	6748	9049	8412	1952	995	340	7961	3601	883

Table 4.2. Training set size of Evolution KG.

	#E	#A	#R	#V	#AT	#RT	#ET
BALL	16,883	6	51	3492	82,500	13,750	10,643
FEBRL	6000	5	1	7497	30,000	0	1112
Cora	940	15	1	1764	7550	0	26123

Methods and parameters.

We implemented various translation based algorithms for generating knowledge graph embedding. We used tensorflow¹⁵ library to generate and optimize embedding for the KG structures defined above. We referred an open-source collection of graph embedding methods called OpenKE¹⁶, to implement the following KG embedding based methods for entity alignment:

- TransE and TransH for ER KG
- KR-EAR and SEEA for Evolution KG

To properly evaluate the contribution of each of the elements introduced in the WERL method we have implemented a one more method called: MERL. For MERL, we explicitly consider weights as 1 i.e. we consider the mean distance instead of weighted distance. Moreover, we applied these two methods on two KGs: a basic ER knowledge graph, in which attributes are also considered as being entities, and to the EKG introduced in Section 4.3. Finally, we applied grid-search on each of the considered methods to find the optimal hyper-parameters, see Section 1 of the Supplementary Material, for fair comparison.

1st experiment. record linkage evaluation.

Overall, EA methods fail when they are applied to record linkage tasks while record linkage methods achieve significantly better performance in all data sets in terms of Precision, Recall and F- score metrics, see Table 3. However, in terms of accuracy it seems that all methods obtain relatively good results although WERL and MERL based methods still achieve the best results. It should be said that these accuracy values can be misleading since we are in a two highly unbalanced classification problem, see Table 3. We note that the major concern is false positives, since incorrect linking can have a chain effect and have negative impact on the concerned research. Since the proportion of true pairs is very low (3% census, 7% febrl, 4% cora) compared to candidate pairs, it is hard to avoid false positives without increasing false negatives.

¹⁵ <https://www.tensorflow.org/install>.

¹⁶ <https://github.com/thunlp/OpenKE>.

Regarding the BALL data set, the WERL (EKG) method provides the best F-Score of 0.93. Among the compared EA methods, TransE has the highest F-Score of 0.21 only. TransH and KR-EAR have F- Scores of 0.17 and 0.18. We note that we get better results when training over evolution triples. Concerning the Febri data set, the MERL EKG provides the best F-score of 0.98. Among EA methods, TransH has the highest F-Score of 0.60, while TransE has a F- Score of 0.58. Applying self-learning to KR-EAR i.e. using the SEEA method improves the performance significantly. We note that the MERL (EKG) method perform better than WERL (EKG). Finally, as it regards the CORA data set, TransE provides the best F-Score of 0.54. SEEA was able to improve the results slightly from KR-EAR. Among our proposed methods, WERL (EKG) provides the highest F-Score of 0.46. We further note that WERL methods do not perform well on sparse dataset.

In summary, WERL (EKG) and MERL (EKG) methods provide satisfactory and stimulating results since with these methods we are able to learn proper embedding vectors for KGs.

Table 4.3. Record Linkage. Results for three data sets, presents Accuracy, Precision, Recall and F-score.

Data set Method	BALL				Febri				Cora			
	Acc.	P	R	F-Score	Acc.	P	R	F-Score	Acc.	P	R	F-Score
TransE (ER)	0.87	0.19	0.25	0.21	0.95	0.70	0.49	0.58	0.99	0.62	0.48	0.54
TransH (ER)	0.81	0.12	0.26	0.17	0.95	0.72	0.52	0.60	0.95	0.39	0.51	0.44
KR-EAR (EAR)	0.79	0.13	0.30	0.18	0.42	0.07	0.65	0.13	0.88	0.20	0.81	0.32
SEEA (EAR)	0.91	0.07	0.10	0.08	0.91	0.44	0.57	0.49	0.97	0.79	0.21	0.33
MERL (ER)	0.98	0.90	0.68	0.77	0.96	0.79	0.92	0.85	0.91	0.23	0.58	0.33
WERL (ER)	0.99	0.99	0.67	0.80	0.95	0.90	0.71	0.79	0.91	0.23	0.58	0.33
MERL (EKG)	0.99	0.76	0.88	0.82	0.99	0.99	0.97	0.98	0.95	0.32	0.39	0.35
WERL (EKG)	0.99	0.98	0.89	0.93	0.97	0.99	0.76	0.86	0.97	0.61	0.36	0.46

2nd experiment. Method robustness.

Additionally, we tested the WERL (EKG) and the MERL (EKG) methods on one more census dataset from Santa Coloma de Cervello and Castellvi de Rosanes towns of Barcelona. The dataset have 8631 records with similar attributes as were provided for Sant Feliu town.

We partition the dataset as Dataset A and B with 4815 and 3816 records. We considered records from census years 1866, 1924, 1936 and 1945 as dataset A . For the a dataset B , we selected the records from census years 1901, 1930, 1940 and 1950. We apply blocking

indexing on the second surname field to yield 83,264 candidate pairs, out of which only 6407 are true links. The results are provided in Table 4 . MERL (EKG) provides the best F-score of 0.66.

Table 4.4. Additional Results for Santa Coloma de Cervelló and Castellví de Rosanes towns.

	Accuracy	Precision	Recall	F-Score
WERL (EKG)	0.96	0.87	0.51	0.64
MERL (EKG)	0.94	0.58	0.77	0.66

4.5 Discussion

Overall, the proposed method achieves satisfactory performance compared to similar methods in record linkage tasks. If we analyse results on benchmark data sets, we notice that sparse data, as in the Cora data set, makes vector embedding learning harder since evolution triples need paired attributes in candidate records and thus the low F-score for these methods. For the febrl data set the most duplicates had surname and lastname swapped. For such errors, WERL suffers from over-fitting and MERL performs better.

Regarding the BALL data set, we notice that the WERL(EKG) method is able to learn transcription errors, which occurred during the digitisation process. Indeed, this method performs better when attribute values evolve in particular order than when they evolve randomly, since it suffers from over-fitting as well.

Moreover in these experiments the WERL (EKG) method generates 1243 false negatives and 96 false positives in the BALL data set. In what follows, we provide the different reasons for their incorrect classification:

Full name: Over 90% of false positives have different full names while over 60% false negatives have the same full names.

Year of Birth: Over 98% false negatives have the same years of births. All 96 false positives also have the same years of birth. Additionally, we had 45 false positives where year of birth is missing for both records.

Civil Status: Over 80% false negatives have the same civil status. We had 38 false positives with the same civil status i.e. about 58% false positives have different civil status.

Occupation: Over 50% false negatives had the same occupation across years. We had 24 false positives with the same occupation i.e. over 70% false positives have different occupations.

Relation: About 65% false negatives have the same relationship with head. We had 5 false positives with the same relationship i.e. about 95% false positives have different relationships across years.

To sum up, we note that WERL EKG linked the record pairs having valid evolution of attribute values while rejecting the record pairs having invalid evolution, even though they had some common attribute values. Next improvements must overcome these problems.

4.6 Conclusions and future work

In this paper we have proposed a KG based method for record linkage tasks, named WERL. We have enriched knowledge representation by introducing evolution relations. These new edges in the knowledge graph encode valid variations (evolution) on individual data. Thanks to this knowledge representation the WERL method is able to learn embedding vectors which better encode census data. Then in the second step, attribute weights are optimized for the record linkage task. This allows us to identify which attributes are more discriminant for record linkage tasks. Thereby, we have taken advantage of knowledge graph representation and the associated entity alignment methods and, at the same time, we have trained a specialized method for record linkage tasks.

We have evaluated the proposed method on benchmark data sets and we have compared it to related methods. The achieved results are satisfactory and stimulating since there is still room for improvement. The proposed method must be applied to larger and more heterogeneous census data sets to evaluate both how it scales when data sizes increase and its robustness to data variability. We are quite confident that knowledge graph methods will behave better than existing approaches.

5 Assaig 3. A tals pares, tals fills: Transmissió intergeneracional a Sabadell, segles XVI - XVII.

5.1 Introducció.

Al llarg de la història les famílies han perseguit la seva continuïtat en el temps, no únicament des d'un punt de vista biològic sinó també social i econòmic. D'aquesta manera, l'anomenada reproducció social seria l'intent de les famílies per afavorir als seus descendents en termes socioeconòmics amb la intenció de perpetuar els seus llinatges al llarg del temps (Bourdieu, 1976; Berkner i Mendels, 1976; Ganzeboom et al., 1991). Al contrari del que es podria pensar, aquest procés no era un fet exclusiu de les famílies i nissagues més benestants sinó que era un objectiu cercat de manera comú en els diferents grups socials (Ferrer, 2009). Per aconseguir aquest fi, les famílies històricament han utilitzat una sèrie d'estratègies basades en el matrimoni i la transmissió intergeneracional de la posició/ocupació (Berkner i Mendels, 1976; Ferrer, 2010; Kurosu, 1996).

La historiografia més clàssica sobre la transmissió intergeneracional de la posició social i/o ocupació ha perpetuat una visió de les societats preindustrials inequívocament immòbils. És a dir, en aquestes societats inevitablement els fills ostentarien les mateixes dedicacions ocupacionals que els seus pares. Aquesta perspectiva s'ha matisat en els últims anys (Boberg-Fazlic et al., 2011; Dambruynes, 1998), encara que és cert que els nivells de transmissió de la posició social en aquestes societats solien ser alts (Pujadas-Mora et al.; 2018; Bearman i Deane, 1992). No obstant això, els mecanismes de transmissió intergeneracional podien variar entre famílies com a conseqüència de les seves pròpies circumstàncies demogràfiques (número de fills/es i el seu gènere), per grups socials (per l'anomenat comportament de classes) o per un determinant context socioeconòmic que seria més o menys favorable (creació d'oportunitats) a la transmissió (Bourdieu, 1976; Laslett, 1983). En aquest sentit, els sistemes hereditaris podien enfortir la transmissió intergeneracional de la posició social, particularment quan el sistema es basava en un sistema d'herència indivisible com en el cas de Catalunya, en el qual generalment s'atorgava als fills (homes) primogènits el privilegi de ser els hereus (Barrera, 1992; Comas d'Argemir, 1988; Ferrer, 2007). En el cas que no existís

descendència masculina, la filla major seria anomenada l'hereva universal o *pubilla* (Ferrer, 2005; Arrizabalaga, 2005).

L'hereu o la pubilla havia de donar continuïtat a la casa, és a dir a la família. S'ha argumentat que la lògica d'aquest sistema comportava que els germans no hereus descendissin socialment en comparació a la posició social de les seves famílies d'origen (Ferrer, 2003 i 2010). Per tal de mitigar aquests efectes, les famílies solien adoptar mecanismes per compensar als no hereus, que podien anar més enllà del pur pagament d'una legítima com la utilització de les xarxes socio-ocupacionals dels pares i/o dels altres germans per tal de trobar una sortida ocupacional adient als descendents que no impliqués un descens social (Pujadas-Mora et al., 2018; Ferrer, 2010b).

L'objectiu principal d'aquest estudi és explicar la dinàmica de transmissió intergeneracional de transmissió de l'ocupació i/o la posició social a Sabadell en els segles XVI i XVII, així com explorar els possibles efectes del sistema d'herència universal, determinat per l'ordre de naixement i el sexe, en la destinació social de fills i filles a partir de la utilització de la BHMD (Barcelona Historical Marriage Database). D'aquesta forma, observarem com d'important era la influència familiar en la reproducció social dels individus, així com establir les diferents estratègies de col·locació de les famílies dependent de la seva posició socioeconòmica.

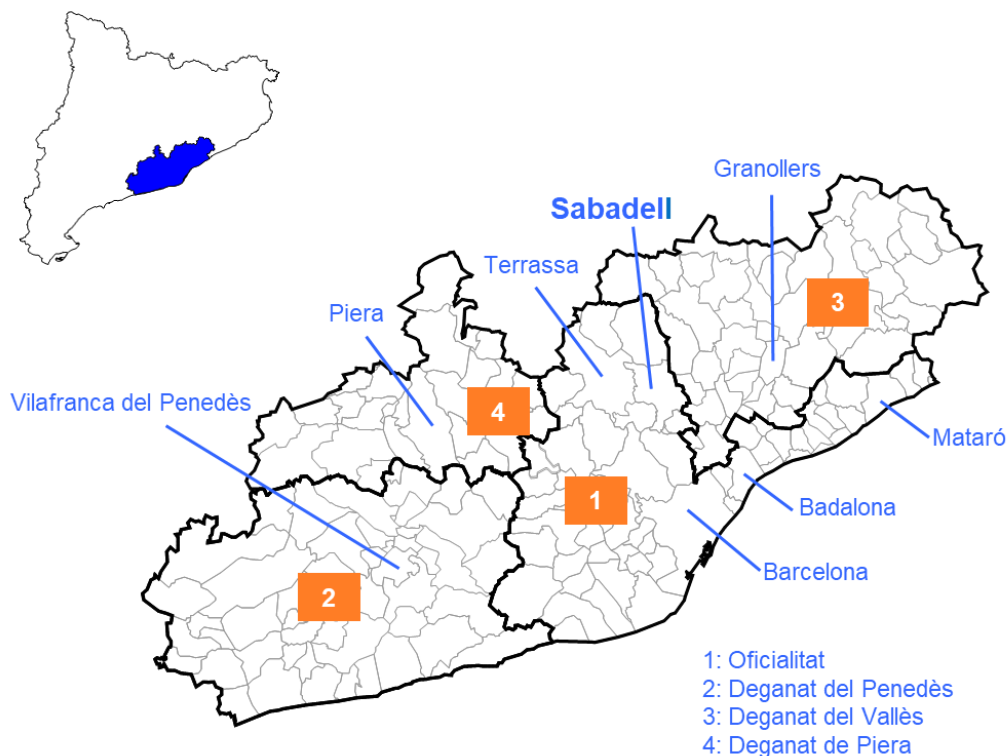
5.2 Els Llibres d'Esposalles de la Catedral de Barcelona.

Els registres sacramentals són la font per excel·lència per a l'estudi de les poblacions en el passat arreu d'Europa. Cal dir que malauradament en el cas de Sabadell no es conserven els registres sacramentals anteriors a 1939, fet que creiem que ha condicionat els estudis de població sobre la ciutat. Afortunadament, els **Llibres d'Esposalles** de la Catedral de Barcelona conservats al seu arxiu capitular poden ajudar en la construcció del coneixement històric de la població de Sabadell. En el marc del projecte "Cinc segles de matrimonis" (*Five Centuries of Marriages*) dirigit per la Professora Anna Cabré, s'ha construït amb els Llibres d'Esposalles del Arxiu de la Catedral de Barcelona una gran base de dades anomenada la **Barcelona Historical Marriage Database** (BHMD) que recull els més de 600.000 matrimonis que es celebraren entre 1451 i 1905 a la Diòcesi de Barcelona. Per tal de comprendre aquesta font, ens hem de remuntar als seus orígens quan el 29 de setembre de 1409 el papa Benet XIII (Pedro Martínez de Luna, 1328-1423) va

concedir a la catedral de Barcelona un impost a satisfer per cada un dels matrimonis que es celebressin a la Diòcesi de Barcelona segons la categoria social dels contraents amb la finalitat de contribuir a la construcció i posterior manteniment de la catedral (Carreras Candi, 1913).

El fons documental d'aquests llibres està format per 291 volums, 244 dels quals corresponen a l'Oficialitat de Barcelona (el deganat principal de la Diòcesi de Barcelona).¹⁷ Aquesta divisió eclesiàstica abastava les comarques actuals del Barcelonès, gran part del Vallès Occidental, del Baix Llobregat i del Maresme contant amb un centenar de parròquies a 1900, entre les quals es troben les parròquies de Sant Feliu i la Puríssima Concepció del Sabadell del moment (Figura 5.1). La resta de llibres, 47, es corresponen amb els deganats del Vallès, del Penedès i de Piera. Concretament, per a la ciutat de Sabadell es registraren des de 1451 a 1905 12.197 esposalles.

Figura 5.1. Oficialitat i deganats de la Diòcesi de Barcelona fins 1905.



Font: Elaboració pròpia

Una de les peculiaritats dels Llibres d'Esposalles es que mantingueren la mateixa estructura gairebé intacte pels cinc segles que es conserven com a conseqüència de la seva

¹⁷ La Diòcesi de Barcelona s'organitzava territorialment amb l'Oficialitat de Barcelona, el deganat del Penedès, el deganat de Piera i el deganat del Vallès).

funció de llibre de comptes. De fet, aquests llibres matrimonials tenien la funció de registrar la taxa que pagava cada matrimoni. Les informacions que apareixen sempre entre 1451 i 1905 són el nom, primer cognom i ocupació del marit, la data en que es pagava l'impost i la seva quantitat. La resta d'informacions han sofert variacions al llarg del temps. El nom de l'esposa no s'esmentà fins el 1485. A partir d'aquell moment i fins a 1643, les mullers únicament s'identificaven pel seu nom de pila ja que sempre anaren referides al seu pare o en cas de ser vídues, al seu difunt marit. Serà a partir d'aquella data que trobarem el primer cognom de l'esposa. Haurem d'esperar fins a 1876 quan apareixerà el segon cognom tant per l'esposa com també pel marit. D'altra banda, el nom dels pares i mares dels contraents es van registrar en dos períodes, entre 1573-1641 i 1750-1905. Pel que fa als seus cognoms, si que es recollia el del pare, però el de la mare es va començar a anotar a partir de 1750.

Pel que fa a les referències geogràfiques, trobem que des de 1481 fins a 1641 es recull l'origen i la residència dels contraents i els seus progenitors. A partir de 1643 es perd gairebé tota referència geogràfica del matrimoni fins que al 1715 es començà a recollir la parròquia de celebració de l'enllaç matrimonial fins 1905. Les ocupacions dels progenitors només es troben durant el període 1573-1641¹⁸. Per aquest motiu, l'estudi que ens proposem sobre la transmissió intergeneracional de la posició / ocupació només és possible pel període esmentat. La taxa matrimonial des de 1575 fins a 1649 es van fixar en set nivells impositius. 1: La noblesa titulada (ducs, comtes, marquesos i vescomtes) pagava 12 lliures. 2: Els cavallers, nobles i senyors de vassall abonaven 2 lliures i 8 sous. 3: Els ciutadans honrats i doctors en medicina i dret pagaven 1 lliura i 4 sous. 4: Els mercaders i notaris de Barcelona havien de sufragar 12 sous. 5: Els mestres gremials, botiguers de teles, notaris reials, causídics i negociants contribuïen amb 6 sous. 6: Els pagesos i artesans pagaven 4 sous. 7: Els pobres, rebien gratuïtament la llicència matrimonial per matrimoni "gràcies" o per "l'amor" de Déu (*Gratia Deo* o *Amore Dei*). No obstant això, la majoria dels individus que reberen aquesta dispensa declararen algun tipus d'ocupació. Els impostos es van registrar en lliures, sous i diners com a unitats de

¹⁸ Aquesta és una de les principals raons per la qual el període abarcat en l'article agafa el 1545-1645, ja que va ser quan millor es va conservar la informació ocupacional/social de pares i sogres. Un altre element a tenir en compte és que ens referim a relacions ocupacionals o social intergeneracionals, només podem tenir en compte a les vies masculines, és a dir, pares i fills. En el cas de les dones només podem observar les tendències de endogàmia i/o exogàmia social, és a dir, entre sogres i gendres, ja que infelïçment en els Llibres d'Esposalles, així com en la major part de fonts històriques de la població, les ocupacions femenines van ser moltes vegades deliberadament no registrades (Humphries i Sarasúa, 2012).

compte, i els valors es van mantenir estables fins a finals del segle XIX (1896), encara que es van pagar amb la moneda de curs legal en cada etapa¹⁹.

Entre 1545 i 1645 es registraren 1.925 llicències matrimoniales en l'àrea de Barcelona que tinguessin almenys un dels cònjuges relacionats amb Sabadell. El fet de què aquest article es centri en la transmissió intergeneracional de l'ofici i/o posició social fa que només es tinguin en compte aquelles llicències matrimoniales en què l'ocupació del pares i dels fills/gendres s'hi registrés. D'aquesta manera, s'han conservat 599 llicències matrimoniales amb les informacions socioeconòmiques completes per pares i fills. Cal tenir present que quan els progenitors no eren nadius o els contraents eren vidus la informació relativa als pares no es registrava.

5.3 Metodologia.

Les bases de dades històriques necessiten ser estandarditzades. La inexistència d'un català estàndard fins a la normativització moderna del català de Pompeu Fabra a principis del segle XX, va afavorir l'aparició al llarg del temps de múltiples variacions escrites d'un mateix nom o cognom. Això fa que una mateixa persona la trobem escrita en diferents variants, així com també les múltiples formes d'esmentar una mateixa ocupació. De tal manera que s'ha realitzat una harmonització dels noms i cognoms i les etiquetes ocupacionals han estat codificades a partir de la *Historical International Classification of Occupations* (HISCO) que atorga un codi específic a cadascuna de les ocupacions i està basada en la classificació ISCO68 de l'Organització Internacional del Treball (OIT) (Van Leeuwen et al., 2002). Una vegada totes les ocupacions s'han codificat van classificar-se a partir de la HISCLASS (Historical International Social Class Scheme) (Van Leeuwen i Maas, 2011) que converteix i agrupa en classes socials els diferents codis HISCO, basant-se en nivells d'especialització, formació i jerarquia ocupacional. A partir d'una adaptació específica a les ocupacions històriques catalanes dissenyat per a la *Barcelona Historical Marriage Database* (BHMD) els codis HISCO també es poden agrupar a partir de sectors i subsectors econòmics (Pujadas-Mora et al., 2014). Així, els anàlisis i resultats obtinguts

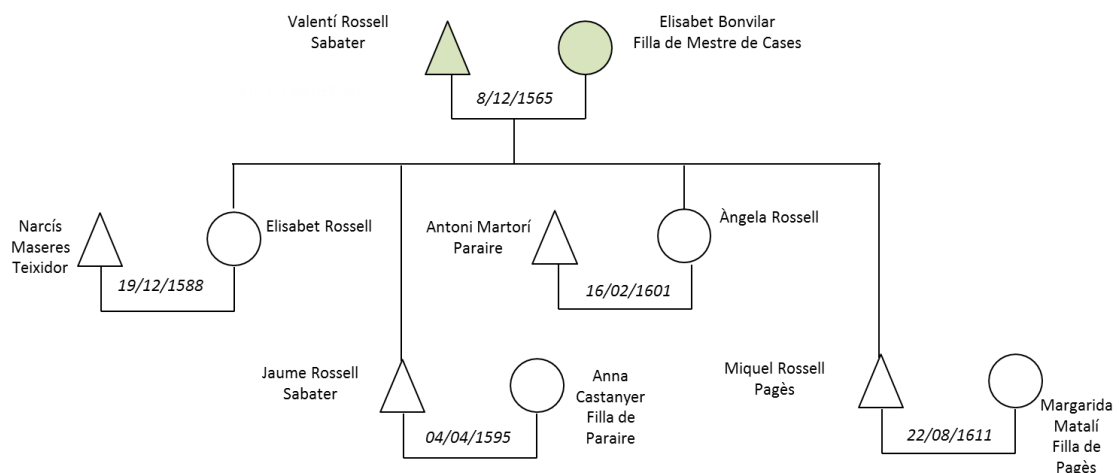
¹⁹ Unes taxes que s'expressen en lliures i sous, fet que les fa comparables al llarg del temps, sense haver de fer conversions depenent de la moneda de l'època. Aquestes 8 categories es mantindran fins a 1857, quan s'elimina la categoria de Mercaders de Barcelona. A partir de 1896 els impostos van començar a registrar-se en pessetes.

tenen una major robustesa estadística i alhora poden ser comparables a nivell internacional.

Per altra banda, atès que l'objectiu d'aquest article és justament analitzar la transmissió intergeneracional de l'ocupació o posició social mitjançant registres matrimonials ha estat necessari realitzar una reconstrucció genealògica. Les dades matrimonials dels Llibres d'Esposalles que abasten els segles XVI i XVII permeten la reconstrucció genealògica de les famílies, independentment de la seva extracció social, un aspecte que generalment no sovinteja en els estudis històrics ja que aquests tipus de reconstruccions han estat més habituals en els grups preeminents, ja sigui perquè deixen més petjada documental o perquè moltes d'aquestes famílies han mostrat una major preocupació en recollir el seu passat com a part del seu comportament de classe (Soria Mesa, 2013). Alhora cal tenir present que les famílies que es poden reconstruir són aquelles que han tingut descendència que també s'ha casat, i que podem entendre com la fracció útil de la descendència.

D'aquesta manera s'ha vinculat els matrimonis dels pares amb els matrimonis dels seus fills al llarg dels segles XVI i XVII. Amb aquest propòsit s'han creat dos tipus de connexions. Per un costat, s'han vinculat els germans/es d'una mateixa família i, per l'altre, s'han enllaçat els matrimonis dels germans/es amb els matrimonis dels seus pares (Figura 5.2.).

Figura 5.2. Matrimonis de la família Rossell-Bonvilar (1565-1611)



Font: Elaboració pròpia a partir de la BHMD.

Metodològicament, aquestes reconstruccions genealògiques s’han realitzat mitjançant el *software* “BuscaDescendències” desenvolupat en el projecte *Five Centuries of Marriages* per investigadors del Centre de Visió per Computador (CVC) i el Centre d’Estudis Demogràfics (CED) de la Universitat Autònoma de Barcelona. Aquest programa funciona utilitzant dos algoritmes (*string distances*) que cerquen una informació comuna, el nom o cognom dels progenitors en aquest cas concret, en diferents llicències matrimonials. Aquest software té incorporat una funcionalitat lingüística per tal de pal·liar els efectes de les similituds en la pronunciació i els errors tipogràfics (Villavicencio et al., 2015). Les vinculacions nominatives també estan subjectes a una limitació temporal atent al fet de que els homes no podien casar-se abans dels catorze anys i les dones abans dels dotze (Gaudemet, 1987).

Pel cas de Sabadell entre 1545 i 1645, el software ha generat 636 *links* individuals entre el matrimoni dels pares i el matrimoni dels seus fills. Aquesta número de *links* s’ha reduït a 256 després d’un procés de validació amb criteris probabilístics tenint en compte variables com les ocupacions, les taxes pagades o la procedència geogràfica. Alhora només s’han seleccionat aquelles llicències matrimonials en les quals tota la informació nominal pels progenitors és igual en la llicència del matrimoni dels pares i dels fills. Finalment, també s’han descartat els emparellaments de registres matrimonials en els quals el temps entre els matrimonis de pares i fills/es fos massa gran, imposant una distància mínima de 15 anys entre esposalles i una màxima de 65 anys. El resultat final de la reconstrucció és un total de 55 famílies de Sabadell entre 1545 i 1645 (Taula 5.1).

Taula 5.1. Reconstrucció Genealògica de Sabadell (1545-1645).

Nº de germans/es	Nº de famílies (55 en total)
2	30
3	12
4+	13

Font: Elaboració pròpia a partir de la BHMD.

Pel que fa a l’anàlisi quantitatiu de la transmissió intergeneracional de l’ocupació i/o del status social s’han fet servir tècniques estadístiques bivariants que permeten quantificar els nivells de heretament del mateix grup social o de mobilitat entre generacions. D’aquesta manera, s’han fet servir taules de doble entrada també conegudes com taules

de mobilitat que permeten comparar l'ocupació/grup social dels pares amb la dels fills/es. A més, també s'han estimat els residus estandarditzats de la distribució de pares i fills en les diferents ocupacions/grups socials per tal quantificar si la coincidència entre els grups socioeconòmics dels pares i fills/es està sobrerepresentada o si al contrari és poc freqüent en relació a l'estructura social de la població analitzada. El funcionament dels residus estandarditzats es basa en la comparació de la freqüència observada, és a dir el nombre de pares i fills en cada categoria social, i la freqüència esperada. En altres paraules, aquesta estimació permet observar si els individus que formaven part d'un determinat grup social depenia o no del seu grup social d'origen. Un valor superior a +2 del residu estandarditzat en una determinada combinació del grup social del pare i del fill representa una combinació molt més freqüent del que s'esperaria per l'estructura social existent en la població estudiada. Un valor per sota de -2 indicaria que és una combinació menys comuna. Els valors entremig mostren que no existí una prevalença específica en la transmissió intergeneracional en determinades ocupacions.

5.4 La transmissió intergeneracional en clau de grup social.

Encara que l'estudi de les genealogies concretes poden il·lustrar d'una manera interessant les dinàmiques existents en la transmissió intergeneracional al Sabadell dels segles XVI i XVII, per tal de poder saber de forma més específica com aquests processos funcionaven, hem utilitzat les anomenades taules de mobilitat social (Lipset i Zetterberg, 1956; Grusky i Hauser, 1984). Aquest tipus de taules permeten estimar els nivells de transmissió intergeneracional per a tots els matrimonis celebrats a Sabadell entre 1545 i 1645. D'aquesta forma, hem categoritzat totes les ocupacions de pares i fills segons sis grups socio-ocupacionals diferents en una taula de doble entrada que presenta a les files els grups als quals pertanyen els pares i a les columnes els dels fills. Aquestes sis categories s'han construït a partir del sistema internacional de classificació social HISCLASS (Van Leeuwen i Mass, 2011) que engloben tots els estrats socials presents a Sabadell. Així s'observarà des d'individus de la baixa noblesa a d'altres que es declaraven treballadors.

El primer element que podem observar en la taula 5.2 és que el pes dels menestrals i pagesos era majoritari en l'estructura socioocupacional de Sabadell, un fet propi de les

societats d'Antic Règim (Brea-Martínez i Pujadas-Mora, 2018; Bearman i Deane, 1992; Boberg-Fazlic et al., 2011). Nogensmenys, si observem el nombre total de pagesos i menestrals pares (files) i fills (columnes) podem veure una inversió en el protagonisme dels grups. Els pares en el moment del seu matrimoni declaren més sovint ser pagesos que els seus fills, els quals mostraren ser més habitualment menestrals. Així, per 319 pagesos pares trobem 222 fills pagesos i per 264 pares menestrals, 335 fills homònims. Això s'explicaria per un context de transició econòmica que es va donar a partir del segle XVII a l'àrea de Barcelona. Molts gremis originalment assentats a Barcelona es varen distribuir per localitats del prelitoral com Sabadell per no haver de suportar una pressió fiscal creixent que es sofria a la ciutat comtal (García Espuche, 1998; Torras, 1984).

Pel que fa als nivells de transmissió intergeneracional de la posició socioeconòmica, aquests eren alts talment com mostra la línia diagonal de la taula 5.2, que ensenya la coincidència entre la mateixa categoria socio-ocupacional de pares i fills. De fet, aquesta coincidència seria del 75% dels casos, una xifra que situa a Sabadell en un nivell similar a d'altres indrets d'Europa al mateix període com Anglaterra (Boberg-Fazlic et al, 2011; Bearman i Deane, 1992).

Taula 5.2. Mobilitat social intergeneracional a Sabadell (1545-1645)

1545-1645		FILL					Total
		Noblesa	Administració i Prof Liberal	Mercaders i Comerciants	Menestrals	Pagesos	
P A R E	Noblesa	0	0	0	0	0	0
	Administració i Prof Liberal	0	6	0	0	0	6
	Mercaders i Comerciants	1	1	3	3	1	9
	Menestrals	0	5	10	233	13	264
	Pagesos	0	2	6	98	208	319
	Treballador-Jornaler	0	0	0	1	0	1
Total		1	14	19	335	222	599

Font: Elaboració pròpia a partir de la BHMD.

En la taula 5.3 s'han estimat els anomenats residus estandarditzats, una tècnica estadística que ens ajuda a saber si la combinació dels grups socioeconòmics dels pares i fills està sobrerrepresentada o si al contrari és poc freqüent en relació a l'estructura social general

de la població analitzada. Si analitzem els resultats dels residus estandarditzats de les combinacions socioocupacionals de pares i fills a Sabadell entre 1545 i 1645, el primer element que cal destacar és que les intensitats de transmissió intergeneracional de la mateixa posició social eren les més altes, com s’hauria d’esperar. Entre els grups amb major transmissió entre pares i fills estaria aquell relacionat amb l’administració i els professionals liberals, els quals “produïen” fins a 15 vegades més fills amb el seu mateix estatus socio-ocupacional del que s’esperaria en una distribució aleatòria. Aquesta intensitat tan alta d’aquest grup era comuna també entre els nobles (molt poc representats a Sabadell) i sobretot entre els pagesos. D’una altra banda, com podem veure entre els dos grups més nombrosos com era el cas de pagesos i menestrals, l’heretament ocupacional dels fills era també important i més freqüent del que s’esperaria. Aquest fet s’entendria no tant per una reticència a què un fill es fes pagès o menestral sinó més bé per una qüestió de limitacions i/o oportunitats. Un fill de pagès que esdevingués artesà comportaria salvar importants obstacles com les rígides regulacions gremials i la manca d’una xarxa en la menestralia, els quals no eren completament insalvables (Torras, 1984; Pujadas-Mora et al., 2018; Ferrer, 2010). En contrapartida, en el període d’estudi a molts d’indrets de la Catalunya vella el mercat de terres agrícoles era molt minso, el que va fer que el preu de les terres s’encarís (Serra i Puig, 2013; García Espuche, 1998).

Taula 5.3. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus socio-ocupacional a Sabadell segons la posició social de pares i fills, 1545-1645

1545-1645		FILL					
		Noblesa	Administració i Prof Liberal	Mercaders i Comerciants	Menestrals	Pagesos	Treballador-Jornaler
P A R E	Noblesa	-	-	-	-	-	-
	Administració i Prof Liberal	-0,10	15,65	-0,44	-1,83	-1,49	-0,28
	Mercaders i Comerciants	8,04	1,72	5,08	-0,91	-1,28	-0,35
	Menestrals	-0,66	-0,47	0,56	7,02	-8,58	-0,28
	Pagesos	-0,73	-2,00	-1,29	-6,02	8,26	0,36
	Treballador-Jornaler	-0,04	-0,15	-0,18	0,59	-0,61	-0,12

Font: Elaboració pròpia a partir de la BHMD.

Fins ara hem presentat els nivells generals de transmissió intergeneracional de la posició social a Sabadell, però és del tot cabdal saber quins eren els destins social segons l'ordre de naixement, en aquest cas l'ordre de casament, sobretot per saber quin era l'estatus socioeconòmic d'aquells que no eren hereus. Com s'ha comentat anteriorment, els nivells de transmissió intergeneracional de la posició social a Catalunya estaven directament determinats per les implicacions del sistema d'herència no divisible. D'aquesta manera, tindria sentit pensar que els hereus tinguessin una major propensió a tenir la mateixa ocupació que els seus pares. Si utilitzem els residus estandarditzats per analitzar la diferència segons l'ordre de casament dels fills, entenent que els primers germans que es casaven solien ser els hereus (Marfany, 2012), es confirma que els germans que es casaven primer mostren una major propensió a presentar la mateixa ocupació o posició social del seu pare en comparació als seus germans que es casaven en segon o tercer lloc (Taula 5.4).

Entre els grups més benestants com els mercaders i comerciants, independentment de l'ordre de casament, la seva transmissió intergeneracional era molt alta. En canvi, els germans que es casaven en segon o tercer lloc en el grup de menestrals o pagesos mostraven un menor heretament en els oficis que el que observem al sector comercial.

Taula 5.4. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus sòcio-ocupacional a Sabadell segons la posició social de pares i fills i per ordre de casament, 1545-1645

1r casat		FILL		
		Mercaders i Comerciants	Menestrals	Pagesos
P A R E	Mercaders i Comerciants	6,25	-0,68	-0,72
	Menestrals	-0,64	2,89	-2,61
	Pagesos	-0,75	-2,35	2,39

2n casat		FILL		
		Mercaders i Comerciants	Menestrals	Pagesos
P A R E	Mercaders i Comerciants	5,37	-1,03	-0,86
	Menestrals	-1,11	2,57	-2,71
	Pagesos	-0,51	-1,94	2,57

3r o + casat		FILL		
		Mercaders i Comerciants	Menestrals	Pagesos
P A R E	Mercaders i Comerciants	-	-	-
	Menestrals	-0,42	2,71	-1,57
	Pagesos	0,19	-1,25	0,73

Font: Elaboració pròpia a partir de la BHMD.

De fet, un major ordre de matrimoni dels germans (segon i principalment tercer o més) mostra que és més probable trobar un pare pagès amb un fill menestral o el contrari. Aquests resultats aporten encara més evidències al fet que en un sistema d'herència indivisible com el que existia a Sabadell i a la resta de Catalunya, determinava el destí dels hereus. En canvi, les famílies per poder situar socialment els altres fills/es no hereus, i sobretot aquells que eren més joves que els hereus i que possiblement ja no tenien els pares vius en el moment del casament com a conseqüència de la baixa esperança de vida del moment, compensaren aquest fet a través d'una estratègia de diversificació econòmica i així ostentaren una ocupació o posició socioeconòmica diferent a la dels seus pares (Pujadas-Mora et al., 2018; Ferrer, 2010; Barrera, 1992). D'aquesta manera, més que una competència entre germans/es s'observaria una cooperació de tots els membres de la família.

Cal destacar un comportament propi en el cas dels menestrals. Aquest col·lectiu no era homogeni ni socialment ni econòmicament sinó més aviat era divers, ja que estava format per diferents ocupacions pertanyents a diferents gremis i en diferents branques productives amb diferents estatus. Per aquest motiu s'ha dut a terme un anàlisi propi dels residus estandarditzats. D'aquesta forma s'han tingut en compte tots els pares menestrals amb fills també menestrals encara que diferenciant per les principals ocupacions artesanes existents a Sabadell entre 1545 i 1645 (Taula 5.5). La transmissió intergeneracional de la mateixa ocupació tenia una intensitat molt més alta del que s'esperaria en una distribució a l'atzar, entre vuit o nou vegades més freqüent del que s'esperaria. En aquest sentit, es pot argumentar que la transmissió intergeneracional de la mateixa ocupació sempre serà major en qualsevol cas per una qüestió relacionada amb la lògica del sistema de l'hereu únic. Així, els hereus sempre tendirien a tenir la mateixa ocupació dels seus pares. De tota manera, entre els paraires, es pot observar que la intensitat de l'heretament ocupacional era menor que en la resta d'ocupacions menestrals, la qual cosa es pot explicar per la posició preponderant en l'economia catalana d'aquests a partir del segle XVII, ja que per a aquest grup ocupacional seria més fàcil col·locar socialment als seus fills en grups de major estatus (Torràs, 1984).

Taula 5.5. Residus estandarditzats sobre la freqüència observada i esperada de transmissió del estatus socio-ocupacional a Sabadell segons la ocupació de pares i fills menestrals, 1545-1645

1545-1645		FILL						Sabater
		Paraire	Teixidor	Sastre	Fuster	Ferrer	Mestre de Cases	
P A R E	Paraire	3,60	-3,79	-0,78	-1,71	-1,71	-1,87	-1,70
	Teixidor	-4,01	9,38	-1,79	0,11	0,11	-1,03	-1,19
	Sastre	-1,93	-0,73	8,26	-0,46	-0,46	-0,51	-0,58
	Fuster	-1,18	-0,52	0,47	9,16	-0,42	-0,46	-0,53
	Ferrer	-0,53	-1,22	-0,75	-0,39	9,76	-0,43	-0,50
	Mestre de Cases	-0,01	-1,93	-1,18	-0,62	-0,62	8,12	0,48
	Sabater	-1,53	-0,27	0,10	-0,50	-0,50	-0,55	8,82

Font: Elaboració pròpia a partir de la BHMD.

Finalment, veiem que entre totes les ocupacions menestrals no eren rares les ocasions en què un fill de menestral, encara que no tingués la mateixa ocupació del seu pare, també tingués la condició de menestral, un fet que s'explica per una xarxa social concomitant en l'entorn productiu dels artesans. En aquest sentit, solament la combinació de pares/fills paraires i teixidors era molt menys intensa del que s'esperaria en una distribució aleatòria, la qual cosa podria recordar els conflictes històrics a Sabadell entre els gremis d'aquestes dues ocupacions tèxtils (Benaul, 2009).

5.5 Conclusions.

L'estudi sobre la transmissió intergeneracional de la posició social i l'ocupació a Sabadell en els segles XVI i XVII ens mostra la importància que tenia la família, tant en un sentit vertical (pares) com horitzontal (germans/es), en la reproducció social dels individus. Aquesta transmissió intergeneracional de la posició socioeconòmica a Sabadell era alt, és a dir, específicament un 75% dels fills tenien la mateixa ocupació que els seus pares, una xifra que també s'observa a altres indrets d'Europa al mateix període.

Pel que fa a la intensitat de la transmissió intergeneracional segons l'ordre de casament dels fills, s'ha pogut comprovar (a partir dels residus estandarditzats) que aquells que es casaven primer tenien una propensió molt més alta d'heretar la mateixa posició social del pare, així com una propensió molt més baixa de mobilitat social intergeneracional. En

aquest sentit, tenint en compte que generalment els primers en casar-se solien ser els hereus, es confirma la tendència de què les diferents activitats econòmiques familiars (la terra pels pagesos o els tallers pels artesans) esdevenien una responsabilitat principalment d'aquells fills que es casaven primer.

Pels altres germans que es casaven en segon o tercer lloc, la lògica del sistema de l'hereu únic, feia que en moltes ocasions els altres germans haguessin de cercar altres sortides socioeconòmiques, un fet que més que casual podia ser una important decisió estratègica per tal de diversificar l'economia familiar. Per això, es donà el cas de què el fill d'un pagès es fes artesà i viceversa, però tal diversificació no estava lliure d'obstacles en els dos grups social majoritaris (pagesos i artesans). D'una banda, cal tenir en compte que el fet de què un fill de pagès fos artesà comportava haver de bregar amb les rígides regulacions gremials i la manca d'una xarxa en la menestralia. Per l'altra banda, en el cas dels fills d'artesans l'accés a la terra fou difícil com a conseqüència d'un mercat de terres molt minso, el que va fer que el preu de les terres s'encarís. A més en el cas específic de Sabadell, en aquest període hem pogut veure que la xarxa menestral no sempre operava de forma positiva ja que els resultats dels residus estandarditzats ens ensenyen una possible conseqüència d'una relació històricament conflictiva entre paraires i teixidors. En aquest sentit, encara que aquests fossin els grups menestrals més comuns, els paraires tendiren a evitar tenir fills teixidors i viceversa.

6 Assaig 4. What counted most for secularization: Politics, economy or family? Shifting seasonality of marriages in the Barcelona area, 1715-1880.

6.1 Introduction.

In Western Europe, secularization was part of modernization process together with the transition from an agrarian to an industrial economy. This process brought along urbanization, democratization, rationalization and individualization (Ester et al. 1993; Verweij et al., 1997; Clark, 2012), spreading mostly since the French Revolution²⁰. In this sense, the increment of marriage frequencies in Lenten days, and by extension a shift on seasonal pattern of weddings, has been seen as an indicator of the evolution of secularization as Lesthaeghe and Wilson (1986) pointed out²¹, capturing those different aspects surrounding it. Lent is a forty days' season of prayer, fasting, sacrifice, penance and almsgiving in preparation before the celebration of Easter. Although the Catholic Church did not prohibit the celebration of marriages during Lent, they were discouraged

²⁰ This is the classical version of the Secularization theory that was conceived as a universal theory, but its validity has only been proofed for Europe: Olivier Tschannen, *Les theories de la secularisation*, (Geneve, 1992). However, this pathway is not valid worldwide; multiple modernizations and secularizations and even a current religious revival have been seen José Casanova, "Rethinking secularization: A global comparative perspective", in Peter Beyer and Lori Beaman ed., *Religion, globalization, and culture*, (Brill, 2007), 101-120. Roland Robertson, "Global millennialism: a postmortem on secularization", in Peter Beyer and Lori Beaman ed., *Religion, globalization, and culture*, (Brill, 2007), 7-34. Peter Berger ed., *The Desecularization of the world: Resurgent religion and world politics*, (Grand Rapids, 1999). Pippa Norris and Ronald Inglehart, *Sacred and secular: Religion and politics worldwide*, (Cambridge, 2011). Moreover, a marginalization and privatization of religion was observed but it was not driven out of societies, being more how forms of religion change David Martin, *On secularization: Towards a revised general theory*, (Farnham, 2017). Charles Taylor, "Religious Mobilizations", *Public Culture*, 18, 2 (2006), 281–300.

²¹ We do not analyse the change of the marriage calendar within Advent due to in Spain it was not a widely followed tradition as several studies have pointed out Eiras Roel, "Demografía rural en la España moderna: Evolución, variants y problemas" in Francisco José Aranda ed., *El mundo rural en la España Moderna* (Albacete, 2004). Àngels Torrents, "Actitudes públicas, actitudes privadas", *Boletín de la Asociación de Demografía Histórica*, 10,1 (1992), 7-29).

until the Second Vatican Council (1962-1965) (Aubert, 2002; De la Cueva & Montero, 2007; Lesthaeghe & Wilson, 1997). Simultaneously, it is a straightforward indicator due to the generalization of the date of celebration in the marriage record either in parish books for preindustrial periods and subsequently in the civil register (Willigan & Lynch, 2003).

All around Europe, studies on the shifting in marriage seasonality within Lent have a longstanding tradition. Some decades ago, Jacques Houdaille and Jean Pierre Sardon showed how the French Revolution modified the wedding calendar in France at the end of the eighteenth century (Houdaille, 1978; Sardon, 1979). This process of change was also observed in Belgium by Ron Lesthaeghe, with differences between Catholics and Protestants along the nineteenth century (Lesthaeghe, 1989). Lesthaeghe and López-Gay observed similar behaviors in the twentieth century, which includes the Spanish case (Lesthaeghe & Lopez-Gay, 2013). More recently, Matsuo Hideko and Koen Matthis for nineteenth - twentieth centuries in West Flanders, Belgium, concluded that the industrial model was not the only cause of the shift in the marriage seasonality; this was also a product of an ideology change over time (Matsuo & Matthys, 2018, 2019). In the case of The Netherlands, Frans Van Poppel relates a transformation of marriage seasonality due to changes in production models and lifestyles (Van Poppel, 1995). Lately for seven Dutch provinces, Theo Engelen show the importance of religion in shaping the marriage calendar until the twentieth century (Engelen, 2017). For Italy, Gabriele Ruiu and Marco Breschi concluded that, in spite of the process of modernization of Italy since its unification, there is still nowadays a dissuasive effect of celebrating marriages during Lent (Ruiu & Breschi, 2015), which coincide with conclusions of the studies by Coppa et al. or Emmanuel Sanna and Maria Enrica Danubio for different Italian regions (Coppa et al., 2001; Sanna & Danubio, 2008). For its part, Enzo Lucchetti et al. observed how marriage seasonality in several towns in France, Italy and Spain varied along the last two centuries due to changes in religion, work and climate (Lucchetti et al., 1996).

In the case of England, Sweden or Germany, in order to understand the processes of modernization more emphasis have been given to by economic factors than to religious ones. Good examples are the seminal study by Ann Kussmaul on preindustrial England or in Martin Dribe and Bart Van de Putte or Christer Lundh for Scania (Sweden) and John E. Knodel in his classical work on fourteen German Village (Kussmaul, 1985; Dribe & Van de Putte, 2012; Lundh, 2003; Knodel, 1988. To this effect, December showed a

steady increase in the number of marriages, which seemed to not be related to Advent but more linked to industrialization and implementation of a new working schedule, favouring days off around Christmas. Besides, these differences between regions mostly catholic or protestant are better observed in places where both religions coexist as in Germany (Knodel, 1988).

In the case of Spain, it is worth mentioning the work from Àngels Torrents on the long term analysis of the marriage calendar for an industrial town in the area of Barcelona, in which the calendar during the seventeenth and eighteenth centuries was determined by agricultural and religious cycles, restrictions that disappeared as the nineteenth century progressed (Torrents, 1992). In the same way, José Miguel Martínez-Carrión, in his study on the South East of Spain, concluded that, in spite of the conditions of the agricultural calendar, the ecclesiastical restrictions on marriage during Lent were significant (Carrion, 1984). Similarly, González-Martín has also shown how the number of marriages during Lent in Andorra increased over time between 1606 and 1960 (González-Martín, 2008).

However, despite the long tradition of the analysis of marriage seasonality in the historical literature, as we have briefly introduced, an economic and social view usually prevailed, leaving out politics and cultural transmission within family as explanatory factors in the statistical analysis²², which appear to be key in assessing the secularization of European societies (Clark, 2019). To this effect, politics capture the degree of separation of church and state being a cornerstone for the European secularization, particularly within Liberalism, the new political doctrine expanding across Europe during the nineteenth century (Didural, 2020; Latreille & Cunneen, 1963; Gorski & Altinordu, 2008). In fact, the current variation across European countries in terms of secularization can be explained as a product of the historical patterns of church-state relations (Casanova, 2007).

Moreover, considering the attitudes, values and behaviours are acquired by instruction or imitation mainly within families (Cavalli-Sforza et al., 1982; Albanese, et al., 2016; Soehl, 2017), the intergenerational transmission of sociodemographic behaviours needs to be considered in any research on marriage patterns, a novel approach not contemplated

²² Some of the studies mentioned above point to the relaxation of ecclesiastical norms or changes in ideology as well as possible causes of the change in the marriage calendar. However, these determinants are not included in the statistical analyses, they are only mentioned in the narrative part of the studies.

previously in the literature marriage seasonality. However, their transmission is not absolute due to each generation needing to adapt to their own time (Schönpflug, 2001; Trommsdorff, 2009). At the same time, the intensity of the transmission within families could depend on gender, birth order, sibship size, social group and the inheritance system as in the intergenerational transmission of social outcomes (Pujadas-Mora et al, 2018; Kolk, 2014; Conley & Glauber, 2008). Besides, during the nineteenth century a progressive loss of the parent's influence on children's social adscription (Knigge et al., 2014; Dribe et al. 2015; Zijdemans, 2009) and the reinforcement of horizontal ties has also been argued an aspect that could also have effects on the nuptial calendar (Boudjaaba, et al. 2010; Johnson & Sabeau, 2011).

This article aims at analysing how politics, economic transformation due to industrialization and family determined the abandonment of the traditional nuptial religious calendar. It will help us to understand the secularization process in the Barcelona area during the eighteenth and nineteenth centuries. For doing so, we use a unique database, the *Barcelona Historical Marriage Database*. The area of Barcelona showed an early industrialization process since the end of the eighteenth century. It was featured by an important decrease of the primary sector and the integration of the factory system in the secondary sector as in the case of the textile activities, transforming the occupational structure, becoming one of the first industrialized zones in Southern Europe (Brea-Martínez & Pujadas-Mora, 2018; Nadal, 1975; Martínez-Galarraga & Prat, 2016).

Although the technical changes inducing the agrarian increase on productivity only arrived in Catalonia in the twentieth century, the growth of agricultural commodities in the nineteenth century was possible thanks to the expansion of croplands, the specialisation on vineyards and the adoption of new cropping (Garrabou & Pujol, 1987). As well, this area faced an early fertility decline within the southern European regions initiated in the first half of the nineteenth century (Cabré, 1999; Weir, 1993; Coale & Watkins, 1997), which has also been considered as an insight of secularization (Norris & Inglehart, 2011). Additionally, the Barcelona area also had episodes of anticlericalism. It happened mainly after 1835, including the burning of convents (Alonso, 2003; Portet Pujol, 2003), the diffusion of anarchist ideology, having a strong importance in the late nineteenth century, and the relaxation of religious norms, moral decline or non-compliance with religious practices, all aspects seen during the nineteenth century (Ealham, 2004; Smith, 2007; Tavera, 2002; Prada Moliner, 2016; González-Bailon &

Murphy, 2013). Altogether, these could be also contributing elements but that would not explain completely secularization.

6.2 Data & Methods.

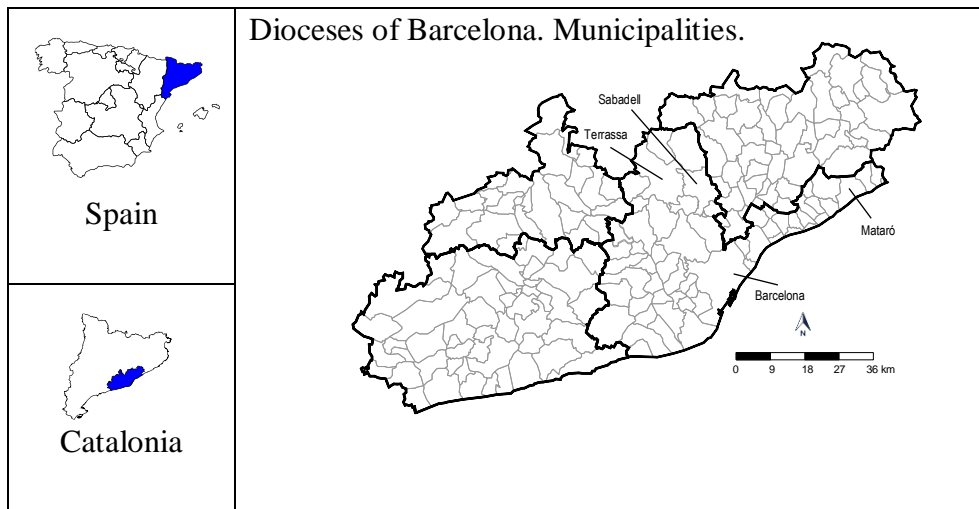
Data.

We have used the Barcelona Historical Marriage Database (BHMD)²³, which gathers more than 600,000 marriage licences registered in the so-called *Llibres d'Esposalles* (Marriage Licence Books) from 1451 and 1905 for all the territories included in the Diocese of Barcelona, in which the *Oficialitat de Barcelona* or Barcelona area was the main part. The Marriage Licence Books are a fiscal source that registered a fee applied to all the marriages celebrated in the Diocese (around 250 parishes in 1900) according to the grooms' occupation or social position. The right for levying this tax was originally granted by Pope Benedict XII (1328-1423), the so-called Pope Luna, an Antipope from the Avignon Papacy. Its main purpose was funding the Barcelona cathedral's construction and maintenance. For this article we will use the marriages licences recorded between 1715 and 1880 due to the objective of the article, which accounted for 284,122 marriages. In these licences the date of payment, which almost coincide with the date of marriage celebration²⁴, the names and surnames of the spouses and their progenitors, the spouses' marital status, the date and parish of celebration and the groom's occupation were generally recorded. During the studied period, the taxes were recorded and ranged among seven socioeconomic categories, varying from the titled nobility in the top to peasants and artisans in the bottom of the categorization. Additionally, there was an exemption fiscal level known as *Amore Dei* devoted to those marriages unable to pay the marriage licence and therefore considered poor. Although, in these cases the groom usually declared to have a trade.

²³ The BHMD was built within the 'Five Centuries of Marriages project', an Advanced Grant from the European Research Council (ERC-2010-AdvancedGrant-269796) directed by Prof. A. Cabré.

²⁴ The date of the marriage licenses was in 90 per cent of cases at most within the first 3 days after the marriage date of celebration. This figure has obtained comparing the dates of the marriage licenses and of the Civil Register for 1860 and 1870 due to historical parish registered in Catalonia suffered important destructions mostly along the twentieth century.

Figure. 6.1. The Barcelona area in the Spanish and Catalan context.



The study area focuses on the so-called *Oficialitat de Barcelona* (Barcelona area), the most populated deanship among the four that compounded the Diocese of Barcelona. The area included cities as Barcelona, Mataró, Sabadell and Terrassa that since the mid-eighteenth century lived a process of protoindustrialisation and early industrialisation in the nineteenth century until turning into the most industrialised area in Spain (Torras, 1998; Nadal, 1978) (See Figure 6.1). The rest of the deanships present important information gaps, which discourages its use.

Data harmonization and codification have been applied to the BHMD. To this effect, the occupational information regarding grooms has been codified through the Historical International Classification of Occupations (HISCO) (Van Leeuwen et al, 2002). Afterwards, occupations were converted into social groups with the Historical International Social Class Scheme (HISCLASS) (Van Leeuwen & Maas, 2011). The twelve categories gathered in HISCLASS have been regrouped in 6 classes given the socioeconomic characteristics from the Barcelona area (See table 6.1).

The geographical locations have been coded following the municipal base of 1860 and geo-referencing all the parishes within the correspondent municipal borders. In this way, we have grouped all the rural parishes as well as the most important urban areas at that time. Additionally, in order to identify the possible effects from the different political periods on marrying or not in Lent, three political categories have been created. The first category gathers the years considered as part of the Old Regime, while the Liberal

political period in the nineteenth century has been divided into two categories, one of them referring to conservative options and the other for progressive ones.

Table 6.1. Codification and classification of the main occupations and social positions in the Barcelona area (1715-1880).

OCCUPATION / SOCIAL POSITION	HISCO	HISCLASS	HISCLASS ADAPTA'
ability*; Government Officials	20210	Higher Managers (1)	Elite (1)
Jurist, Physician	12000, 6105	Higher Professionals (2)	Liberal Professionals
Merchants, Retail traders	41020, 41030	Lower Managers (3), Lower Professional and clerical (4), Lower clerical and Sales (5), Foremen (6)	Trade (3)
Weaver, Tailor, Blacksmith	75400, 79100, 83110	Skilled workers (7), Lower skilled workers (9)	Artisans (4)
Peasant, Farm Servant	61110, 62120	Farmers (8), Lower skilled farm workers (10)	Farmers (5)
Labourer, Factory worker	99910, 99930	Unskilled workers (11), Unskilled farm workers (12)	Labourers (6)

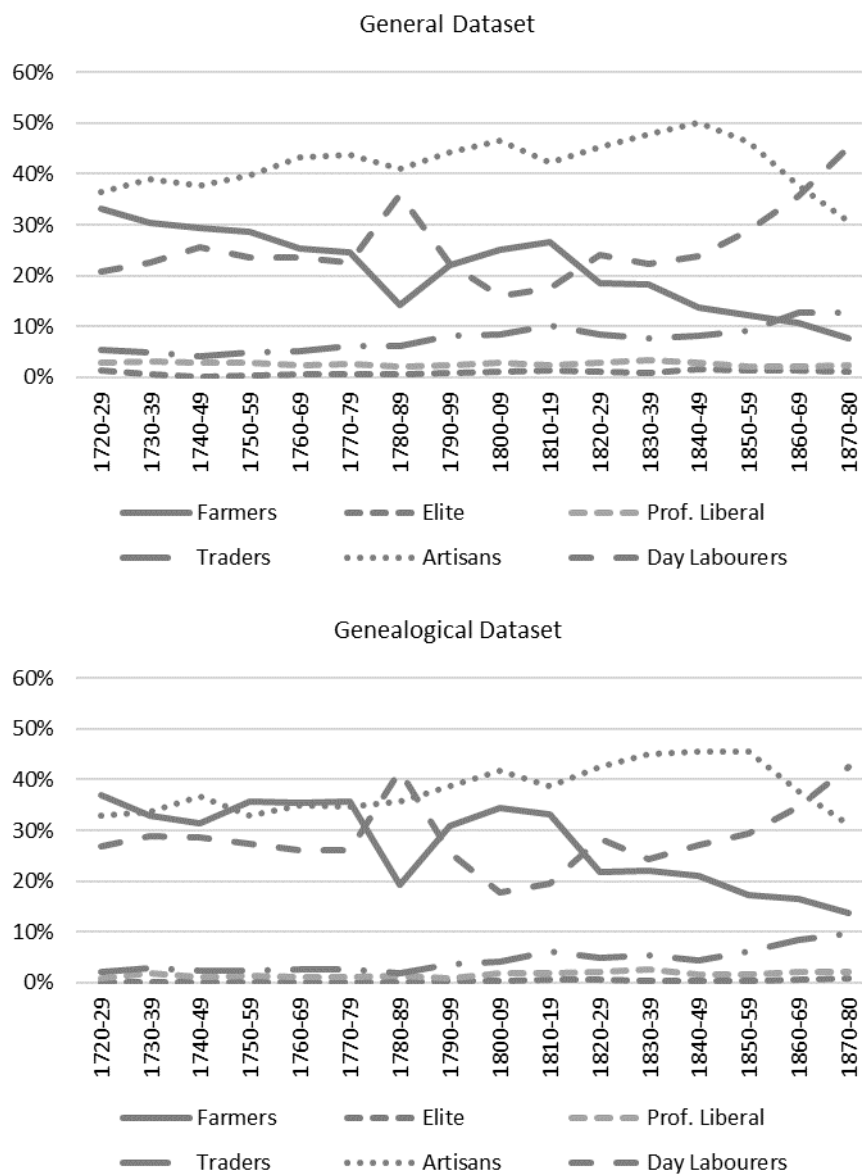
The Old regime gathers all the years between 1715 and 1807 coinciding with the beginning of the Peninsular War (1808-1814). The Liberal Progressive periods would encompass the *Cortes of Cádiz* period in 1810, which is considered the take-off of the Liberal revolution in Spain, the Liberal Triennium (1820-1823) and the more progressive and liberal governments during the minority of Isabella II (1833-1843). During these periods the abolishment of the Holy Inquisition was decreed (even if briefly restored between 1814 and 1820 during the years of the Absolutism establishment), confiscation of ecclesiastical properties as the one of promoted by the prime minister Juan Álvarez Mendizábal in 1836-37 and the tithes' suppression also in 1836 were observed. Moreover, during the so-called Six-democratic (*Sexenio Democrático*) years (1868-1874) that came after the revolution of 1868, there was the implementation of the religious freedom and the suppression of Jesuits from many institutions as well as an important debate about civil marriages, the secularization of cemeteries and the reform on the Church's privileges.

There were also periods in the nineteenth century when conservative options ruled the political power, even restoring absolutism. Such periods were the restoration after the Peninsular war (1814-1820), in the so-called Ominous Decade (1823-1833) which re-

established King Ferdinand VII. The period 1843-1868 also was dominated by political conservatism, during the reign of Isabella II and mainly in the so-called Moderate Decade, when the relations with the Holy See were resumed through the signing of the Concordat in 1851, declaring the confessional status of Spain, which still exists nowadays. Finally, the last period encompassed within the Conservative years is the so-called Bourbon Restoration (1874-1931) that with the constitution from 1876 implemented the religious tolerance albeit keeping the catholic characteristic of the State.

The nominative information contained in each marriage licence allows genealogical reconstitutions, a useful tool for testing the existence or not of intergenerational transmission of behaviours regarding marriage within families. Besides, only through genealogical data, it is possible to encompass the dynamics of the members of the same family on marriage seasonality and the intergenerational transmission of the nuptial calendar. In this way, a subset of families was reconstructed through the nominal record linkage of parent's marriages with children's ones (see Figure 6.2). The assumption behind it stems from the idea that the same individual may appear as a spouse, prior to being a parent in their children's marriages. Therefore, this reconstitution was carried out using the software *BuscaDescendències* (Looking for offspring), developed into the "Five Centuries of Marriages" project. The software runs an algorithm with three similarity measures between strings taking into account the length of each string and the position of each letter. The string comparison is based on the "Levenshtein Distance" and a combination of the so-called "Bag Distance" and "Longest Common Substring distances". Additionally, the algorithm is adapted to Catalan phonetics. At the same time, restrictions for the linkage were imposed taking into account data plausibility between marriages, setting a minimum difference of 15 years between the marriage of parents and children and a 50-years maximum difference in the extreme cases. Record linkage procedures can lead to the presence of overlinks, features likely to occur because the nominal information linked deal with many common Catalan and Spanish names and surnames, which may connect unrelated individuals sharing the same names and surnames (Pujadas-Mora et al., 2018). The total number of links generated automatically surpassed 400,000 in a period that comprised the years of 1715 and 1880.

Figure 6.2. Social group structure in the total of cases and by children’s occupation in the genealogical sample (1750-1880).



In order to ensure the most reliable data as possible, several steps were used after the automatic linkage process minimising the consequences from overlinks. We also carried out a complete check procedure with historical and probabilistic criteria. It was done in order to obtain the most feasible linkage considering the geographic and time distance between marriages of parents and children and their social distance. The spatial distance between the place where the marriages of parents and children took place should follow a logical sequence according to the mobility patterns of the study period. The occupations of parents and their descendants and the amount of tax paid in each marriage should not

be should not differ more than two tiers in the fee scale described. So, 98,991 valid links were verified in the total number of father–children matches.²⁵

The choice for families with at least two siblings or more was done in order to better capture common dynamics regarding family influence and to test for aspects such as sibship size and sibling’s marriage order. Hence, we show results in this article only for family reconstitutions with two siblings or more sharing the same parents, dropping the total number of marriages used in the analyses to 60,984. However, analyses on the probability of marrying in Lent have been run as robustness checks (not reported here), giving similar results as the ones shown in the paper.

A total amount of 23,439 different “families” with at least 2 children from 1750 to 1880 was analysed (See table 6.2). We should bear in mind that the risk of having selection effects always exists when dealing with nominal linked data (Knigge et al., 2014). However, it is worth considering that the socioeconomic composition of the genealogical source coincides mainly with the structure of the general database gathering all cases. The trends in labour market evolution are the same, which could ensure that we are not incurring in deep issues of a selection effect (See Figure 6.2).

Table 6.2. Number of families per number of Siblings.

Number of married siblings per family	Number of families
2	14,099
3	5,830
4	2,285
5	820
6	283
7	84
8	28
9	8
10	2

²⁵ For more details on the record linkage procedures, please see: Francisco Villavicencio, Joan Pau Jordà, and Joana Maria Pujadas-Mora. “Reconstructing lifespans through historical marriage records of Barcelona from the sixteenth and seventeenth centuries” in Gerrit Bloothoof, Peter Christen, Kees Mandemakers and Marijn Schraagen, *Population reconstruction*, (Cham, 2015), 199-216.

Another data limitation that should be taken into account in this study is a sample-selection issue, mainly regarding migrants. In this way, we cannot observe in intergenerational terms the behaviour towards Lent for individuals that migrated outside of the Barcelona area, which limits results only for those who moved within and whose parents were also from inside this territory, as well as for those who were less mobile. However, the geographic area studied is broad, comprising around 90 municipalities at the end of the nineteenth century, and during the studied period, cities like Barcelona, Mataró, Sabadell and Terrassa acted mainly as attraction poles for migratory flows.

Additionally, unlike civil registers, another important issue regarding the Marriage Licence Books is the lack of information on ages for the spouses, which is obviously a drawback for any demographic analysis, impeding the assessment of this kind of determinant on the marriage trends. Nonetheless, this inconvenience is compensated by the sizable amount of cases in the Marriage Licence Books and their wide temporal and geographical coverage. Besides, as we do have information of the civil status of grooms and brides upon marriage, we use variables differing first married individuals from remarried ones as proxy for age differences, as the latter would tend to be older than the former.

Methods.

In order to cover the aim of the article, we have followed an analytical strategy based on two main steps. The first step dealt with an accurate measurement of marriage seasonality and Lent disrespect in the Barcelona area during the period 1715-1880 using descriptive statistics. The measurement of matrimonial seasonality has been done through two different indices. The first index, the so-called Seasonality Index (SI), captures the levels of seasonality within marriages by means of the monthly observed and expected marriages (Henry, 1980). Moreover, due to the objective of this study is mainly assessing trends of marriages in Lent; we have also computed the Lent Seasonality Index (LSI). As known, Lent is a liturgical period based on the lunar calendar, which implies a yearly variability on its beginning, ranging from the 4th of February to the 10th of March. The LSI takes into account the exact days of Lent in order to avoid using only the month of March, which is frequently used as a proxy to Lent and does not include all the Lenten

days from other months (Van Poppel, 1995; Lesthaeghe, 1989). Both SI and LSI can be formulated as follows:

$$SI = \frac{\frac{Mx}{Dx}}{\frac{My}{Dy}} \qquad LSI = \frac{\frac{Mq}{Dq}}{\frac{My}{Dy}}$$

Where Mx is the number of marriages per month, My are marriages per year, Dx is the number of days of a given month, Dy is the number of days per year, Mq is the number of marriages in Lent and Dq the number of days in Lent.

In the second part of our analytical strategy, we tested the different determinants that could influence the likelihood of marrying in Lent in the Barcelona area. We contrasted those individuals marrying in Lent to those not as the dependent variable, assessing socioeconomic and political period's features for the total cases (1715-1880) and for the demographic and familial effects in the genealogical reconstituted data (1750-1880) (see table 6.3). In this regard, we used two different typology of models. In first place, we ran logistic regressions for a dataset including all the marriages celebrated in the Barcelona area during our studied period. Secondly, we also applied random intercept logistic models (multilevel) taking into account the hierarchical structure of our genealogical data, in order to not underestimate the standard errors of regression coefficients (Hox, 2002). Accordingly, as some individuals belonged to the same families sharing a series of characteristics and constraints that could make them more similarly prone or not to marrying in Lent. Hence, the choice of a multilevel approach allows us to control for between-family variance and statistical non-independence of individuals, usually not controlled in normal logistic regressions.

Among the independent variables, we created a categorical one defining the political period in which individuals married based on the different political phases that alternated in Spain during the nineteenth century (see Data section). In this way, the Old regime was assumed to last until 1807, while the Conservative and Progressive periods alternated along the nineteenth century. In addition, we also added different interactions between the political periods, the individual's social background and their demographic and familial characteristics in order to define if different combinations could multiply or nuance the likelihood of respecting or not Lent when marrying. Finally, as Lenten

marriages would be expected to have a linear trend increase over time, all our models have included decadal dummies to control for time trend.

Tables 6.3. Descriptive statistics

Descriptives (General Dataset)				
Total number of Marriages 1715-1880				
n= 284,122				
	Range		Mean	SD
	Minimum	Maximum		
Marrying in Lent (dependent)	0	1	0.08	0.27
Tax-exempted (<i>Amore Dei</i>)	0	1	0.21	0.41
Widow/er	0	1	0.21	0.41
Social Groups	1	6		
Farmers (1)			0.18	
Elite (2)			0.01	
Liberal Professional (3)			0.03	
Traders (4)			0.09	
Artisans (5)			0.41	
Labourers (6)			0.28	
Barcelona	0	1	0.66	0.47
Years	1715	1880		
Political periods				
Old Regime	1715	1807	0.37	
Liberal Revolution (Liberal)	1808	1814	0.02	
Restoration (Conservative)	1815	1820	0.03	
Liberal Triennium (Liberal)	1821	1823	0.02	
Ominous Decade (Conservative)	1824	1833	0.06	
Isabella II (Liberal)	1834	1843	0.07	
Moderate Decade (Conservative)	1844	1854	0.10	
Liberal Union (Liberal)	1855	1868	0.16	
Six Democratic (Liberal)	1869	1873	0.06	
Borbonic Restoration (Conservative)	1874	1880	0.10	
Political periods categorical	1	3		
Old regime – 92 years (1)			0.36	
Conservative – 30 years (2)			0.32	
Progressive – 34 years (3)			0.32	

Descriptives (Genealogical Dataset)

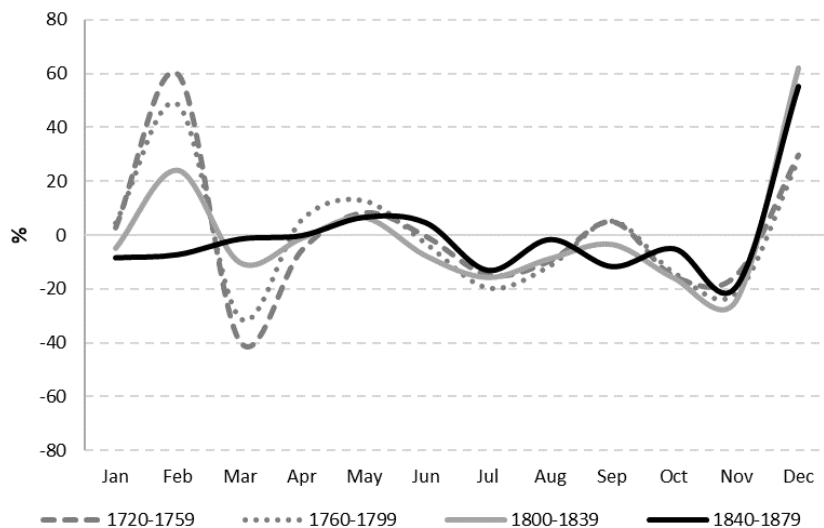
Number of Marriages 1750-1880				
n= 60,984				
	Range or N		Mean	SD
	Minimum	Maximum		
Years Parents	1715	1850		
Years Children	1750	1880		
Political periods Parents				
Old regime (1)			0.51	
Conservative (2)			0.29	
Progressive (3)			0.21	
Political Periods Children				
Old regime (1)			0.32	
Conservative (2)			0.26	
Progressive (3)			0.42	
Marrying in Lent (dependent)	0	1	0.1	0.29
Father marrying in Lent	0	1	0.08	0.27
Difference in years between father's and children's marriage date	11	50	32.34	6.65
Marrying at the same month Parents-Children	0	1	0.1	0.29
Sibship size	2	5	2.91	1.02
Marriage Order	1	3	1.86	0.77
1st			0.38	
2nd			0.38	
3rd or +			0.239	
Female	0	1	0.51	0.5
Intergenerational class mobility	0	1	0.48	0.5
Marrying in a different parish than parents	0	1	0.51	0.5
Father <i>Amore Dei</i>	0	1	0.09	0.29
Children <i>Amore Dei</i>	0	1	0.18	0.39
Social Group Father				
Farmers (1)			0.33	
Elite (2)			0.001	
Liberal Professional (3)			0.01	
Traders (4)			0.03	
Artisans (5)			0.35	
Labourers (6)			0.27	
Social Group Children				
Farmers (1)			0.23	
Elite (2)			0.004	
Liberal Professional (3)			0.02	
Traders (4)			0.05	
Artisans (5)			0.39	
Labourers (6)			0.30	

6.3 Results.

6.3.1 Measuring the likelihood of marrying in Lent in the Barcelona Area.

Since the end of the eighteenth century, the Barcelona area faced the take-off and progressive establishment of industrialization. This is clear when throughout this period the fluctuation of marriages per months changed progressively, turning from a concentration of marriages in February and the avoidance of March during the 1720-1759 to an almost unseasonal marriage calendar except for the concentration of celebrations in December during the 1840-1879 (See Figure 6.3). This change in the marriage calendar towards a higher concentration in December has been observed also in other countries as Sweden (Dribe & Van de Putte, 2012).

Figure 6.3. Index of marriage seasonality by months in the Barcelona area (1720-1880).

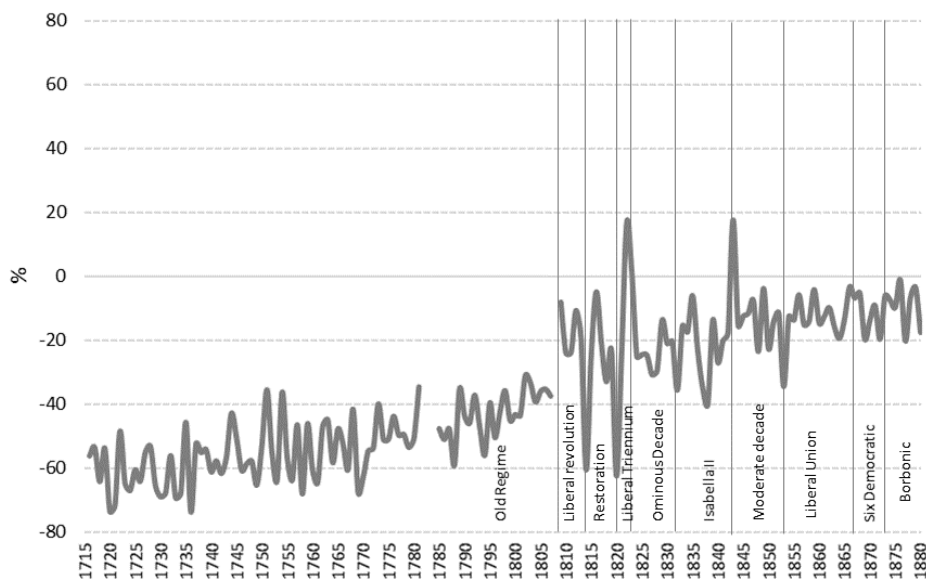


In the long-term, the incidence of marriages in Lenten days increased progressively from the beginning of the eighteenth to the end of the nineteenth century (See Figure 6.4). Particularly, the intensity of growth in the slope for the Lent's index (LSI) became higher from the 1780s onwards, when the occupational structural change resulting from the industrialization of the Barcelona area as it was already noticed. This was the period when for the first time the sector leading the labour market was the secondary sector in detriment of the primary one, a symbolic fact of the Catalan early industrialization (Brea-Martínez & Pujadas-Mora, 2018), which would give insights of the link between

secularization and economic modernization (Lesthaeghe & Wilson, 1997; Dribe & Van de Putte, 2012).

Furthermore, it is possible also to observe different peaks of positive and negative values in the LSI index during the first half of the nineteenth century. For instance, the highest value around 1808 in the first year of the so-called Peninsular War (1808-1814) when the Napoleonic troops invaded Spain or the alternation of conservative and more progressive political options that ruled discontinuously Spain throughout the nineteenth century. In this sense, for the absolutist restoration periods covering 1814-1820 or 1823-1833, the propensity of marrying in Lenten days dropped; whilst during progressive periods like those of 1820-1823 or in 1833-1843, marriages in Lent increased (See Figure 6.4). In fact, this suggests that secularization was also a political issue (Lesthaeghe, 1989; Houdaille, 1978).

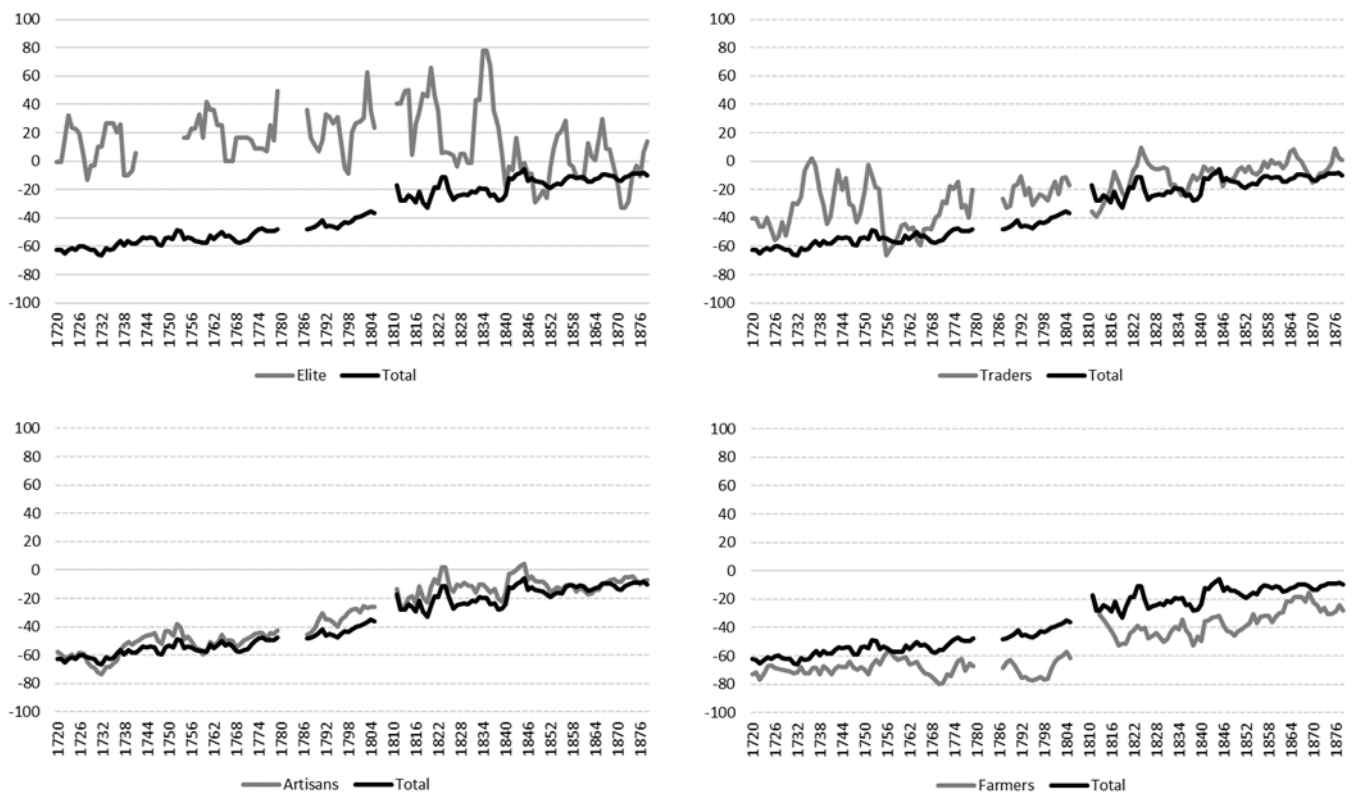
Figure 6.4. Index of relative frequency of Lenten marriages in the Barcelona area (1715-1880)



The elite never showed a seasonal attitude towards Lent, while farmers avoiding Lenten days for marriage was in any period a more likely practice (See Figure 6.5). Traders and artisans had a clear evolving marriage seasonality from avoiding Lenten days for marriage to a more flexible behaviour in the nineteenth century especially in the case of the commercial social group. Although the difference in Lent respect presented by the different social groups, a general increase in the LSI is observed, which means a common

movement towards the disrespect of Lent. All social groups presented an important increment of marriages in Lent since the 1780s, a behaviour presented also by rural and agrarian population (farmers), which may be showing not only an economic-oriented change but also a cultural/religious shift in marriage aptitude, as noted in other rural catholic societies from the Netherlands (Engelen, 2017) (See Figure 6.5).

Figure 6.5. Index of Relative frequency of Lenten marriages in the Barcelona area (1715-1880) by some social groups.



So far, we have only observed the evolving trends on Lenten marriages through aggregate data on social groups. However, in order to disentangle more precisely the determinants of lower or higher propensity of marrying during Lent, it is necessary to go deeper. In this sense, we have used the individual (family) level data from the marriage licence books to model the social and demographic characteristics of those marriages celebrated in Lenten days.

6.3.2 The determinants of the likelihood of marrying in lent in the Barcelona area

One of the main issues when disentangling the determinants behind the increasing incidence of Lenten marriages is discerning between time trends and/or political, cultural or societal factors. On the one hand, as already observed in the descriptive figures above,

Lenten marriages increased as a function of time, as the second half of the nineteenth century shows a higher share of Lenten marriages than the eighteenth century for instance. In this way, the progressive industrialization in the Barcelona area would obviously have played an important role showing that more industrialized areas saw higher likelihoods of marrying in Lent, which may reinforce what we already know about urbanization and industrialization regarding secularization (Van Poppel, 1995; Dribe & Van de Putte, 2012). On the other hand, although industrialization evolved progressively across the nineteenth century, Spanish politics were characterized by a continuous alternation of conservative and progressive periods, which at the same time would be driving the increase on Lenten Marriages, a factor that has not been studied so far in the literature.

We start using the general sample of marriages testing the dependent variable of marrying in Lent by decades (Model 1) and political periods (Model 2). In Model 1 we can see that as time passed by, the likelihood of Lenten marriages increased although not monotonically from decade to decade in the nineteenth century (see probabilities on figure 6.6). Moreover, if we observe the odds ratios in Model 2 regarding political periods (chronologically ordered) it is worthy noticeable that even with a certain increasing trend towards the end of the nineteenth century, earlier liberal periods as the Liberal Triennium showed higher incidences of Lent disrespect (figure 6.7). This suggests that this specific political period could have been the most important one towards the speed of secularization in the long-term. Albeit, the social groups to which grooms belonged were also determinant on the likelihood of getting married in Lent. Farmers were in any model those with less propensity to Lenten unions while better-off groups such as the Elite, Liberal Professionals or Traders showed the highest values.

Finally, we included interactions of the different social groups with a three-category variable regarding the political sign of the different periods -Old regime, Conservative or Liberal options- in order to capture how the political contexts could influence the likelihood of marrying in Lent and if it differed by social groups (see Table 6.4, Figure 6.8). The results suggest for all social groups an important increase in the likelihood of Lenten marriages in both Conservative and Liberal periods in comparison to the Old Regime. However, Farmers and the Elite (characterised by different groups from the Nobility) displayed a clear gradient, increasing the probability of marrying in Lent during Liberal periods in comparison to Conservative times (See figure 6.8).

Table 6.4. Logistic regressions on the likelihood of marrying in Lent by social groups, decades and political periods.

	Model 1	Model 2
Farmers	<i>ref</i>	<i>ref</i>
Elite	1.763*** (0.116)	1.762*** (0.116)
Prof. Liberals	1.605*** (0.0758)	1.603*** (0.0757)
Trade	1.643*** (0.0498)	1.641*** (0.0498)
Artisans	1.531*** (0.0356)	1.531*** (0.0357)
Day Labourers	1.379*** (0.0343)	1.378*** (0.0344)
Old regime		<i>ref</i>
Liberal Revolution		1.469** (0.210)
Restoration		1.449* (0.232)
Liberal triennium		4.017*** (0.813)
Ominous decade		3.013*** (0.605)
Isabella II		3.027*** (0.632)
Moderate decade		2.972*** (0.635)
Liberal union		3.335*** (0.725)
Six democratic		3.001*** (0.678)
Borbonic		3.094*** (0.708)
Decadal Dummies	YES	YES
<i>N</i>	275516	275516

Figure 6.6. Probability of marrying in Lent by decades and social groups

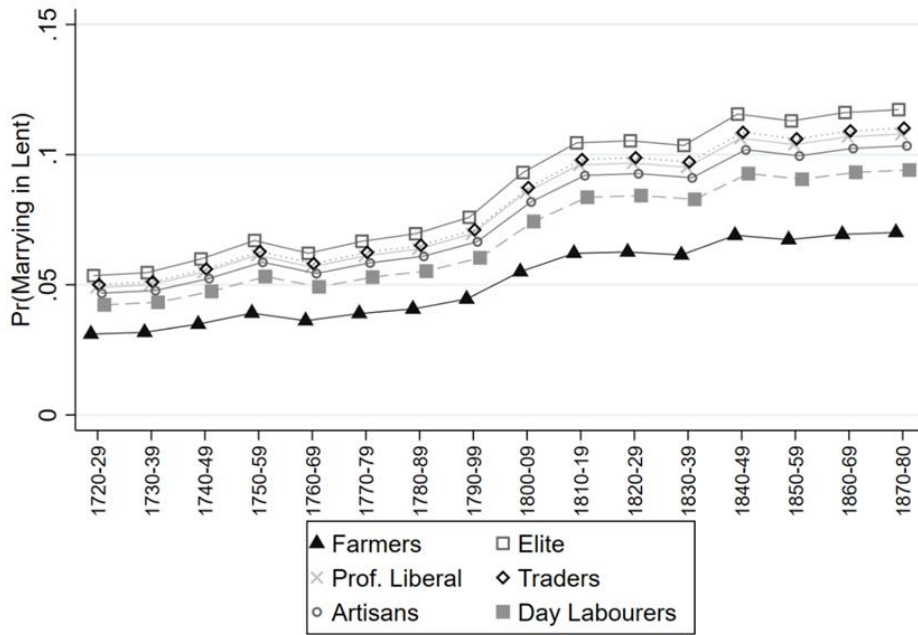


Figure 6.7. Probability of marrying in Lent by political periods and social groups

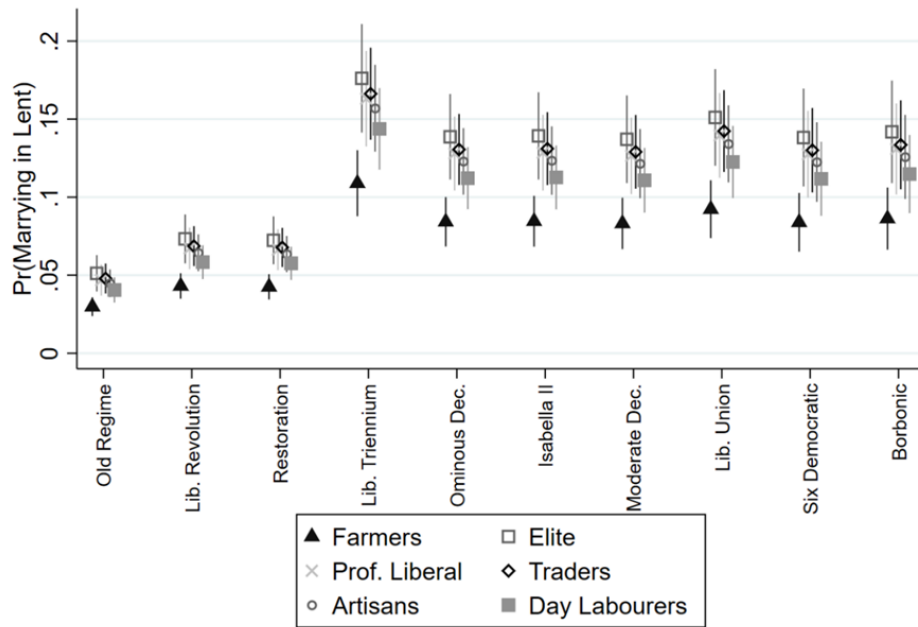
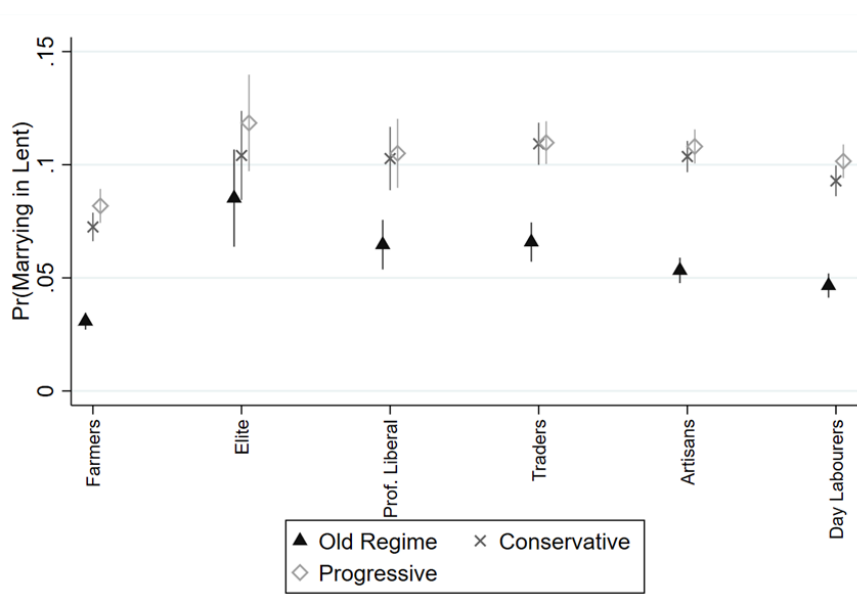


Table 6.5. Logistic regression on the likelihood of marrying in Lent, interactions of social groups and categories of political sign.

	Model 3
Farmers X Old Regime	<i>ref</i>
Farmer X Conservative	2.458*** (0.242)
Farmer X Liberal	2.803*** (0.288)
Elite X Old Regime	2.931*** (0.397)
Elite X Conservative	3.656*** (0.510)
Elite X Liberal	4.230*** (0.581)
Prof. Liberals X Old Regime	2.174*** (0.181)
Prof. Liberals X Conservative	3.602*** (0.421)
Prof. Liberals X Liberal	3.693*** (0.449)
Traders X Old Regime	2.216*** (0.134)
Traders X Conservative	3.859*** (0.388)
Traders X Liberal	3.880*** (0.396)
Artisans X Old Regime	1.770*** (0.0723)
Artisans X Conservative	3.637*** (0.348)
Artisans X Liberal	3.813*** (0.373)
Day Labourers X Old Regime	1.536*** (0.0709)
Day Labourers X Conservative	3.221*** (0.314)
Day Labourers X Liberal	3.556*** (0.350)
Decadal dummies	YES
<i>N</i>	275516

Figure 6.8. Probability of marrying in Lent by social groups and political sign category.



Next, in models 4 and 5 and 6 we add other socioeconomic and demographic variables. Among them we control if individuals marrying were exempted from paying the marriage tax as a proxy of contextual poverty (see the data section), if individuals were first married or remarried, which in the absence of an information in the age of individuals becomes the best proxy for age at marriage, given that presumably remarried individuals would be older. Besides, we also include the social group to which individuals belonged as well as a control for marriages happening within Barcelona and the political period during which marriages were celebrated. In this sense, being exempted from the marriage licence fee (declared *ad Amore Dei*) had a slight effect on increasing the likelihood of getting married during Lent, although it was more important whether widows or widowers were getting married. These two variables *per se* may show insights that in some cases, people with more vulnerable situations as poverty or widowhood could be less selective in choosing a marriage date, or in the contrary also preferred marrying in the periods when celebration were sober, as Lent was expected to be a period when sumptuary behaviours would be avoided. Overall, in model 4, albeit odds ratios point out an effect of urban areas (Barcelona against others) and remarried individuals in the likelihood of marrying in Lent, the differences between political periods (Old Regime vs Conservative and Liberal periods) is still highest.

Table 6.6. Logistic regressions on the likelihood of marrying in Lent by social groups and sociodemographic characteristics.

	Model 4	Model 5 (1807-1880)	Model 6
Farmers	<i>ref</i>	<i>ref</i>	<i>ref</i>
Elite	1.589*** (0.105)	1.406*** (0.108)	1.586*** (0.105)
Prof. Liberals	1.456*** (0.0699)	1.293*** (0.0769)	1.454*** (0.0698)
Trade	1.467*** (0.0462)	1.361*** (0.0519)	1.467*** (0.0462)
Artisans	1.403*** (0.0338)	1.348*** (0.0414)	1.405*** (0.0339)
Day Labourers	1.286*** (0.0329)	1.243*** (0.0398)	1.288*** (0.0329)
Barcelona	1.255*** (0.0212)	1.224*** (0.0240)	1.256*** (0.0212)
<i>Amore Dei</i>	1.072*** (0.0193)	1.010 (0.0213)	1.071*** (0.0193)
Remarried	1.268*** (0.0212)	1.185*** (0.0238)	
Old Regime	<i>ref</i>		
Conservative	2.068*** (0.184)		
Progressive	2.192*** (0.200)		
Conservative		<i>ref</i>	
Progressive		1.067*** (0.0212)	
1 st Marriage X Old Regime			<i>ref</i>
1 st Marriage X Conservative			2.246*** (0.201)
1 st Marriage X Progressive			2.359*** (0.216)
Remarried X Old Regime			1.575*** (0.0496)
Remarried X Conservative			2.564*** (0.236)
Remarried X Progressive			2.822*** (0.265)
Decadal dummies	YES	YES	YES
<i>N</i>	275516	171883	275516

Odds ratios displayed; Standard errors in parentheses

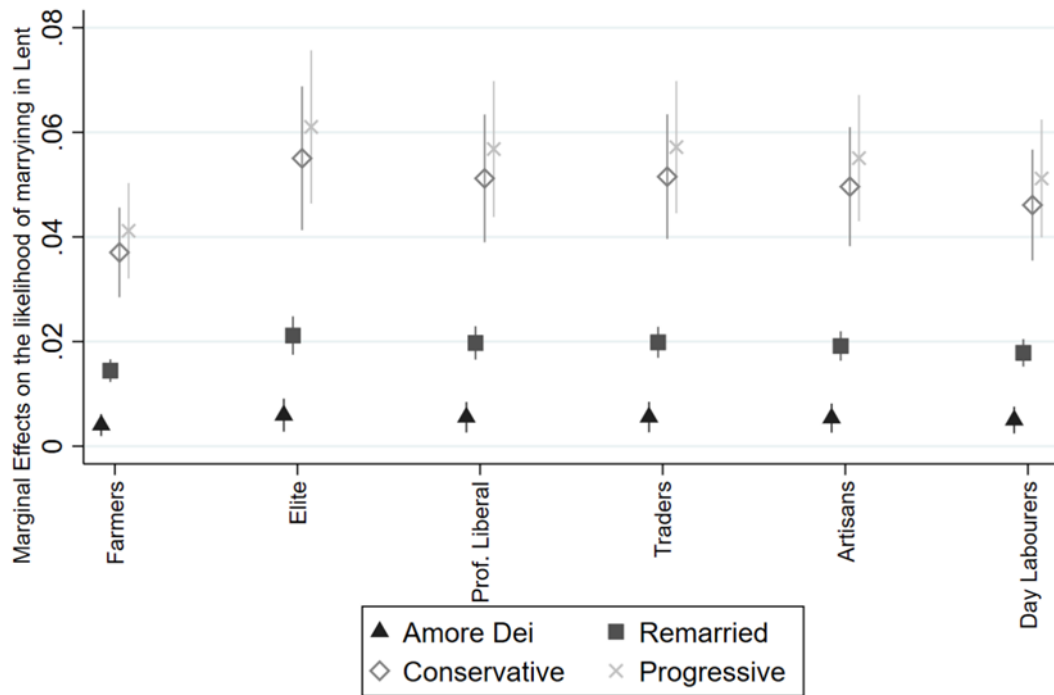
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Author's elaboration

These elements might be observed in figure 6.9 where the marginal effects of those independent variables are plotted and compared. Accordingly, the effects of poverty (*Amore Dei*) on the likelihood of marrying in Lent is really low, as poor individuals were only about 1% (0.01) more likely to have Lenten Marriages than others across any social group. On the other hand, remarried individuals had overall 2% higher chances of Lenten

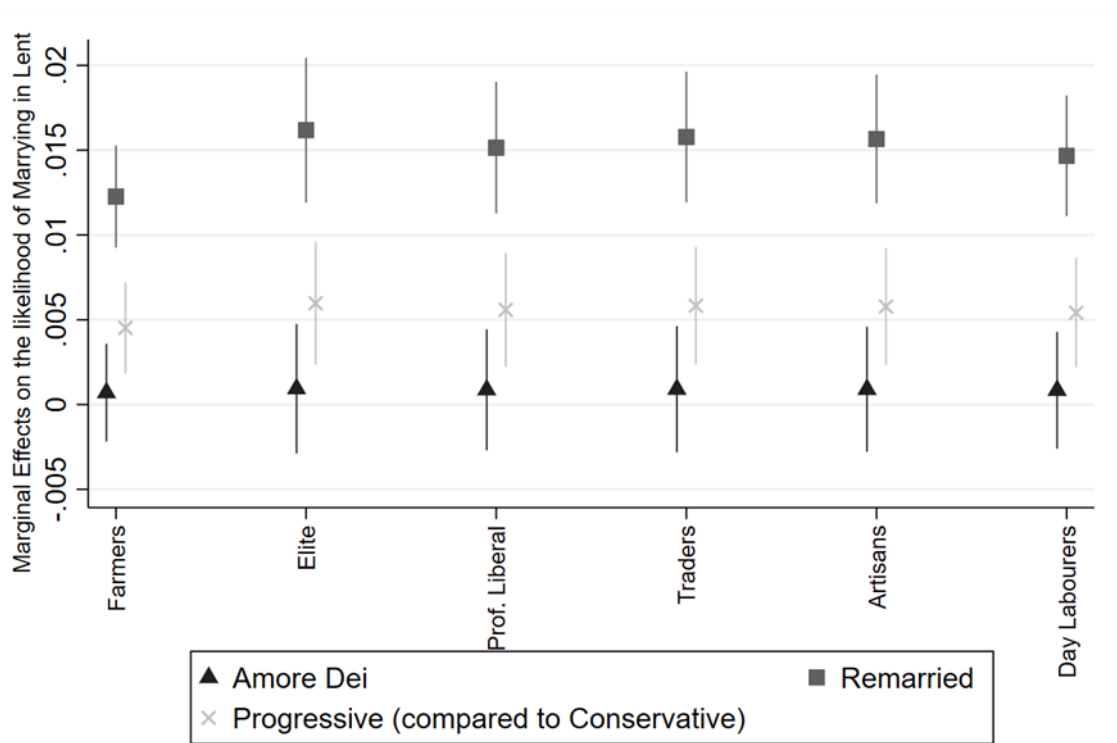
marriages (although a slightly lower effect for farmers). Nevertheless, including the entire dataset (1715-1880) show us that on overall it was mainly the political sign the responsible for differences in the likelihood of marrying in Lent. Couples marrying during conservative and progressive periods had between 4% and 6% more chances of marrying in Lent than individuals getting married during the Old Regime (See figure 6.9).

Figure 6.9. Marginal effects on the likelihood of marrying in Lent (1715-1880).



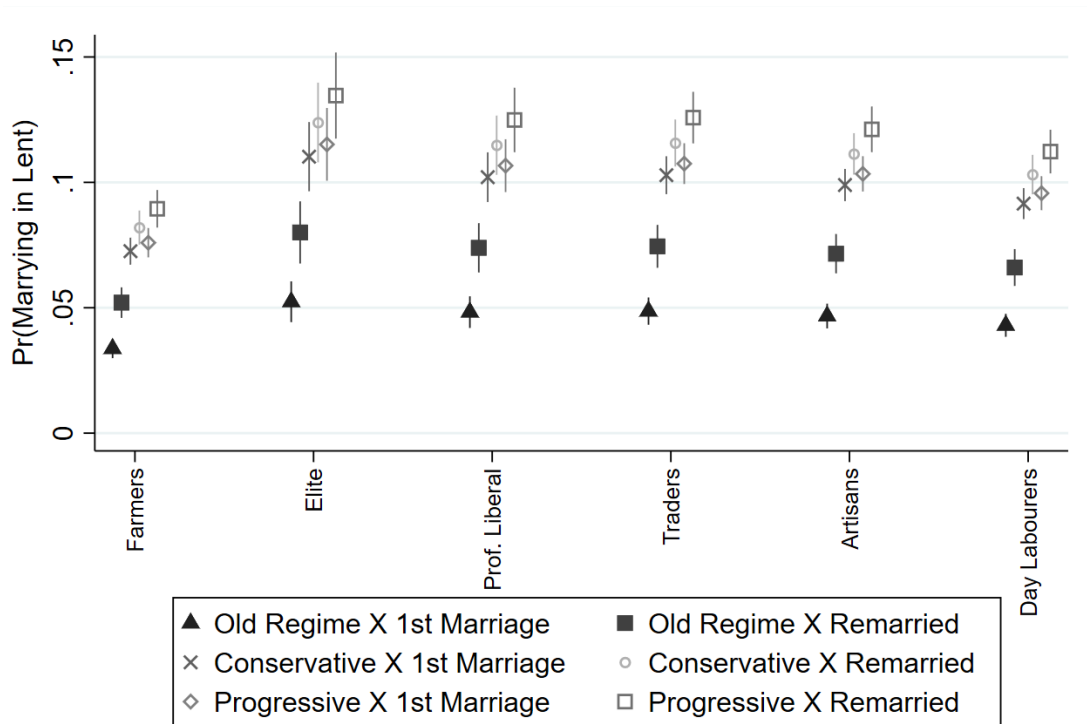
In model 5 (see Table 6), we reduced our sample to include only the periods from 1808 onwards in order to just compare the conservative and progressive periods. As it can be noticed, the social gradient in Lenten Marriages still existed (with farmers with the lowest likelihood), although the differences between groups reduced. In this sense, when plotting the marginal effects only for the conservative periods (See figure 6.10), we may see that the effects of poverty (*Amore Dei*) are even lower. But remarried increased chances of marrying in Lent a 1.5%. On the other hand, individuals marrying in progressive periods were less than 1% more likely to marry in Lent than those in conservative periods. These last results can also be observed in model 6 (see table 6) as we interact the marital status upon marriage with all political periods, which shows that on overall the fact of being remarried counted more for the chances of marrying in Lent.

Figure 6.10. Marginal effects on the likelihood of marrying in Lent (1807-1880).



To sum up, the results above show us that the political periods played a role on increasing Lenten marriages mainly from the first liberal period as the Liberal Triennium (1820-1823), which set a point of no return on declining the levels of marriages in Lent and that affected equally all social groups. This reduction was mainly due to the gap between farmers and the other social groups. Nevertheless, these results suggest that the most important aspect driving differences in chances of marrying in Lent was being widow or widower. This would show a greater haste in getting married, a kind of marriage that frequently was denounced with the noise of cowbells, pots or pans (Mantecón, 2013), due to it is a time without solemnization in the performance of the nuptial mass (Breschi & Ruiu, 2016). Accordingly, the importance of remarriage can be observed in the interaction between the three political signals and civil status (first married and remarried) on model 6. As it can be observed in figure 6.11, the probabilities of marrying in Lent for these interactions show that again although the secularization process started in the liberal age in the nineteenth century was crucial for increasing the likelihood of marrying in Lent across social groups, the incidence of Lenten marriages among remarried individuals was always more important.

Figure 6.11. Probability of marrying in Lent by social groups, interaction of civil status and political sign category.



6.4 Was secularization also a family matter?

Another aim of this article is to estimate how individuals in families were coping with secularization; thus, we investigate if aspects such as marriage order²⁶ or gender affected the willingness to break the traditional marriage calendar, along with the fact of belonging to a particular social group. Moreover, a possible decline of intergenerational transmission of values and behaviors, due to the weakening of family ties as result of the industrialization, migration and urbanization, could also be seen in the increase of marriages in Lent. In this way, we have modelled the likelihood of marrying in Lent using the genealogical database (see Data and methods section), which displays parent-children

²⁶ Given the fiscal nature of the marriage licenses, in which the age at marriages was not reported, when coping with siblings, marriage order has been used as a proxy of birth order. Obviously we have to be cautious when assimilating both orders, although from previous studies in Catalonia it has been seen that heirs, usually the eldest siblings, were the first to marry for instance. Julie Marfany, “Choices and constraints: Marriage and inheritance in eighteenth- and early nineteenth-century Catalonia”, *Continuity and Change*, 21, 1 (2006), 73–106. Pujadas-Mora, Brea-Martínez, Jordà and Cabré ‘The apple never falls’.

links nested in families, applying random intercept logistic models (multilevel) (Hox, 2002).

Table 6.7. Random-intercept logistic regressions on the likelihood of marrying in Lent by social groups and familial characteristics.

	Model 7	Model 8
1 st Married Child	<i>ref</i>	<i>ref</i>
2 nd Married Child	1.006 (0.0416)	1.007 (0.0423)
3 rd Married Child	0.889 (0.0544)	0.883* (0.0548)
4 th or + Married Child	0.891 (0.0751)	0.881 (0.0752)
Parents-Children marriage distance in years	1.010** (0.00339)	1.011** (0.00345)
Sibship size	1.015 (0.0209)	1.033 (0.0216)
Female	0.973 (0.0365)	0.970 (0.0370)
Different Parish than parents	0.897*** (0.0294)	0.928* (0.0312)
Amore Dei (Parents)	1.102 (0.0635)	1.056 (0.0614)
Amore Dei (Ego)	1.103* (0.0449)	1.089* (0.0451)
Intergenerational Social Mobility	1.036 (0.0338)	1.083* (0.0364)
Parents-Children marrying at same month	0.767*** (0.0465)	0.782*** (0.0478)
Parents marrying in Lent	1.153* (0.0751)	
Farmer father X Not marrying in Lent		<i>ref</i>
Elite father X Not marrying in Lent		1.018 (0.256)
Prof. Liberal father X Not marrying in Lent		1.198 (0.207)
Trade Father X Not marrying in Lent		1.397*** (0.140)
Artisan Father X Not marrying in Lent		1.435*** (0.0620)
Day Labourer father X Not marrying in Lent		1.187*** (0.0554)
Farmer Father X Marrying in Lent		0.845 (0.141)
Elite father X Marrying in Lent		2.013 (0.756)
Prof. Liberal father X Marrying in Lent		1.260 (0.676)
Trade Father X Marrying in Lent		1.503 (0.502)
Artisan Father X Marrying in Lent		1.783*** (0.164)
Day Labourer father X Marrying in Lent		1.368* (0.182)
Decadal dummies	YES	YES
Random Intercept at family level	1.248*** (0.0762)	1.247*** (0.0768)
<i>N</i>	60984	59494

Odds ratios displayed; Standard errors in parentheses

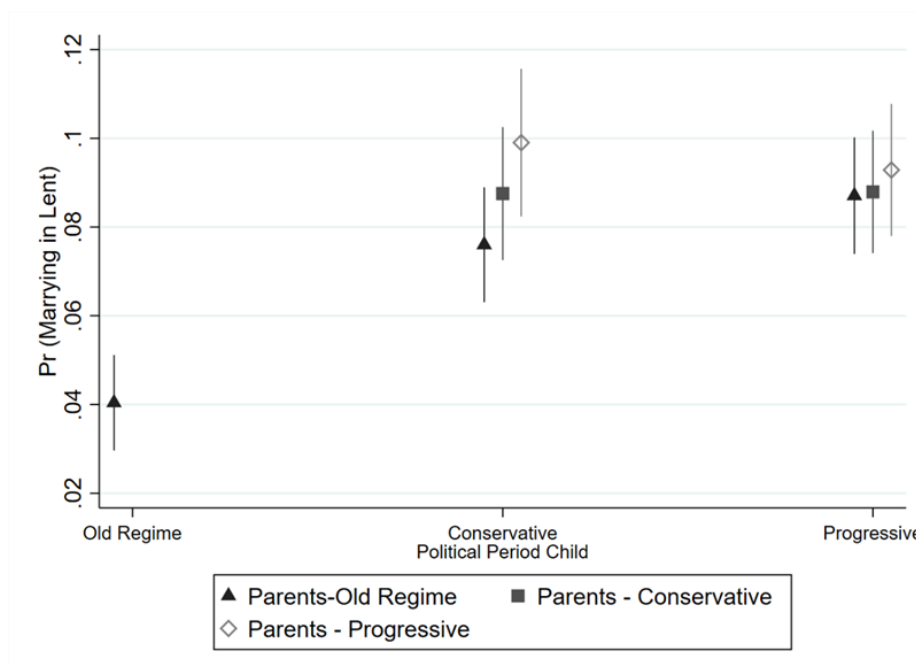
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Author's elaboration

Using the genealogical sample, in Model 7, we have modelled demographic variables like sibship size, the siblings' marriage order and gender, as well as the distance in years between parents and children marriages or if there was intergenerational social mobility between parents and children. In the first place, the likelihood of marrying in Lent was higher if parents did so too.

However, being female or marrying in a different parish than the progenitors (a proxy for migratory status) did not have a positive effect on the chances of marrying in Lent. In this regard, parent's behaviours on overall findings show that were influential for children in the choice of disrespecting Lent.

Figure 6.12. Probability of marrying in Lent by the political sign period of children and political sign period of parents.



Furthermore, considering whether the parents married in Lent along with the social group in Model 8 show that, though not all the interactions were statistically significant, there are substantial findings to argue that marriages in Lent were likely to be intergenerationally transmitted at different levels by social groups. In this way, for most parents' social groups the interaction with having parents who married in Lent increased the likelihood of marrying in Lent too, mainly for artisans. Nevertheless, as the genealogical sample includes both parents and children getting married in all the different political periods (Old Regime, Conservative and Progressive), it could be expected that

the likelihood of getting married in Lent would be higher if parents and children did so during the nineteenth century, for instance in progressive periods.

Hence, in Model 9 (see table 8) we look at periods of marriage for both parents and children, which suggests that those parents who married in progressive periods were more likely to have children marrying in Lent, as was expected. However, given that one of the aims of this study is to shed light on the determinants of secularization through marriage, we argue that the politically conservative periods negatively influenced the rate of marrying in Lent.

Table 6.8. Random intercept logistic models on the likelihood of marrying in Lent by social groups, familial characteristics and political sing categories

	Model 9	Model 10
1 st Married Child	<i>ref</i>	<i>ref</i>
2 nd Married Child	1.005 (0.0417)	1.001 (0.0423)
3 rd Married Child	0.888 (0.0544)	0.884* (0.0550)
4 th or + Married Child	0.893 (0.0753)	0.882 (0.0755)
Parents-Children marriage distance in years	1.011** (0.00354)	1.012*** (0.00359)
Sibship size	1.014 (0.0209)	1.028 (0.0216)
Parents Old Regime X Child Old Regime	<i>ref</i>	
Parents Old Regime X Child Conservative	1.973** (0.411)	
Parents Old Regime X Child Progressive	2.295*** (0.446)	
Parents Conservative X Child Conservative	2.309*** (0.525)	
Parents Conservative X Child Progressive	2.320*** (0.521)	
Parents Liberal X Child Conservative	2.653*** (0.600)	
Parents Liberal X Child Progressive	2.467*** (0.557)	
Different Parish than parents	0.895*** (0.0294)	0.931* (0.0313)
Amore Dei (Parents)	1.102 (0.0636)	1.069 (0.0628)
Amore Dei (Ego)	1.105* (0.0451)	1.103* (0.0469)
Intergenerational Social Mobility	1.034 (0.0337)	0.998 (0.0342)
Female (Ego)	0.975 (0.0367)	0.961 (0.0368)
Parents-Children marrying at same month	0.768*** (0.0465)	0.776*** (0.0477)
Parents marrying in Lent	1.148* (0.0747)	

<i>Continuation</i>	Model 9	Model 10
Child Farmer		1
Child Elite		2.668*** (0.533)
Child Prof. Liberal		1.742*** (0.212)
Child Trader		1.630*** (0.128)
Child Artisan		1.645*** (0.0800)
Child Day Labourer		1.356*** (0.0721)
Child Old Regime X Parents not marrying in Lent		<i>ref</i>
Child Old Regime X Parents marrying in Lent		1.312 (0.253)
Child Conservative X Parents not marrying in Lent		2.240*** (0.458)
Child Conservative X Parents marrying in Lent		2.589*** (0.596)
Child Progressive X Parents not marrying in Lent		2.340*** (0.467)
Child Progressive X Parents marrying in Lent		2.507*** (0.542)
Decadal Dummies	YES	YES
Random Intercept at family level	1.245*** (0.0760)	1.221** (0.0759)
<i>N</i>	60984	58961

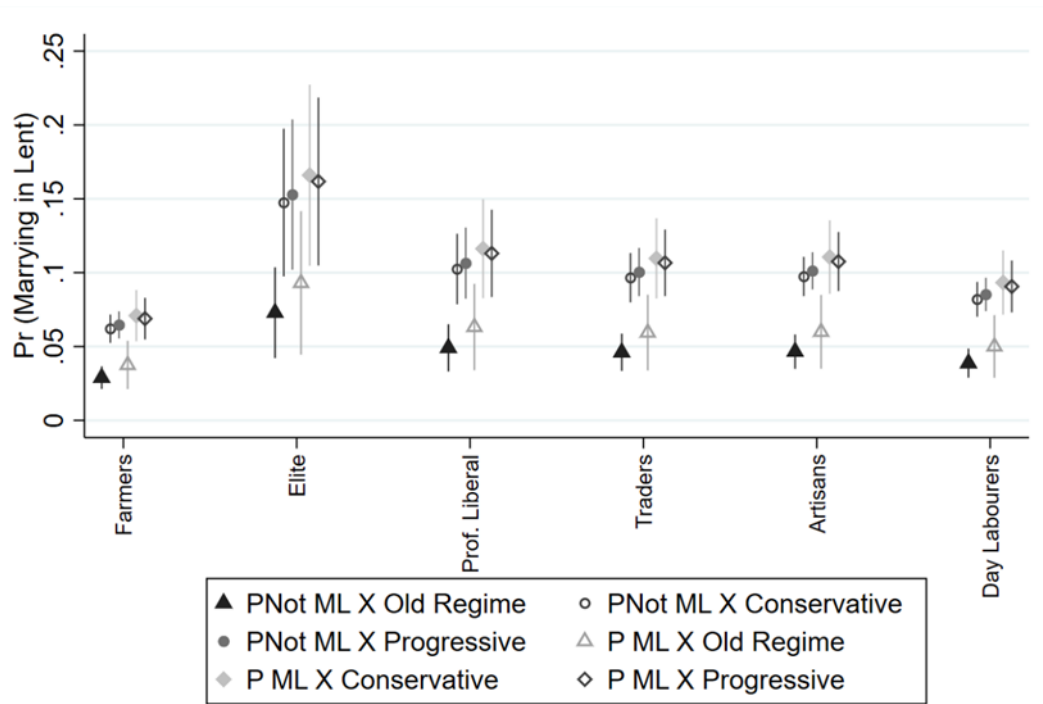
Odds ratios displayed; Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In this sense, we compute the probabilities of getting married in Lent from the interactions of the marriage political period of children and parents (Figure 6.12), where it can be seen that parents married in progressive periods influenced most their children independently of the actual period when children got married (either Liberal or Conservative). Moreover, this influence could also foster the contrary, given parents marrying at conservative political periods had children in progressive periods with similar probability of disrespecting Lent than parents that married during the Old regime.

Finally, in model 10 (see table 8) we tried to disentangle if the fact of having parents that married in Lent could be more important than the political period where children married. Thus we interacted the variable regarding if parents married in Lent with the political periods, controlling also for the social groups children belonged to. The results in odds ratios as well as in probabilities (see figure 6.13) showed that for all social groups, despite the political period when children would marry, the intergenerational transmission of the likelihood of marrying in Lent seemed always more important.

Figure 6.13. Probability of marrying in Lent by the political sign period of children and interaction of the political sign period of parents and if parents married in Lent.



6.5 Conclusions

A progressive change in marriage seasonality can be argued through the increase in the number of marriages that took place during Lent in the area of Barcelona, as has been also pointed out in previous studies for different countries. However, respect for the precept of avoiding Lenten marriages did not totally vanish. In this progressive change of seasonality, there was a clear social class behaviour, which depended on the individual social group of origin, both for grooms and for brides. Farmers maintained the traditional religious calendar, even when in the nineteenth century the primary sector was decreasing in the Barcelona area. In contrast, upper-classes or artisans presented greater chances of getting married during Lent. This different feature can be explained by two reasons: class behaviour regarding marriage and structural changes in the labour market. Traditionally, farmers tended to marry more homogeneously than the rest of social groups, except the nobles, even during the nineteenth century. Conversely, other groups showed higher incidences of intermarriages.

Through this study, we have also shed light on the family determinants that drove this change of nuptial calendar. In this regard, there were substantial effects indicating that secularization's behaviors on marriage were prone to be intergenerationally transmitted. Nonetheless, this was also related to the social group and sibship size but less to the sibling's marriage order and gender. Although the social destination of individuals was important, their social backgrounds were even more decisive.

The political period was also crucial in the movement towards secularization. Accordingly, liberalism helped to break the traditional calendar of marriage. This influence was clear, independently of whether liberalism was more progressive, as at the beginning of the nineteenth century, or more conservative, as in the second half of the century. Moreover, progressive political periods such as the Liberal Triennium (1820-1823) or the Revolutionary six-year period (1868-1874) for instance, which performed different actions as abolishing the Inquisition or establishing religious freedom, sped up an irreversible secularization that was already taking part in the society, which might have been boosted by several episodes of anticlericalism experienced mainly after 1835 and a relaxation of religious norms. By contrast, this trend would be slowed down in the political moment of the restoration of absolutism (1814-1820) in the first half of the nineteenth century. In the long term, the introduction of liberalism represented a clear breakdown in terms of Lent and marriages, starting a point of no return because the number of marriages celebrate during Lent increased over time. At the end, the individual characteristics such second marriages or the intergenerational transmission of the marriage behaviour explained more the propensity to marry in Lent.

It is worth dissociating individual behaviour from that of the ruling political group in the year of marriage or the power maintained by the Church. Otherwise, ignoring these points could lead to the erroneous view that secularization within the Spanish society was non-existent, opposed to what has been observed in the trends analysed within this study. In this sense, Catalonia, and more precisely the area of Barcelona and mainly the city of Barcelona, are likely to have been one of the forerunners in Lent disrespect when compared to other Spanish or even other European Catholic regions. We must bear in mind that the area of Barcelona showed an early industrialization in comparison with other Spanish or southern European regions, partly because Catalonia was considered the 'factory' of Spain and the importance of the Barcelona Port as diffuser and transmission channel for new ideas and behaviours.

All in all, it is clear that the political period and the economic phase, as well as individual behavior in terms of cultural or religious values, were concomitant in the progressive secularization faced by the Barcelona area. All of these forces were communicating vessels of desacralisation within the Spanish society. As well, we believe that this article opens other questions to be answered, as in the case of the exact determinants for the concentration of marriages in December, which will be contemplated in our future research agenda.

7 Conclusions.

Aquesta tesi mostra els avantatges de la construcció de grans bases de dades demogràfiques individuals que serveixen per interpretar amb més precisió els comportaments demogràfics i socials en el passat. De fet, la implantació de grans infraestructures pel buidatge de dades és ja una realitat fruit de la col·laboració entre diferents disciplines científiques. La demografia històrica i la visió per computador, juntament amb la participació del voluntariat social a través del paradigma de *crowdsourcing*, ho han fet possible. D'aquesta manera, a partir dels padrons municipals d'habitants com a pal de paller, des de l'àrea de Demografia Històrica del Centre d'Estudis Demogràfics i el Grup de d'Anàlisi de Documents del Centre de Visió per Computador, teixint tota una xarxa de dades per comprendre millor els nostres avantpassats. A més a més de la informació demogràfica, també s'hi està afegint informació econòmica, procedent de les anomenades cèdules personals. De tal manera, que ja no només disposem d'una dada individual d'una persona, sinó que aquesta està connectada amb tot el seu entorn familiar, social i territorial. Per altra banda, aquestes bases individuals permeten introduir la perspectiva de *life course*, que avui en dia és molt habitual en la recerca en demografia històrica arreu d'Europa. Una perspectiva individual que ens ofereix un major ventall de possibilitats per analitzar els comportaments de la societat enfront les dades agregades. Ens permet arribar a la interpretació del conjunt a partir de la singularitat específica de cada individu.

La construcció d'aquestes bases es recolza en quatre fases. La primera és el procés de digitalització dels fons arxivístics conservats, relacionats amb la població, que esdevé fonamental no tant sols per la seva preservació sinó també és imprescindible per l'aplicació dels mètodes de visió per computador. Amb el pas del temps, el paper i la tinta esdevenen suports fràgils en perill de desfer-se, a més a més la manipulació per part dels investigadors/es, encara posa més en risc de desaparició d'aquesta documentació. L'escaneig i el posterior tractament de la digitalització, mitjançant la catalogació i anàlisi dels seus continguts, són part de la solució. Un altre handicap per a la conservació de tot aquest patrimoni documental són les condicions en que es troben alguns arxius, tant siguin eclesiàstics com civils, en l'àmbit local. És cert, que des de l'administració s'està portant a terme una important tasca per recopilar i conservar en bones condicions molta

documentació. Els arxius comarcals son una peça cabdal en tot aquest procés. El seu esforç en la conservació, catalogació i difusió de la documentació permet als investigadors, no només obtenir dades pels seus estudis, sinó que també permeten descobrir noves fonts d'informació que enriqueixen la recerca. Com a mostra d'exemple a seguir, és la línia d'actuació portada a terme per part de l'Arxiu Comarcal del Baix Llobregat, tant en la digitalització com en la difusió del seu fons, que ens ha facilitat en tot moment la nostra tasca de recerca.

La següent fase, un cop assegurada la conservació de la font documental, és el buidatge de la font. Una tasca, la transcripció, que sempre es molt costosa en temps de dedicació. Però per dur-ho a terme eficientment, es fonamental poder escurçar el temps de construcció de les bases de dades. I és aquí quan entren en funcionament les tècniques de visió per computador per tal d'aplicar mètodes d'extracció massiva per obtenir el contingut de la font. En altres paraules, passar de la informació de text que hi ha a la imatge, a la base de dades. Es veritat que en els darrers temps, els ordenadors son capaços de llegir textos mecanografiats mitjançant sistemes de reconeixement òptic de caràcters, els anomenats OCR, i que també s'està avançant molt en el reconeixement d'escriptura manual, malgrat que en aquest darrer punt encara ens trobem en una fase d'evolució. Doncs l'especificitat de cada persona en realitzar el traç en el moment d'escriure, fa que els softwares i algorismes per transcriure correctament les dades no siguin tan precisos i s'hagin d'adaptar a les diferents formes d'escriptura. Es per aquest motiu que la transcripció manual esdevé necessària sobretot pels processos d'entrenament i aprenentatge dels algorismes de lectura de manuscrits. Per altra banda, es evident que la tasca de transcripció dels fons demogràfics es immensa. Per aquest motiu, el *Crowdsourcing* i la col·laboració ciutadana són la clau de volta per donar resposta aquesta dificultat. La compartició d'un treball molt gran en petites unitats per realitzar per part de diferents persones i l'accés a les noves tecnologies, ens permet poder abordar la construcció de grans bases de dades demogràfiques que fins ara semblaven impossibles. Al mateix temps, tota aquesta informació transcrita servirà per "entrenar" els algorismes d'Intel·ligència Artificial que han de permetre la lectura automàtica de manuscrits històrics en un futur no molt llunyà.

La tercera fase és harmonitzar i estandarditzar les dades transcrites per aconseguir que aquestes grans infraestructures d'informació demogràfica siguin comparables a nivell internacional. Les bases de dades han de poder ser fàcilment explotables, i això passa en

primer lloc per una neteja i depuració de les errades i/o incongruències que es deriven tant de la font original com del procés de transcripció. I en segon lloc, per a la creació de noves variables que facilitin el seu anàlisi. D'aquesta manera calen confeccionar marcadors que permetin relacionar i enllaçar les informacions individuals amb les de llar. Alhora que també es necessari poder enllaçar les dades dels individus entre les diferents fonts d'informació. En darrer lloc s'ha de redactar tot un conjunt de documentació (Metadata) que cal adjuntar, explicant com s'ha construït la base de dades i quins han estat els processos i les presses de decisions en cada moment per la seva confecció.

I finalment, tota aquesta informació ens ha de servir per poder investigar i avançar en el coneixement del nostre passat demogràfic. Necessitem crear nous bancs de dades històrics per fer-nos noves preguntes i verificar hipòtesis. Serà a partir de grans bancs de dades demogràfiques individuals i la utilització de tècniques metodològiques d'anàlisi, que podrem arribar a obtenir millors conclusions per interpretar el nostre passat demogràfic del conjunt de la societat.

Les aportacions dels dos primers assajos mostren les experiències de la creació de grans infraestructures d'extracció massiva d'informació demogràfica. Una tasca que passa per la col·laboració entre demògrafs i enginyers informàtics, així com també amb l'aportació ciutadana. Les experiències en crowdsourcing dutes a terme fins ara en el marc dels projectes de l'Àrea de Demografia Històrica del Centre d'Estudis Demogràfics i el *Document Analysis Group* del Centre de Visió per Computador, marquen el camí a seguir per realitzar un buidatge massiu de fonts demogràfiques. Com més agents s'involucren en la tasca de reconstrucció de la història de la població més ràpid i millors bases de dades s'obtindran.

En l'assaig *The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth – twentieth centuries)* s'explica com s'ha dut a terme el procés d'extracció massiva de dades demogràfiques per a la creació de la base de dades del Baix Llobregat: Baix Llobregat (BALL) Database. Un banc de dades que compta amb 152 registres de població (Padrons municipals d'habitants i Censos de Població), conservats i digitalitzats a l'Arxiu Comarcal del Baix Llobregat, des de 1828 fins a 1950 i que conté més de 220.000 registres individuals. A través d'aquestes fonts d'informació podem seguir els individus al llarg del temps, i observar d'on eren, si entremig s'han casat, han canviat de feina o de residència, han tingut fills, han enviudat, etc... i alhora, enllaçar-los amb la seva estructura familiar i el seu entorn socioeconòmic.

Disposem de la digitalització de les fonts demogràfiques feta per l'Arxiu Comarcal del Baix Llobregat de 9 municipis de la comarca (Begues, Castellví de Rosanes, Collbató, Corbera de Llobregat, El Papiol, Molins de Rei, Sant Feliu de Llobregat, Santa Coloma de Cervelló i Torrelles de Llobregat) amb la participació de 119 voluntaris/es que mitjançant una plataforma de *crowdsourcing* i amb el suport de tècniques de visió per computador han reconstruït gran part de la població del Baix Llobregat. Les creacions de formularis de transcripció *ad hoc*, per cada padró o cens de població, a través d'aquesta plataforma online en la qual el transcriptor/a podia accedir des de casa seva quan volgués, ha fet que es reduïssin molt els temps de construcció de la base. Paral·lelament, es van implementar dos videojocs, per avançar en les diferents formes de validació de les transcripcions automàtiques realitzades pels algoritmes d'Intel·ligència Artificial. Aquest procés de gamificació amplia i diversifica les formes de transcripció semi-automàtica. En aquest sentit, cal destacar el gran salt qualitatiu que s'està produint en la transcripció automàtica de manuscrits històrics en la darrera dècada, malgrat que l'heterogeneïtat en les diferents formes d'escriptura fa que encara estiguem en el procés per aconseguir una transcripció literal de tot un text, però sí que podem indexar o cercar paraules concretes, entre moltes d'altres funcionalitats.

Al mateix temps, s'ha creat una base de dades longitudinal de Sant Feliu de Llobregat, que conté 59.084 observacions individuals i 12.748 llars des de 1828 fins 1940. La vinculació automàtica s'ha creat en dues fases, la primera mitjançant amb dos algoritmes de *string distance* (Bag distance i Levenshtein distance) que permeten enllaçar aquells camps que malgrat no s'hagin transcrit igual són el mateix (per exemple les múltiples possibilitats d'escriure el cognom Ferrer: Farre, Ferre, Farrer,...) i una segona fase utilitzant criteris probabilístics que han permès el linkatge de la mateixa persona en diferents padrons o censos. De tal manera que podem fer el seguiment de 10.405 persones, com a mínim en 2 registres diferents.

Per últim, cal esmentar la importància del que s'anomena ciència ciutadana en tots aquests processos de creació de bases de dades. Apropar la ciència a la població en general és també una forma de fer ciència. És necessari que hi hagi un retorn a la societat de la tasca feta a la universitat. Per això destaquem la creació d'una pàgina web que permet fer tant cerques onomàstiques com visualitzacions interactives (mapes, piràmides de població, ...) amb les dades històriques dels municipis transcrits.

El segon assaig *Knowledge graph based methods for record linkage*, es centra en un aspecte molt més específic de la creació de les bases de dades. Posa de relleu la importància d'avançar en noves metodologies per obtenir una millor precisió alhora de vincular registres individuals amb la finalitat de reconstruir trajectòries vitals. Un cop tenim la base de dades transcrita cal establir mecanismes àgils i eficients que ens permetin reconstruir línies de vida (*life course*) dels diferents individus registrats. És per aquest motiu que proposem un mètode anomenat WERL (Weighted embedding based record linkage) en el qual mitjançant les tècniques de *Knowledge graph* realitzem un pas més endavant en la vinculació automàtica d'individus mitjançant algorismes d'Intel·ligència Artificial.

La demografia històrica utilitza cada vegada més la dada a nivell individual per entendre els comportaments de la població històrica com a conseqüència de la predominança dels anàlisis de *life course*. De tal manera que els processos de *record linkage* són essencials per vincular les diferents observacions d'una persona al llarg del temps i l'espai. Els mètodes utilitzats fins ara no són del tot precisos i generen sovint resultats que contenen molts falsos positius o duplicats. Per tant, la nova metodologia proposada s'estructura en dos passos. El primer pas és que el mètode pugui aprendre a incorporar vectors que codifiquin millor les dades procedents dels registres de població. Així el model té en compte els possibles canvis que tingui una persona al llarg del temps. Per exemple si ens referim a l'estat civil d'una persona que ha estat casada, en el futur no pot ser soltera però sí que pot continuar estant casada o ser viuda. El segon pas, una vegada millorada la vinculació dels registres, es realitza una ponderació de pesos entre els possibles candidats. Això ens ha permès identificar quines variables són més discriminants alhora de realitzar tasques de vinculació com són la coincidència total dels noms i cognoms, l'any de naixement, l'estat civil, l'ocupació o la relació amb el cap de la llar.

Per a la validació del mètode s'ha realitzat una comparativa amb tres bases de dades de referència: la Baix Llobregat Demographic Database (BALL) utilitzant 16 padrons de població de Sant Feliu de Llobregat desde 1828 fins a 1940; la Freely Extensible Biomedical Record Linkage (Ferbl) amb dades mèdiques de pacients i finalment la Cora data set (Cora) una base de dades de cites bibliogràfiques. Els resultats obtinguts han estat molt satisfactoris malgrat que s'haurien d'aplicar en conjunts de dades més grans i heterogènies per tal de testar amb major profunditat la solidesa del mètode. No obstant això, s'ha vist com la proposta del WERL millora les metodologies existents.

El tercer i quart assaig, són una mostra dels múltiples estudis que podem realitzar a partir de les bases de dades creades. Així, en el tercer assaig analitzem els processos de transmissió intergeneracional a l'Antic Règim, *De tals pares, tals fills: transmissió intergeneracional a Sabadell, segles XVI i XVII*. Un assaig que es desenvolupa en el marc del premi de la 9^a edició de la Beca Miquel Carreras, convocada pel Museu d'Història de Sabadell i la Fundació Bosch i Cardellach. El treball mostra la importància que tenia la família en la reproducció social dels individus. La transmissió intergeneracional de la posició socioeconòmica a Sabadell era alta: concretament un 75 % dels fills tenien la mateixa ocupació que els seus pares, una xifra que també s'observa en altres indrets d'Europa en el mateix període. Aquesta transmissió intergeneracional també s'ha pogut comprovar segons de l'ordre de casament dels fills, a partir dels residus estandarditzats, i s'ha vist com aquells fills que es casaven primer tenien una propensió molt més alta d'heretar la mateixa posició social del pare. En aquest sentit, tenint en compte que generalment els primers en casar-se solien ser els hereus, es confirma la tendència que les diferents activitats econòmiques familiars (la terra per als pagesos o els tallers per als artesans) esdevenien una responsabilitat principalment d'aquells fills que es casaven primer. Per als altres germans que es casaven en segon o tercer lloc, la lògica del sistema de l'hereu únic feia que en moltes ocasions haguessin de cercar altres sortides socioeconòmiques, un fet que més que casual podia ser una important decisió estratègica per tal de diversificar l'economia familiar. Per això es donà el cas que el fill d'un pagès es fes artesà i viceversa, aquesta diversificació no estava lliure d'obstacles en els dos grups socials analitzats (pagesos i artesans).

D'una banda, cal tenir en compte que el fet que un fill de pagès fos artesà comportava haver de bregar amb les rígides regulacions gremials i la manca d'una xarxa en la menestralia. De l'altra, en el cas dels fills d'artesans l'accés a la terra fou difícil com a conseqüència d'un mercat de terres molt minso, cosa que va fer que el preu de les terres s'encarís. A més, en el cas específic de Sabadell, en aquest període hem pogut veure que la xarxa menestral no sempre operava de forma positiva, ja que els resultats dels residus estandarditzats ens ensenyen una possible conseqüència d'una relació històricament conflictiva entre paraires i teixidors. En aquest sentit, encara que aquests fossin els grups menestrals més comuns, els paraires tendiren a evitar tenir fills teixidors i viceversa.

En el quart assaig, *What counted most for secularization: Politics, economy or family? Shifting seasonality of marriages in the Barcelona area, 1715-1880*, realitzem un anàlisi

dels factors que van tenir més influència en el procés de secularització de la societat a l'àrea de Barcelona mitjançant els matrimonis registrats a la Barcelona Historical Marriage Database. L'augment del nombre de matrimonis celebrats durant el període de Quaresma observat a partir de finals del segle XVIII es pot interpretar com l'inici del procés de secularització de la societat, com s'ha vist en altres països. Al nord d'Europa, també s'ha utilitzat el període d'Advent per dur a terme aquest anàlisi, però tant la historiografia com els nostres anàlisis exploratoris han demostrat que en els països del sud, les restriccions alhora de contraure matrimoni durant l'Advent no eren tan estrictes, per la qual cosa no l'hem utilitzat en els nostres anàlisis. En aquest treball cal destacar les novetats analítiques aportades alhora d'abordar el tema de la secularització respecte a la literatura existents fins ara. Tradicionalment s'havia analitzat mitjançant l'anàlisi de l'evolució de l'estacionalitat dels matrimonis i el canvi ocupacional. No obstant això, l'assaig fa un pas més endavant i no només estudia l'estacionalitat sinó que observa les influències de la família i el context polític en el calendari nupcial a partir d'una aproximació estadística multinivell.

En aquest procés de canvi, hem pogut demostrar que hi ha un comportament de classe social. Les dades reflecteixen que els pagesos són el grup social que mantenen les tradicions del calendari eclesiàstic que desaconsellava casar-se a Quaresma, malgrat que a finals del segle XIX, aquest sector estava perdent pes a l'àrea de Barcelona. Per l'altra banda, els comerciants i els artesans són el grup que presenta uns canvis més importants alhora de contraure matrimoni durant la Quaresma.

Aquest assaig també posa de manifest quins són els determinants familiars que porten a aquest canvi de comportament en el calendari nupcial. Podem observar com la transmissió intergeneracional va ser un factor substancial, juntament amb el grup social i la grandària de la família. Per altra banda, el període polític en el moment de la celebració del matrimoni també resulta ser un factor determinant en el procés de secularització. Els períodes amb governs més progressistes que acostumen a debilitar el poder de l'Església creen moments més proclius als canvis en el calendari tradicional. Així, durant el Trienni Liberal (1820-1823) o el sexenni revolucionari (1868-1874) ens trobem amb un augment del nombre de matrimonis durant el temps de Quaresma, mentre que per contra, durant els períodes de la restauració absolutista (1814-1820) o la Dècada Omniosa (1823-1833) es produeix un descens.

8 Futures línies de treball.

Cal continuar apostant per la recerca realitzada dins del marc dels diferents projectes duts a terme pel Grup de Demografia Històrica del Centre d'Estudis Demogràfics i el Centre de Visió per Computador per a la construcció d'una gran infraestructura de bases de dades sobre població històrica per comprendre millor el llegat demogràfic. De tal manera que els reptes de futur passen per:

- Continuar explorant les múltiples possibilitats que ofereixen les bases de dades de la BHMD i la BALL Database dins del marc dels estudis com ara la pobresa i la desigualtat social, els moviments migratoris intra i inter municipals, o els anàlisis de localització i distribució espacial de les ocupacions o la pobresa. Al mateix temps, cal seguir avançant amb les línies de recerca iniciades referents als processos intergeneracionals dins la família.
- Establir sinèrgies amb arxius comarcals i municipals, així com també amb arxivers i centres d'estudis locals per a la difusió dels mètodes de transcripció i la creació de bases de dades històriques de la població. D'aquesta manera es cerca ampliar el nombre de municipis que conformen la BALL Database.
- Digitalitzar tot tipus de fonts demogràfiques amb dades individuals, tenint com a prioritat els padrons i censos de població a nivell municipal. Aquestes fonts demogràfiques han de ser el pal de paller dels futurs bancs de dades, dels qual anirem enllaçant noves dades d'altres fonts d'informació, com ara arxius de salaris de grans empreses; informació antropomètrica a partir dels registres de quintos del servei militar, o informació sanitària, social,... Aquestes bases de dades, també es poden completar amb vinculacions amb arxius fotogràfics (col·leccions històriques de fotografies) o sons (memòria oral). Cal teixir una veritable xarxa per tal d'obtenir un Big Data del passat.
- Recuperar i actualitzar les bases de dades demogràfiques individuals ja existents, concretament de les fitxes de família realitzades en diferents àrees de la geografia espanyola, perquè no caiguin en l'oblit i es pugui realitzar una nova interpretació de la seva informació amb els nous mètodes i tècniques d'anàlisis existents.
- Continuar amb la col·laboració amb altres branques del coneixement com els enginyers en visió per computador per a la millora de les tècniques de transcripció automàtica de manuscrits històrics; així com també la sofisticació d'algoritmes de

vinculació que permetin relacionar les dades individuals d'èpoques i registres diferents. Aquests avenços també passen per una millora de la gamificació de la validació i harmonització de les dades que ens permetrà escurçar el temps i els esforços en la construcció de bases de dades. De tal manera que cada vegada siguem més àgils en la construcció de nous bancs de dades.

- Elaboració d'un manual de bones pràctiques per la creació de bases de dades demogràfiques, per tal de contribuir en la unificació de criteris i així obtenir bases de dades comparables internacionalment.

9 Bibliografia.

- Albaigès (2005) *Gran llibre dels cognoms Catalans*, Edicions 62, Barcelona
- Albanese, G.; De Blasio, G.; Sestito, P. (2016) My parents taught me. evidence on the family transmission of values, *Journal of Population Economics* 29 (2): 571-592.
- Alcover, A.M.; Moll, F.B. (1993) *Diccionari català-Valencià-Balear: Inventari Lexicogràfic i Etimològic de la llengua catalana en totes les seves formes literàries i dialectals*. Editorial Moll, Mallorca.
- Alonso, G. (2003) La secularización de las sociedades europeas, *Revista historia social*, XLIII: 137-157.
- Alter, G.; Mandemakers, K. (2014) The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical life course studies*: 1.
- Alter, G.; Newton, G.; Oepen, J. (2020) Re-introducing the Cambridge Group Family Reconstitutions. *Historical life course studies*, 9: 24-48.
- Arrizabalaga, M. P. (2005). Succession strategies in the Pyrenees in the 19th century: The Basque case. *The History of the Family*, 10(3), 271–292.
- Aubert, P. (ed.) (2002) *Religión y sociedad en España, siglos XIX y XX*, Madrid.
- Barbi, E., Bertino, S. and Sonnino, E., (2004) *Inverse projection techniques: old and new approaches*. Springer Science & Business Media.
- Barrera-Gonzalez, A. S. (1992). Eldest and younger siblings in a stem-family system: The case of rural Catalonia. *Continuity and Change*, 7(03), 335–355.
- Bearman, P. S.; Deane, G. (1992). The structure of opportunity: Middle-class mobility in England, 1548-1689. *American Journal of Sociology*, 98(1), 30–66.
- Benaül, JM. (director) (2009). *El Gremi de Fabricants de Sabadell 1559-2009*. Fundació Gremi de Fabricants de Sabadell. Sabadell.
- Bengtsson, T.; Dribe, M; Svensson, P. (2012). *The Scanian Economic Demographic Database: Version 3.0* (Machine-readable database). Lund, Lund University, Centre for Economic Demography.
- Bengtsson, T., M. Dribe, L. Quaranta and P. Svensson (2014). *The Scanian Economic Demographic Database. Version 4.0* (Machine-readable database). Lund: Lund University, Centre for Economic Demography.

- Berkner, L.; Mendels, F. (1976). Inheritance Systems, family structure and demographic patterns in Western Europe, 1700-1900. TILLY, C. (ed.) *Historical Studies in Changing Fertility*, Princeton, Princeton University Press. 209-223.
- Billari, F.; Zagheni, E. (2017) Big Data and Population Processes: A revolution?. In Petrucci, A. & Verde, R. (eds.) SIS 2017. *Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*, Firenze University Press: 167-178.
- Bloothoof, G. (1998) Assessment of systems for nominal retrieval and historical record linkage. *Computers and the Humanities*. 32 (1): 39-56.
- Boberg-Fazlic, N.; Sharp, P.R.; Weisdorf, J. (2011). “Survival of the Richest? Social Status, Fertility, and Social Mobility in England 1541-1824”. *European Review of Economic History*, 15(3), 365-392.
- Borderías, C. (2012). La reconstrucción de la actividad femenina en Cataluña circa 1920. *Historia Contemporánea*. 44:17-47.
- Borderías, C.; Ferrer-Alòs, Ll. (2017) The stem family and industrialization in Catalonia (1900–1936). *The History of the Family*. 22(1): 34-56.
- Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. (2013) Translating embeddings for modeling multi-relational data, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 26: 2787–2795.
- Bosch, M. (2010). Movilidad y herencia de los masovers catalanes. La familia Culubret, siglos XVII-XIX. *Historia Agraria*, 52: 45-74.
- Botticini, M. (1999). “A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town”. *Journal of Economic History*, 59(1): 104-121.
- Boudjaaba, F.; Dousset-Seiden, C., Mouysset, S. (2016) *Frères et sœurs du Moyen Age à nos jours*. Bern.
- Bourdieu, P. (1976). Marriage strategies as strategies of social reproduction. R. Forster; O. Ranum (Eds.), *Family and society. Selections from the Annales, Economies, Societes, Civilisations* (pp. 117–144). Baltimore, MD: The Johns Hopkins University Press.
- Bourdieu, P. Et al. (2009). Vive la difference? Intergenerational social mobility in France and the United States during the nineteenth and twentieth centuries. *Journal of Interdisciplinary History* 39: 523-557.
- Bourdieu, J.; Kesztenbaum, L.; Postel-Vinay, G. (2014). The TRA Project, a Historical Matrix. *Population*, 69(2): 191-220.

- Brands, K. (2014). Big Data and Business Intelligence for Management Accountants. *Strategic Finance*, 96(6), pp. 64–5.
- Brea-Martínez, G.; Pujadas-Mora, J.M. (2018) Transformación y desigualdad económica en la industrialización en el área de Barcelona, 1715-1860. *Revista de Historia Económica-Journal of Iberian and Latin American Economic History*. 36(2): 241-273.
- Breschi, M., Fornasin, A.; Manfredini, M. (2020). The Richness of Italian Historical Demography. *Historical Life Course Studies*
- Breschi, M.; Ruii, G. (2016) Superstitions, religiosity and secularization: An analysis of the periodic oscillations of weddings in Italy, *Genus*, 72, (1): 1-7.
- Cabre, A. (1999) *El sistema català de reproducció. Cent anys de singularitat demogràfica*. Barcelona: Editorial Proa.
- Calvi, G.; Blutrach-Jelín, C. (2010) Sibling relations in family history: conflicts, cooperation and gender roles in the sixteenth to nineteenth centuries. An introduction, *European Review of History—Revue Européenne d'histoire*, 17, (5): 695–704.
- Canyameres, E. (1994). La Immigració francesa al Vallès Occidental (S. XVI-XVII) quatre exemples locals: Sabadell, Sant Cugat del Vallès, Sant Julià d'Alta i Sentmenat. *Arraona: revista d'història*, 15: 9-25.
- Carbonell i Porro, J. A. (1995) Desenvolupament de la indústria tèxtil a Sant Feliu de Llobregat (1861 - 1923). In *El pas de la societat agrària a industrial al Baix Llobregat*, edited by Angel Calvo i Calvo. Barcelona, Centre d'Estudis Comarcals del Baix Llobregat, Publicacions de l'Abadia de Montserrat, 405-425.
- Carreras Candi, F. (1913). Les obres de la Catedral de Barcelona 1298-1445, *Butlletí de la Reial Acadèmia de Bones Lletres de Barcelona*. 7(49):22-30.
- Carreras, A. (1990) *Industrialización española: estudios de historia cuantitativa*. Madrid, Espasa Calpe.
- Carreras, A.; Prados de la Escosura, L.; Rosés, J.R. (2005). Renta y riqueza. Cap. 17 en Carreras, Albert; Tafunell, Xaviers (coord.) *Estadística Históricas de España, Siglos XIX-XX*. Fundación BBVA, 1297-1376.
- Cavalli-Sforza, L.; Feldman, M.; Chen, K.H.; Dornbusch, S.M. (1982) Theory and observation in cultural transmission, *Science*, 218, 4567: 19-27.
- Charbonneau, H. (1975). *Vie et mort de nos ancêtres: Étude démographique*. Université de Montréal.

- Chen, J.; Riba, P.; Fornés, A.; Mas, J.; Lladós, J.; Pujadas-Mora, J.M. (2018) Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*.
- Cherlin, A.J. (2001) *Public and private families: A reader*. New York, McGraw-Hill
- Christen, O. (2008) Febrl - an open source data cleaning, deduplication and record linkage system with a graphical user interface, in: *KDD*, 2008: 1065–1068.
- Christen, P. (2012) *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications*. Berlin, Springer Science & Business Media.
- Ciuffreda, A. (1997). Lo zio Prete. Individui, famiglie e parentela nella trasmissione dell'ufficio sacerdotale in una comunità pugliese in età moderna. *Diseguaglianze: stratificazioni e mobilità sociale nella popolazione italiana*, 553-575.
- Clark, J. (2012) Secularization and modernization: The failure of a 'grand narrative', *The Historical Journal*, 55 (1): 161-194.
- Coale, A.; Cotts Watkins, S. (1986) *The decline of fertility in Europe: the revised proceedings of a Conference on the Princeton European Fertility Project*. Princeton, Princeton University Press.
- Coleman, J.S. (1986) Social Theory, Social Research, and a Theory of Action. *American Journal of Sociology*. 91(6): 1309-335.
- Comas D'argemir, D. (1988). Household, family, and social stratification: Inheritance and labor strategies in a Catalan village (Nineteenth and Twentieth Centuries). *Journal of Family History*, 13(1), 143–163.
- Congost, R. (2003). Property rights and historical analysis: What rights? What history?. *Past & Present*, (181): 73-106.
- Conley, D.; Glauber, R. (2008) All in the family? Family composition, resources, and sibling similarity in socioeconomic status, *Research in Social Stratification and Mobility*, 26, (4): 297-306.
- Coontz, S. (2005). *Marriage, a history: From obedience to intimacy or how love conquered marriage*. New York: Viking.
- Coppa, a.; Di Donato, L.; Vecchi, F.; Danubio, M.E. (2001) Seasonality of marriages and ecological contexts in rural communities of central-southern Italy (Abruzzo), 1500-1871, *Collegium antropologicum*, 25, (2): 403-412.
- Coromines, J. (1997) *Onomasticon Cataloniae*, Editorial Curial, Barcelona

- Dambruynne, J. (1998). Guilds, Social Mobility and Status in Sixteenth-Century Ghent. *International Review of Social History*, 43: 31-78.
- Danubio; M.E.; Amicone, E. (2011) Biodemographic study of a central apennine area (Italy) in the 19th and 20th centuries: Marriages seasonality and reproductive isolation. *Journal of biosocial science*, 33: 427-449.
- De la Cueva, J.; Montero, F. (2007) *La secularización conflictiva: España: 1898-1931*, Madrid.
- Díaz Rodríguez, R. (2018). *Connection of a historical social network with Wikimedia community*. Treball final de Grau. Universitat Autònoma de Barcelona.
- Didarul, Md (2020) Understanding secularisation as indicating the process of the separation of the political and religious.” *OSF Preprints*, 1 Apr.
- Desjardins, B. (1998). Le registre de la population du Québec ancien. *Annales de démographie historique*, 2: 215–226.
- Dillon, L.; Amorevieta-Gentil, M.; Caron, M.; Lewis, C.; Guay-Giroux, A.; Desjardins, B., Gagnon, A. (2018) The Programme de recherche en démographie historique: past, present and future developments in family reconstitution. *The History of the Family*. 23(1): 20-53.
- Dong, H., Campbell, C., Kurosu, S., Yang, W., Lee, J.Z. (2015). New sources for comparative social science: Historical population panel data from East Asia. *Demography*, 52 (3): 1061-1088
- Draisbach, U.; Naumann, F. (2010) *DuDe: the duplicate detection toolkit*, ACM - VLDB.
- Dribe, M.; Lundh, C. (2010). Marriage choices and social reproduction: The interrelationship between partner selection and intergenerational socioeconomic mobility in 19th-century Sweden. *Demographic Research*, 22:347–382.
- Dribe, M.; Helgertz, J.; Van de Putte, B. (2015) Did social mobility increase during the industrialization process? A micro-level study of a transforming community in southern Sweden 1828–1968, *Research in Social Stratification and Mobility*, 41: 25-39.
- Dribe, M.; Van de Putte, B. (2012) Marriages seasonality and the industrious revolution. Southern sweden, 1690-1895, *The Economic History Review*, 65, (3): 1123-1146.
- Dupâquier, M. (1977) Le mouvement saisonnier des mariages en France (1856-1968)’ *Annales de démographie historique*: 131-149.
- Ealham, C. (2004) *Class, culture and conflict in Barcelona, 1898-1937*, London.

- Edvinsson, S. (2000). The Demographic Data Base at Umeå University—a resource for historical studies. In *Handbook of international historical microdata for population research*, edited by Hall, Patricia Kelly, Robert McCaa, and Gunnar Thorvaldsen. Minneapolis, Minnesota Population Center, 231-248.
- Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. (2007) Duplicate record detection: a survey, *IEEE Trans. Knowl. Data Eng.* 19: 1–16.
- Engelen, T. (2017) What the Seasons Tell Us. The Monthly Movement of Marriages, Economic Modernization, and Secularization in the Netherlands, 1810-1940, *Historical Life Course Studies*, 4: 165-180.
- Esping-Andersen, G. (1999). *Social foundations of postindustrial economies*. OUP Oxford.
- Ester, P.; Halman, L.; de Moor, R. (eds.) (1993) *The individualizing society: Value change in Europe and North America*. Tilburg.
- Fabra, P. (1937 [1917]) *Diccionari ortogràfic*, IEC, Barcelona.
- Fabra, P. (1983 [1932]) *Diccionari General de la Llengua Catalana*, IEC, Barcelona.
- Fabra, P. (1995 [1918]) *Gramàtica Catalana*, IEC, Barcelona.
- Fabra, P. (2013 [1913]) *Normes Ortogràfiques*, IEC, Barcelona
- Faragó, T. (1994) Seasonality of marriages in Hungary from the Eighteenth to the Twentieth Century, *Journal of Family History*, 19, (4): 333-350
- Fauve-Chamoux, A. (2005). A comparative study of family transmission systems in the central Pyrenees and northeastern Japan. *The History of the Family*, 10(3), 231–248.
- Fauve-Chamoux, A. (2009) *The stem family in Eurasian perspective: Revisiting house societies, 17th-20th centuries*. Vol. 11. Peter Lang.
- Fellegi, I.P.; Sunter, A.B. (1969) A theory for record linkage, *J. Am. Stat. Assoc.* 64: 1183–1210.
- Ferrer Alós, Ll. (1992). Edat de casament i celibat definitiu a la Catalunya Central (1803-1807). *Manuscripts: revista d'història moderna*, (10): 259-286.
- Ferrer Alós, Ll. (1998). Germans al celibat, germanes al matrimoni. El paper dels cabalers en la reproducció social dels grups socials benestants a Catalunya (segles XVIII i XIX). X. Roigé; A. García; M. Mascarell (eds.) *Antropologia del parentiu*, Barcelona, Icaria, Institut Català d'Antropologia, 211 – 249.
- Ferrer Alós, Ll. (2003). Segundones y actividad económica en Cataluña (siglos XVIII-XIX): Reflexiones a partir de la familia Berenguer de Artes. *Revista de Demografia Histórica*, 21(2), 93–126.

- Ferrer Alós, Ll. (2004). Kinship as a mechanism in the social structuring of rural Catalonia (eighteenth and nineteenth centuries). *Journal of Family History*, 29(2), 135–152.
- Ferrer Alós, Ll. (2005). When there was no male heir: The transfer of wealth through women in Catalonia (the pubilla). *Continuity and Change*, 20(1), 27–52.
- Ferrer Alós, Ll. (2007). *Hereus, pubilles i cabalers: el sistema d'hereu a Catalunya*. Catarroja: Afers.
- Ferrer Alós, Ll. (2009). Achieving Well-Being in Spain through the Single Heir System (18th – 19th Centuries. M. Duraes et al. *The Transmission of Well-Being. Gendered Marriages Strategies*. Nova York, Peter Lang.
- Ferrer Alós, Ll. (2010). Matrimonial regimes in Catalonia and Spain. The logic of family systems. Many paths to happiness?. M.P. Arrizabalaga et al.(eds) *Many Paths to Happiness? Studies in population and family history. A festschrift for Antoinette Fauve-Chamoux*. Amsterdam, Aksant.
- Ferrer, Ll. (2010b). Les clàusules dels capítols matrimonials. Els capítols matrimonials, una font per a la història social. R. Ros Massana (ed.) *Els capítols matrimonials: Una font per a la història social*. Biblioteca d'Història Rural. Girona: 55-69.
- Ferrer Alós, Ll. (2018). Estructura agraria, familia troncal, mercado de trabajo y redes sociales en el mundo rural, Cataluña siglos XVIII-XIX. *Mundo Agrario*, 19(40): e079-e079.
- Fleury, M.; Henry, L. (1956) *Des registres paroissiaux à l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien*. Paris.
- Fox, S.; Do, T. (2013). Getting real about Big Data. Applying critical realism to analyse Big Data hype. *International Journal of Managing Projects in Business*, 6(4): 739–60.
- Franck, R.; Galor, O. (2015). Industrialisation and Fertility Decline. *Working Paper, No. 2015-6*. Brown University, Department of Economics.
- Freedman, R. (1979) Theories of fertility decline: A reappraisal. *Social forces*. 58(1): 1-17.
- Fu, Z.; Christen, P.; Zhou, J. (2014) A graph matching method for historical census household linkage, *Advances in Knowledge Discovery and Data Mining*: xx–xx
- Ganzeboom, H.; Treiman, D.; Ultee, W. C (1991). Comparative intergenerational stratification research: Three generations and beyond. *Annual Review of Sociology*, 17(1): 277-302.

- García Espuche, A. (1998). *Un siglo decisivo: Barcelona y Cataluña, 1550-1640*, Vol. II. Madrid: Alianza editorial.
- García Pérez, M.S. (2007) El padrón municipal de habitantes: origen, evolución y significado. *Hispania Nova: Revista de historia contemporánea*. 7:1-8.
- García Pérez, M.S. (2007) Tratamiento y resolución de las descripciones definidas y su aplicación en sistemas de extracción de información, *Hispania Nova*.
- García Ruipérez, M. (2012). El empadronamiento municipal en España. Evolución legislativa y tipología documental. *Documenta & instrumenta*, 10: 45-86.
- Garrabou, R.; Pujol Andreu, J.; Colomé i Ferrer, J. (1991) Salariis, ús i explotació de la força de treball agrícola (Catalunya 1818-1936). *Recerques: història, economia, cultura*. 24:23-51.
- Garrabou, R.; Pujol, J. (1987) El canvi agrari a la Catalunya del segle XIX, *Recerques* 19: 35-83.
- Gifre, P. (2010). El procés final d'implantació dels capítols matrimonials (finals de segle XVI–començament de segle XVII. R. Ros Massana, (ed.) *Els capítols matrimonials, una font per a la història social*, Girona: Biblioteca d'Història Rural, 55-69.
- Goiser, K.; Christen, P. (2006) Towards automated record linkage. In *Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61*:23-31. Australian Computer Society, Inc.
- González-Bailón, S.; Murphy, T. (2013) The effects of social interactions on fertility decline in nineteenth-century France: an agent-based simulation experiment, *Population Studies*, 67, (2): 135-155.
- González-Martín, A. (2008) Ecological and cultural pressure on marriage seasonality in the principality of Andorra, *Journal of Biosocial Science*, 40: 1-18.
- Goody, J. (1970). Marriage prestations, inheritance and descent in pre-industrial societies. *Journal of Comparative Family Studies*, 1(1): 37-54.
- Goody, J. (1971). Class and marriage in Africa and Eurasia. *American Journal of Sociology*, 76(4): 585-603.
- Goody, J. (1983). *The development of the family and marriage in Europe*. Cambridge. Cambridge University Press.
- Gorski, P.; Altmordu, A. (2008) After secularization?. *Annual Review of Sociology*, 34: 55-85.
- Grusky, D.; Hauser, R. (1984). Comparative social mobility revisited: Models of convergence and divergence in 16 countries. *American Sociological Review*, 49, 19-38.

- Guan, S.; Jin, X.; Jia, Y.; Wang, Y.; Shen, H.; Cheng, X. (2017) Self-learning and embedding based entity alignment, in: *2017 IEEE International Conference on Big Knowledge (ICBK)*: 33–40.
- Guimarães, P.E. (2013). *Explorações sobre a mobilidade social em Portugal (1860-1960)*. Anexo I. Caracterização da AHP-MS V 2.0. Progress report nº 2 of the project. Evora.
- Hall, P.; McCaa, R.; Thorvaldsen, G. (ed.). (2000) *Handbook of international historical microdata for population research*. Minneapolis, Minnesota Population Center.
- Haryanto, S. (2019) The four debates on secularism and secularization in the sociology of religion, *Journal of Social Sciences*, 5, (1) 392-401.
- Henripin, J. (1954). *La population canadienne au début du XVIIIe siècle: Nuptialité, fécondité, mortalité infantile*. Paris: Presses universitaires de France.
- Henry, L. (1969a). Schémas de nuptialité: déséquilibre des sexes et âge du mariage. *Population*, 24(6): 1067-1122.
- Henry, L. (1969b). Schémas de nuptialité: déséquilibre des sexes et célibat. *Population*, 24(2): 457-486.
- Henry, L. (1972). Fécondité des mariages dans le quart sud-ouest de la France de 1720 à 1829. *Annales*, 27(3): 612–640.
- Henry, L. (1980) *Tecniques d'analyse en démographie historique*, Paris.
- Henry, L.; Blayo, Y. (1975). La population de la France de 1740 a 1860. *Population (French Edition)*, 30: 71–122.
- Hernández Andreu, J. (2009) Los orígenes de la fiscalidad española contemporánea”. *Revista empresa y humanismo*, 12 (1): 63-88.
- Herzog, T. (2007) Nombres y apellidos: ¿cómo se llamaban las personas en Castilla e Hispanoamérica durante la época moderna? *Jahrbuch für Geschichte Lateinamerikas*. 44(1):1-35.
- Herzog, T., Scheuren, J.; Winkler, W. (2007) *Data quality and record linkage techniques*. New York. Springer Science & Business Media.
- Houdaille, J. (1976). La fécondité des mariages de 1670 à 1829 dans le quart nord-est de la France. *Annales de Démographie historique*: 341–391.
- Houdaille, J. (1978) Un indicateur de pratique religieuse: la célébration saisonnière des mariages avant, pendant et après la révolution française (1740-1829), *Population*, 33, (2): 367-380.

- Hox, J. (2002) Sample sizes for multilevel modeling, in Joerg Blasius, Joop Hox, Edith deLeeuw and Peter Schmidt eds., *Proceedings of the fifth international conference on logic and methodology*, Cologne: 1–19.
- Jomphe, M. (2012) *Structure of the Balsac database and content of the main tables*. Document n°:I-C-245, Université du Québec.
- Jaro, M.A. (1995) Probabilistic linkage of large public health data files, *Stat. Med.* 14, 4 91–4 98.
- Johnson, C.; Sabeau, D. (eds.) (2011) *Sibling relations and the transformations of European kinship, 1300-1900*, New York.
- Jordà, J.P. (2016) *Aproximación a las migraciones históricas a través del estudio de la información nominal*. Doctoral Thesis. Barcelona. Universitat Autònoma de Barcelona.
- Jordà, J.P.; Valls, M.; Pujadas-Mora, J.M. (2013). Apellidos y migraciones. Estudio a través de los fogatges catalanes de 1497 y 1553. *Revista de Demografía Histórica*. 31(1):105-130.
- Kalmijn, M. (1998). “Intermarriage and homogamy: causes, patterns, trends. *Annual Review of Sociology*, (24):395-421.
- Kistermann, F. (2005) Hollerith punched card system development (1905-1913), *Annals of the History of Computing*, 27: 56–66.
- Knigge, A.; Maas, I.; Van Leeuwen, M.; Mandemakers, K. (2014) Status attainment of siblings during modernization, *American Sociological Review*, 79, (3): 549-574.
- Knodel, J. (1988) *Demographic Behavior in the Past: A Study of Fourteen German Village Populations in the Eighteenth and Nineteenth Centuries*, Cambridge.
- Kolk, M. (2014) Multigenerational transmission of family size in contemporary Sweden”, *Population Studies*, 68, (1): 111-129.
- Kurosu, S. (1996). Leaving home in a stem family system: Departures of heirs and non-heirs in preindustrial Japan. *The History of the Family*, 1(3): 329–352.
- Kussmaul, A (1985) Time and space, hoofs and grain: the seasonality of marriage in England, *The Journal of Interdisciplinary History*, 15, (4): 755-779.
- Kuznets, S.; Murphy, J. (1966) *Modern economic growth: Rate, structure, and spread*. Vol. 2. New Haven: Yale University Press.
- Lambert, P.S.; Zijdeman, R.; Van Leeuwen, M.; Maas, I.; Prandy, K. (2013) The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. 46(2):77-89.

- Laslett, P. (1983). A one-class society. Social divisions and power relations amongst nobility, gentry, townsmen and peasants. P. Laslett (Ed.), *The World We Have Lost: Further explored*. London: Routledge. 22-52.
- Latreille, A.; Cunneen, J.E. (1963) The catholic church and the secular state: the church and the secularization of modern societies, *CrossCurrents*, 13, 2: 217-248.
- Lee, J.; Campbell, C.; Chen. S. (2010) China multi-generational panel dataset, Liaoning (CMGPD-LN). In *User guide: 1749-1909*. Inter-university Consortium for Political and Social Research.
- Lee, J.Z.; Campbell, C.; Chen, S. (2010) China multi-generational panel dataset, Liaoning (CMGPD-LN). In *User guide: 1749-1909*. Inter-university Consortium for Political and Social Research.
- Lesthaeghe, R. (1989) *Marriage seasonality, moral control and reproduction in Belgium (1600-1900)*, Brussels.
- Lesthaeghe, R.; López-Gay, A. (2013) Spatial continuities and discontinuities in two successive demographic transitions: Spain and Belgium, 1880-2010, *Demographic Research*, 28: 77-136.
- Lesthaeghe, R.; Wilson, C. (1997) Modes of production, secularization and the pace of the fertility decline in Western Europe, 1870-1930, in John. Coale and Susan Cotts Watkins eds., *The decline of fertility in Europe*, Princeton.
- Li, Z.; Zhao, S.; Ding, X.; Liu L. (2017) Knowledge base for event evolutionary principles and patterns *EEG. 2017*: 40–52.
- Lin, Y.; Liu, Z.; Luan, H., Sun, M.; Rao, S.; Liu, S. (2015) Modeling relation paths for representation learning of knowledge bases, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*: 705–714.
- Lin, Y.; Liu, Z.; Sun, M. (2016) Knowledge representation learning with entities, attributes and relations, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*: 2866–2872.
- Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. (2015) Learning entity and relation embeddings for knowledge graph completion, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*: 2181–2187.
- Lipset, S.; Zetterberg, H. (1956). A theory of social mobility. Lipset, S.; Bendix, R. (ed.) *Social Mobility in Industrial Society*. Berkeley: University of California Press.

- Lladós, J.; Rusinol, M.; Fornés, A.; Fernández, D.; Dutta, A. (2012) On the influence of word representations for handwritten word spotting in historical documents. In *International journal of pattern recognition and artificial intelligence* 26 (5),1263002.
- Luccheti, E.; Manfredini, M.; Boetsch, G.; Bley, D.; Aluja, P.; Pena, J.; Revello, D.; Melleri, R.; Sevin, A. (1996) Changes in marriages seasonality among some european rural populations, *International Journal of Anthropology*, 11: 73-81.
- Lundh, C. (2003) Swedish marriages. Customs legislation and demography in the eighteenth and nineteenth centuries, *Lund Papers in Economic History*, 88: 1-63.
- Mandemakers, K. (2002) Building life course datasets from population registers by the Historical Sample of the Netherlands (HSN). *History and Computing*. 14(1-2):87-107.
- Mantecón Movellán, T. (2013) Cencerradas, cultura moral campesina y disciplinamiento social en la España del Antiguo Régimen, *Mundo agrario* 14, (27): 1-29.
- Marfany, J. (2012). *Land, Proto-Industry and Population in Catalonia, c. 1680-1829: An Alternative Transition to Capitalism?* Surrey: Ashgate Publishing, Ltd.
- Martínez Carrión, J.M. (1984) Estacionalidad y cambio demográfico. La transformación del "ciclo vital anual" en tierras albacetenses durante los siglos XIX y XX *Al-Basit: Revista de estudios albacetenses*, 13: 87-136.
- Martínez Cuadrado, M. (1976). El sistema Social. *La burguesía conservadora (1874-1931)*. Alianza Universidad, Madrid: 216-283.
- Martínez-Galarraga, J.; Prat, M. (2016) Wages, prices, and technology in early C atalan industrialization. *The Economic History Review*. 69(2):548-574.
- Mas, J.; Fornés, A.; Lladós, J. (2016) An interactive transcription system of census records using word-spotting based information transfer. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*: 54-59.
- Matsuo, H.; Matthys, K. (2018) Trends in secularization and marriage seasonality in the province of West Flanders, Belgium, in the long 19th century, in Paul Puschmann and Tim Riswick eds., *Building Bridges. Scholars, History and Historical Demography*, Nijmegen: 574-589.
- Matsuo, H.; Matthys, K. (2019) Rural and urban differences in marriage seasonality in first marriages: Northern Belgium from early 19th early 20th century, *Chaire Quetelet*, Date: 2019/11/07 - 2019/11/08, Location: Louvaine-le-Neuve.
- Matthijs, K.; Hin, S.; Kok, J.; Matsuo, H. (2016) *The future of historical demography. Upside down and inside out*. Leuven/Den Haag: Acco.

- Matthijs, K.; Moreels, S. (2010). The Antwerp COR*-database: a unique Flemish source for historical-demographic research. *The History of the Family*, 15(1): 109-115.
- Melis Maymar, F. (2019). Distribución personal y provincial de la renta en 1926 según el Impuesto de Cédulas Personales. *Papeles de Trabajo 3/2019*. Instituto de Estudios Fiscales, Madrid.
- Moll, B. (1959) *Els llinatges catalans. Catalunya, País Valencià i Illes Balears*. Editorial Moll, Mallorca
- Monge, A.; Elkan, C. (1997) *An efficient domain-independent algorithm for detecting approximately duplicate database records*.
- Morschheuser, J.; Koivisto, J. (2016) Gamification in crowd-sourcing: a review, *HICSS*, :4375-4384.
- Muñoz Pradas, F. (1997). Fluctuaciones de precios y dinámica demográfica en Cataluña (1600–1850). *Revista de Historia Económica -Journal of Iberian and Latin American Economic History*, 15(3): 507-543.
- Nadal, J. (1975) *El fracaso de la revolución industrial en España: 1830-1914*. Barcelona: Ariel.
- Nadal, J. (1992) *Moler, tejer y fundir: estudios de historia industrial*. Barcelona: Ariel.
- Ní Bhrolcháin, M. (2000). La flexibilité du marché matrimonial. *Population* 55(6): 899-939.
- Nickel, M.; Rosasco, L.; Poggio, T. (2016) Holographic embeddings of knowledge graphs, in: *30th Conference on Artificial Intelligence, 2016*: 1955–1961.
- Nickel, M.; Tresp, V.; Kriegel, H.P. (2011) A three-way model for collective learning on multi-relational data, in: *28th International Conference on International Conference on Machine Learning*: 809–816.
- Nikolov, A.; d'Aquin, M.; Motta, E. (2012) Unsupervised learning of link discovery configuration, in: *9th International Conference on The Semantic Web: Research and Applications*:119–133.
- Peytavi Deixona, J. (2010) *Antroponímia, poblament i immigració a la Catalunya moderna: l'exemple dels comtats de Rosselló i Cerdanya (segles XVI-XVIII)*. Vol. 20. Barcelona: Institut d'Estudis Catalans.
- Porter, T. (1995) *Trust in Numbers. the Pursuit of Objectivity in Science and Public Life*, Princeton University Press.
- Portet Pujol, J. (2003) L'exclaustració de religiosos de l'any 1835. *AUSA*, XXI: 29-49.

- Portillo Navarro, M.J. (1997). Evolución del sistema impositivo español desde 1845 hasta la ‘reforma tributaria silenciosa’ de Flores de Lemus. *Anales de Derecho*. Universidad de Murcia, 15: 129-149.
- Prada Moliner, A. (2016) En torno a la Revolución Liberal y la Iglesia española del siglo XIX, *Ler Historia* 69: 31-50.
- Prandy, K. (2000) Class, the stratification order and party identification. *British Journal of Political Science*. 30(2):237-258.
- Pujadas-Mora, J.M., Romero Marín, J.; Villar, C. (2014) Propuestas metodológicas para la aplicación de HISCO en el caso de Cataluña, siglos XV-XX. *Revista de Demografía Histórica*. 32(1):181-220.
- Pujadas-Mora, J.M.; Brea-Martínez, G.; Jordà, J.P.; Cabré, A. (2018) The apple never falls far from the tree: siblings and intergenerational transmission among farmers and artisans in the Barcelona area in the sixteenth and seventeenth centuries”, *The History of the Family*, 23, (4): 533-567.
- Pujadas-Mora, J.M.; Fornes, A.; Lladós, J.; Brea-Martínez, G.; Valls-Figols, M. (2019) The Baix Llobregat (BALL) Demographic Database, between Historical Demography and Computer Vision (nineteenth-twentieth centuries) In *Nominative Data in Demographic Research in the East and the West: monograph* :29–61.
- Pujadas-Mora, J.M.; Fornés, A.; Lladós, J.; Cabré, A. (2016) Bridging the gap between historical demography and computing: tools for computer-assisted transcription and analysis of demographic sources. In *Future of historical demography: upside down and inside out*, edited by Matthijs, Koenraad, Saskia Hin, Jan Kok, and Hideko Matsuo. Leuven ACCO: 222-231.
- Raman, V.; Hellerstein, J. (2001) Potter’s Wheel: an Interactive Data Cleaning System, *VLDB J.* 2: 381–390.
- Raphael, L. (2008) *Ley y orden. dominación mediante la administración en el siglo XIX, Siglo XXI*, Madrid.
- Ravikumar, P.; Cohen, W.W. (2004) A hierarchical graphical model for record linkage, in: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, AUA Press: 454–461.
- Reher, D. S. (1986). Desarrollo urbano y evolución de la población: España 1787–1930. *Revista de Historia Económica-Journal of Iberian and Latin American Economic History*, 4(1): 39-66.

- Reher, D.S. (2004) Family ties in western Europe. In *Strong family and low fertility: A paradox?*, edited by Gianpiero Dalla Zuanna and Giuseppe A. Micheli. Dordrecht: Springer, 45-76.
- Reher D.S.; Sanz Gimeno, A. (2007) Rethinking historical reproductive change: Insights from longitudinal data for a Spanish town, *Population and Development Review*, 33 (4): 703-727
- Reher, D.S.; Valero Lobo, A. (1995) *Fuentes de información demográfica en España*. Vol. 13. Madrid: Centro de Investigaciones Sociológicas.
- Ristad, E.S.; Yianilos, P.N. (1998) Learning string-edit distance, *IEEE Trans. Pattern Anal. Mach. Intell.* 20: 522–532.
- Ros, C. (2006). Les estratègies familiars i la mobilitat social dels menestrals a Catalunya en els segles XVII-XIX: El cas de Mataró. *Butlletí de la Societat Catalana d'Estudis Històrics*, 16, 171–190.
- Rubio, A.; Lizondo, M.R. (1997) *Antroponímia valenciana del segle XIV: nòmimes de la ciutat de València (1368-69 i 1373): estudi i índexs*. Universitat de València, València.
- Ruggles, S. (2014). Big microdata for population research. *Demography*. 51(1):287-297.
- Ruggles, S.; Roberts, E.; Sarkar, S.; Sobek, M. (2011) The North Atlantic population project: Progress and prospects. *Historical Methods*. 44(1):1-6.
- Ruggles, S.; Fitch, C.; Roberts, E. (2017) Historical Census Record Linkage, *Technical Report*, 2017-3, University of Minnesota.
- Ruggles, S.; Roberts, E.; Sarkar, S.; Sobek, M. (2011) The North Atlantic population project: Progress and prospects. *Historical methods* 44 (1): 1-6.
- Ruiu, G.; Breschi, M. (2015) For the times they are a changin' - the respect for religious precepts through the analysis of the seasonality of marriages. Italy, 1862-2012, *Demographic Research*, 33: 179-210.
- Ruiu, G.; Gonano, G. (2015) Seasonality of marriages in Italian regions: an analysis from the formation of the Italian kingdom to the present', *Italian Review of Economics, Demography and Statistics*, 69, 1: 135-142.
- Sanna, E.; Danubio, M.E. (2008) Seasonality of marriages in sardinian pastoral and agricultural communities in the nineteenth century, *Journal of Biosocial Science*, 40, (4): 577-586.

- Sardon, J.P. (1979) Mariage et révolution dans une petite ville de vigneron: Argenteuil (1780-1819)”, *Population*, 34, (6): 1162-1167.
- Schönplflug, U. (2001) Intergenerational transmission of values: The role of transmission belts, *Journal of Cross-Cultural Psychology*, 32: 174-185.
- Schürer, K. (2007) Focus: Creating a Nationally Representative Individual and Household Sample for Great Britain, 1851 to 1901-The Victorian Panel Study (VPS). *Historical Social Research/Historische Sozialforschung*. 32(2):211-331.
- Smith, A. (2007) *Anarchism, revolution, and reaction: Catalan labour and the crisis of the Spanish state, 1898-1923*, New York.
- Soehl, T. (2017) Social reproduction of religiosity in the immigrant context: The role of family transmission and family formation—evidence from France, *International Migration Review* 51, 4: 999-1030.
- Soria Mesa, E. (2013). *La nobleza en la España moderna: cambio y continuidad*. Madrid. Marcial Pons Historia.
- Suchanek, F.; Abiteboul, S.; Senellart, P. (2011) PARIS: probabilistic alignment of relations, instances, and schema, *CoRR* (2011). 1111.7164
- Sun, Z.; Hu, W.; Zhang, Q.; Qu, Y. (2018) Bootstrapping entity alignment with knowledge graph embedding, in: *27th International Joint Conference on Artificial Intelligence*: 4396–4402.
- Szołtysek, M.; Gruber, S. (2015) Mosaic: recovering surviving census records and reconstructing the familial history of Europe. *The History of the Family* 21, (1): 38-60.
- Tavera, S. (2002) La historia del anarquismo español: una encrucijada interpretativa nueva, *Ayer*, 45, (1):13-37.
- Tejada, S.; Knoblock, C. A.; Minton, S. (2001) Learning object identification rules for information integration, *Inf. Syst.* 26: 607–633.
- Thorvaldsen, G., Andersen, T., & Sommerseth, H. L. (2015). Record linkage in the historical population register for Norway. In *Population reconstruction*: 155-171. Springer, Cham.
- To Figueras, L. (1993). Le mas catalan du XIIIe s.: Genese et evolution d’une structure d’encadrement et d’asservissement de la paysannerie *Cahiers de civilisation medievale*, 36(142), 151–177.
- Torras, J. (1984). Especialización agrícola e industria rural en Cataluña en el siglo XVIII. *Revista de Historia Económica* 2 (03): 113-127.

- Torras, J. (1998) Small Towns, Craft Guilds and Proto-Industry in Spain, *Jahrbuch für Wirtschaftsgeschichte / Economic History Yearbook*, 39, (2): 79-96.
- Torrents, A. (1992) Actitudes públicas, actitudes privadas', *Boletín de la Asociación de Demografía Histórica*, 10 (1): 7-29.
- Torrents, A. (1993). *Transformacions demogràfiques en un municipi industrial català: Sant Pere de Riudebitlles, 1608-1935*. Tesis doctoral, Departament d'Història Contemporània. Universitat de Barcelona.
- Tribó, G. (1989) *Evolució de l'estructura agrària del Baix Llobregat, 1860-1931*. Doctoral Thesis doctoral. Departament d'Història Contemporània. Barcelona. Universitat de Barcelona.
- Trivedi, R.; Dai, H.; Wang, Y.; Song, L. (2017) Knowevolve: Deep temporal reasoning for dynamic knowledge graphs, *arXiv:1705.05742*.
- Trommsdorff, G. (2009) Intergenerational relations and cultural transmission" in Ute Schönplflug (ed.), *Cultural transmission: Psychological, developmental, social, and methodological aspects*, Cambridge: 126-160.
- Valero Escandell, J.R. (1986) La implantación del registro civil en España (problemas de utilización en estudios demográficos), *Historia Contemporánea*: 87-99.
- Van Leeuwen, M.; Maas, I. (2011). *HISCLASS (Historical International Social Class Scheme)*. Leuven, Leuven University Press.
- Van Leeuwen, M.; Maas, I.; Miles, A. (2002) *HISCO: Historical international standard classification of occupations*. Leuven, Leuven University Press.
- Van Poppel, F. (1995) Seasonality of work, religion and popular customs: the seasonality of marriage in the nineteenth and twentieth centuries Netherlands", *Continuity and Change*, 10, (2): 215-256.
- Verweij, J.; Ester, P.; Nauta, R. (1997) Secularization as an economic and cultural phenomenon: A cross-national analysis. *Journal for the Scientific Study of Religion*: 309-324.
- Vézina, H.; Bournival, J.S. (2020). An overview of theBALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*. Vol. 11, Special Issue 2.
- Vilalta, M.J. (1997). "La família en la Lleida del siglo XVI". Chacón, F.; Ferrer, Ll. (Cord.) *Familia, casa y trabajo*. Murcia, Universidad.

- Vilanova Ribas, M. (1973). Estructura demogràfica y económica en La Escala (Provincia de Gerona) en 1930-1940. (Aplicación de los datos procedentes del Padrón de Cédulas Personales). *Estudis: Revista de Historia Moderna*, 2:259-305.
- Villavicencio, F.; Jordà, J.P.; Pujadas-Mora, J.M. (2015) Reconstructing lifespans through historical marriage records of Barcelona from the sixteenth and seventeenth centuries. In *Population reconstruction*, Bloothoof, Gerrit, Peter Christen, Kees Mandemakers, and Marijn Schraagen, (eds). Springer, Cham, 199-216.
- Vohland, K.; Land-Zandstra, A.; Ceccaroni, L.; Lemmens, R.; Perelló, J.; Ponti, M.; Samson, R.; Wagneknecht, K. (eds.) (2021) *The Science of Citizen Science*. Springer, Switzerland.
- Wang, Z., Zhang, J.; Feng, J.; Chen, Z. (2014) Knowledge graph embedding by translating on hyperplanes, in: *AAAI*: 1112–1119.
- Ward, J.S.; Barker, A. (2013) Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.
- Weir, D. (1993) Family reconstitution and population reconstruction: two approaches to the fertility transition in France 1740-1911. In *Old and New Methods in Historical Demography*, edited by David Reher and Roger Schofield. Oxford, Clarendon Press: 145-58.
- Willigan, D.; Lynch, K. (2003) *Sources and methods of Historical Demography*, London.
- Winkler, W. (1999) The state of record linkage and current research problems, *Stat. Med.* 14.
- Woolf, S. (1989). Statistics and the modern state. *Comparative Studies in Society and History*. 31(3):588-604.
- Wrigley, E.A.; Schofield, R., (1989) *The population history of England 1541-1871*. Cambridge University Press.
- Yule, G.U.; Kendall, M.G. (1919) *An Introduction to the Theory of Statistics*, Charles Griffin and Company.
- Zijdeman, R. (2009) Like my father before me: Intergenerational occupational status transfer during industrialization (Zeeland, 1811–1915), *Continuity and Change*, 24, (3): 455-486.
- Zofio Llorente, J. C. (2005). *Gremios y artesanos en Madrid, 1500–1650: La sociedad del trabajo en una Ciudad cortesana preindustrial*. Madrid: Instituto de Estudios Madrileños.



UAB
Universitat Autònoma
de Barcelona

**CED**
Centre d'Estudis
Demogràfics