UNIVERSITÀ DI PISA
Corso di Laurea Magistrale in
Informatica Umanistica

Tesi di Laurea

MRCA e MCRA: proprietà statistiche delle genealogie nobiliari europee nell'età moderna

Relatore:
Prof. Paolo Rossi

Candidato:
Giorgio Spugnesi

Anno Accademico 2013-2014

A mia moglie Chiara,
presente della mia genealogia personale

Ringraziamenti

Il primo e più importante ringraziamento va a mia moglie Chiara per avermi sempre incoraggiato e sostenuto anche quando, come è mia abitudine, ripetevo che non avrei potuto farcela. Grazie per i fine settimana sacrificati allo studio, per aver spesso dovuto fare a meno dell'uso del computer, per essere stata presente nei momenti in cui io ero “altrove”.

Un ringraziamento particolare va anche ai miei genitori che mi hanno trasmesso il valore della conoscenza e mi hanno sempre sostenuto nelle mie peregrinazioni formative.

Sul versante accademico, ringrazio di cuore il mio relatore, il prof. Paolo Rossi. Sono rimasto affascinato dal suo approccio alle discipline umanistiche fin dall'esame che ho sostenuto con lui. Senza la sua disponibilità e la sua passione, questa ricerca non avrebbe mai visto la luce. Grazie per essere stato sempre presente e per aver vissuto in prima persona il travaglio di questa ricerca. In alcuni momenti si sarebbe potuto pensare che, alla fine, avremmo dovuto discuterla entrambi, questa tesi.

Grazie anche a tutti i compagni di studio conosciuti virtualmente sull'ormai scomparso forum www.infouma.net: senza di loro avrei avuto non poche difficoltà a completare il percorso di studi senza frequentare i corsi.

Infine, anche se non leggerà mai queste righe, mi sento in dovere di ringraziare Leo van De Pas per il titanico lavoro di raccolta delle genealogie rese disponibili sul suo sito www.genealogics.com senza le quali tutto questo lavoro sarebbe stato impossibile.

Indice generale

Introduzione

Il presente lavoro di tesi intende affrontare, attraverso l'uso di tecnologie informatiche, l'analisi di alcune proprietà statistiche degli alberi genealogici ed in particolare validare attraverso i dati l'esistenza di un modello nella individuazione dell'MRCA (Most Recent Common Ancestor) ovvero il più recente antenato comune all'interno di un gruppo di soggetti.

Se la modellazione matematica del MRCA vanta una tradizione ormai consolidata, soprattutto in discipline come la biologia e in particolare la genetica che affrontano il problema dal punto di vista cromosomico, sono pochi gli studi che analizzano la coalescenza degli antenati dal punto di vista della relazione parentale genitore-figlio intesa come processo sociale. E sono praticamente assenti studi in grado di verificare il modello accedendo ad un numero di dati sufficientemente ampio da fornirne una prova sperimentale. Il presente lavoro intende offrire un primo contributo nel colmare questa lacuna, sia creando una base dati ampia e sufficientemente attendibile da poter essere utilizzata come benchmark, sia verificare su di essa, in modo sperimentale, alcune intuizioni ancora in attesa di conferma.

In un'ottica più ampia, il lavoro rientra inoltre nel progetto Hochadelsdorf[1] teso ad analizzare le dinamiche familiari all'interno di un gruppo chiuso quale la nobiltà europea dell'età moderna. Il progetto si pone lo scopo di effettuare

[1] Si veda la presentazione del progetto http://www.df.unipi.it/~rossi/Hochadelsdorf%20project.pdf

un'analisi statistica diacronica delle genealogie e delle relazioni familiari all'interno di questo "villaggio"[2], puntando ad ottenere misure quantitative ed a fornire una interpretazione dinamica delle loro caratteristiche. La base dati che è stata costruita come oggetto del presente lavoro potrà in seguito costituire la popolazione del "villaggio" su cui compiere tali studi.

Dopo aver affrontato, nel Capitolo 1, una panoramica degli studi condotti sulle genealogie nell'ambito della fisica statistica e le motivazioni che spingono i fisici ad occuparsi di un fenomeno apparentemente più consono alle scienze umanistiche, nel Capitolo 2 saranno presentate le procedure di estrazione del campione da due delle principali basi dati genealogiche online. La presenza di siti web che raccolgono una enorme quantità di dati genealogici, più o meno attendibili, è stata infatti il requisito fondamentale per poter cominciare a studiare le genealogie da un punto di vista quantitativo basandosi su dati e non solo su modelli.

Nel Capitolo 3 vengono descritti gli algoritmi utilizzati per l'estrazione e l'elaborazione dei dati, a supporto della natura informatica di questo lavoro di tesi e nell'ottica di fornire una "cassetta degli attrezzi" applicativa per l'approccio a quella forma particolare di albero binario che è l'albero genealogico.

Il Capitolo 4 è dedicato alla analisi del MRCA (Most Recent Common Ancestor). Dopo una formulazione teorica alla luce della letteratura in materia, afferente prevalentemente all'ambito biologico e genetico, saranno presentate le

[2] Hochadelsdorf significa infatti "villaggio dell'alta nobiltà"

tre metriche utilizzate nell'individuazione dell'antenato comune più recente (anno di nascita, generazione minima, generazione media) e saranno commentati i risultati ottenuti dall'analisi degli alberi degli antenati sia suddivisi in generazioni genealogiche che anagrafiche.

Nel Capitolo 5 sarà introdotto il concetto di MCRA (Most Common Recent Ancestor), non un semplice anagramma della sigla ma una chiave di lettura del fenomeno della ripetizione degli antenati che è stato oggetto di studio a partire dagli anni '90 del XX secolo e sul quale si attendono ancora verifiche sperimentali.

Infine, nelle Conclusioni si aprirà la strada a futuri sviluppi della ricerca sulle genealogie e alle sue possibili applicazioni nell'analisi delle dinamiche di gruppi chiusi di popolazioni. Alcune di queste ricerche si sperano essere naturale completamento del lavoro avviato con il presente elaborato di tesi.

Capitolo 1:

Lo studio delle genealogie, tra fisica e informatica

L'approccio alle genealogie dal punto di vista della fisica statistica potrebbe apparire piuttosto inconsueto, pensando che sia materia più adatta agli storici. In realtà l'argomento vanta una trattazione piuttosto ampia che prende avvio da alcune considerazioni che andremo ad analizzare.

Le genealogie, come i fenomeni critici in microfisica, sono sistemi privi di scala[3] ovvero sono sistemi nei quali non esistono valori medi né le loro distribuzioni si concentrano in intervalli particolari. Le "misure" pertanto saranno confrontabili indipendentemente dalla dimensione dell'albero genealogico osservato. Le genealogie sono inoltre caratterizzate dalla universalità ovvero la caratteristica per la quale differenti distribuzioni empiriche possono ricondursi alla stessa forma matematica, indipendentemente dai parametri specifici del campione. Questo fenomeno si manifesta prevalentemente nella teoria statistica della consanguineità tra antenati, esplorata per la prima volta da Derrida e colleghi [Derrida et al. 1999]. Tale studio è condotto su una simulazione numerica basata su un modello teorico di popolazione chiusa con riproduzione sessuale e generazioni non sovrapposte. Lo studio tuttavia è privo di una verifica sperimentale ampia, limitandosi all'analisi di un solo albero genealogico, quello di Edoardo III d'Inghilterra

[3] Sui fenomeni privi di scala si veda [Newman 2005]

(1312 – 1377). Il presente lavoro ha tra i suoi scopi anche quello di fornire un campione significativo su cui verificare in futuro le teorie formulate.

Un altro aspetto degli alberi genealogici di interesse per la fisica è il loro essere sistemi autosimilari: ogni albero è strutturalmente simile in ogni sua posizione. Qualsiasi frammento di albero genealogico è infatti simile a qualunque altro e all'albero stesso, indipendentemente dalle dimensioni considerate. Tale proprietà è presente in molti fenomeni naturali, non solo in fisica ma anche in biologia. Potremmo dire che l'albero genealogico è un frattale.

La costruzione di un albero genealogico è inoltre un processo di branching[4] ovvero un processo che riguarda oggetti che possono riprodursi generando oggetti dello stesso tipo, in questo caso esseri umani. È un fenomeno studiato in fisica nelle reazioni a catena o nei processi a cascata come la propagazione dei neutroni in una reazione nucleare.

Infine, essendo le genealogie sistemi caratterizzati da grandi numeri, possono essere, e vengono, trattate con i metodi propri della fisica statistica.

Lo studio degli alberi genealogici quindi, quando non sia affrontato con l'intento di ricavare informazioni storiche su un certo soggetto o una certa famiglia, è tradizionalmente campo d'azione dei fisici.

Sono infatti numerosi gli articoli di fisici, e non solo di genetisti, sulla distribuzione dei cognomi [Fox et al. 1983][Rossi 2013], sull'estinzione delle famiglie [Lotka 1931], su isonimia e consanguineità [Crow et al. 1965] [Lasker 1977].

[4] Sui processi di branching si veda [Athreya 2004]

Diverso è il discorso per quanto riguarda gli informatici. Come si approcciano alle genealogie?

Se gli alberi binari fanno parte della teoria informatica, se non altro per gli algoritmi di ricerca, diversa è la loro applicazione in ambito genealogico. Non sembrerebbero esistere pubblicazioni in materia. Tuttavia, come vedremo nel prossimo capitolo, la genealogia si è avvalsa dello sviluppo degli strumenti informatici per compiere un balzo avanti in termini di diffusione e di possibilità di mettere in relazione i dati.

Più che all'elaborazione, l'informatica ha contribuito alla conservazione e alla consultazione dei dati genealogici, attraverso l'uso dei database relazionali e la creazione di pagine web che ne consentissero l'accesso e l'interrogazione.

L'attenzione e l'interesse, in tempi recenti, per i big data coinvolge in un certo senso anche i database genealogici che vanno sempre più arricchendosi di informazioni che hanno bisogno di essere processate ed estratte perché possano acquistare un valore statistico oltre quello documentale.

Trattandosi di dati decisamente umanistici, non solo perché costituiscono informazioni storiche ma anche perché coinvolgono individui realmente vissuti e relazioni matrimoniali e genitoriali, la figura dell'informatico umanistico dovrebbe trovarsi a proprio agio in questo contesto ed essere in grado di fornire il proprio contributo alla ricerca attingendo ad un ventaglio di discipline ampio.

Rispetto ai modelli puramente teorici, nelle genealogie reali fenomeni come seconde nozze, sovrapposizione di generazioni, omonimie spesso anche tra

padre e figlio, fanno parte del campione e possono necessitare di un supplemento di analisi che può richiedere anche competenze non puramente informatiche.

Capitolo 2:

Genealogie digitali ed estrazione del campione

L'avvento del digitale e la diffusione di Internet hanno fatto sì che le genealogie di personaggi più o meno famosi cominciassero ad essere disponibili in formato elettronico non solo agli storici o agli appassionati di araldica più o meno improvvisati ma a chiunque intenda ricostruire la Ahnentafel (o tavola degli antenati) di un personaggio e indagarne le relazioni di parentela con altri soggetti. Questa analisi è resa agevole dall'uso di database relazionali che permettono di immagazzinare le informazioni e porle tra loro in relazione.

Gli schemi tradizionali (Ahnentafel, alberi dei discendenti, ecc.) divengono adesso generabili in modo semplice ed immediato e sono al contempo possibili interrogazioni sui dati maggiormente dinamiche, come il calcolo automatico della parentela, o approfondite, come la possibilità di ottenere informazioni di natura storica o prosopografica se aggiunte ai soggetti.

Infine, la presenza di una struttura nei dati e la loro disponibilità in formato elettronico hanno permesso, ed è un aspetto fondamentale per il presente lavoro, di estrarre informazioni di tipo statistico.

Accanto a questi vantaggi si collocano però dei *caveat* che è bene tener presenti nell'accedere alle risorse genealogiche online. I dati infatti sono spesso di bassa qualità, derivanti da fonti secondarie o, ancor peggio, privi di fonti

documentate e dichiarate. Fenomeno ancor peggiore è la derivazione dei dati da altri siti web con la conseguenza di far propagare in modo incontrollabile l'errore.

Se da un punto di vista storico, questo approccio può risultare intollerabile ai fini della ricerca, da quello statistico, pur rappresentando un errore sistematico non sempre trascurabile, esso è compensato dalla quantità di dati che diviene possibile processare in modo automatico.

Non mancano comunque basi dati costruite con serietà ed onestà intellettuale oltre che decisamente ampie. Molte di queste raccolgono genealogie familiari, messe insieme da soggetti interessati ad individuare i propri antenati ma senza alcuna, o con scarsa, utilità storica.

Vi sono infine alcune basi dati nate con lo scopo (titanico) di mettere insieme le genealogie dell'intera nobiltà europea. Sono opere costruite con criteri scientifici, spesso dichiarati in modo esplicito nel sito che le ospita, e attingendo a fonti quasi sempre dichiarate.

In questo contesto emergono due basi dati che per ampiezza, completezza e attendibilità del dato e possibilità di ricerca sono state scelte per ottenere la base dati necessaria al presente lavoro. In particolare, i dati sono stati estratti, con i procedimenti che descriveremo in seguito, dal sito Genalogics[5] e quindi integrati con alcuni dati derivanti da Roglo[6].

Non ci dilungheremo a descrivere i due siti e le relative basi dati sia perché

[5] http://www.genalogics.com
[6] http://www.roglo.eu/roglo

entrambi forniscono sufficienti informazioni utili a valutare la qualità dei dati e comprendere le procedure attuate per la loro costituzione sia perché essi sono serviti esclusivamente come fonte cui attingere con tecniche di data mining, purtroppo non interamente automatizzato, per la costituzione della base dati da studiare.

Entrambi i siti non consentono né l'accesso diretto al database tramite script né l'utilizzo di strumenti di data mining automatizzati. È stato pertanto necessario generare manualmente le tavole degli antenati di interesse, interrogando il sito web, salvare tali tavole in formato testuale e operare il mining sui file di testo, un processo piuttosto lungo e laborioso, data anche l'enorme quantità di dati trattata, di cui analizzeremo in seguito i dettagli e le problematiche.

È stato d'aiuto il fatto che entrambi i siti facciano uso, nell'individuazione degli antenati, del sistema Sosa-Stradonitz uno dei più usati sistemi di numerazione per le genealogie, ideato nel XVII secolo da Jeronimo de Sosa e reso popolare, nella seconda metà dell'800, da Stephan Kekulé von Stradonitz. Il sistema si basa sul numero 2 e sulle sue potenze: il numero del padre è sempre il doppio di quello del figlio e quello della madre, il doppio più uno. Gli uomini, in questo modo, hanno sempre numero pari e le donne dispari. Inoltre ogni generazione è una successiva potenza di 2 e il numero totale di persone fino alla generazione n è 2^n-1. Data la natura matematica del sistema, è stato semplice navigare, con sistemi informatici, all'interno degli alberi individuando legami e relazioni di parentela in modo automatico ed elaborare i tracciati

basandosi su questo indicatore numerico.

Preso atto della possibilità di avere accesso ad una grande quantità di dati, il primo passo è stato determinare il campione iniziale, ovvero il gruppo degli individui di cui andare a costruire le Ahnentafeln.

Lo studio prevede l'analisi di un campione che rappresenti in sé un gruppo ben definito e sufficientemente chiuso come lo è stata la nobiltà europea nell'età moderna. L'avvento della Grande Guerra ha certamente determinato la disgregazione di tale gruppo, sia dal punto di vista del ruolo sociale e politico dei suoi membri, sia da quello delle politiche matrimoniali tese a consolidare tale potere. Pertanto il secondo ventennio del '900 segna il limite superiore nella scelta del campione. Da esso tuttavia non ci si è scostati molto per avere margine sufficiente ad individuare 16 generazioni di antenati non troppo lacunose. Il campione selezionato è composto da 200 individui dei quali solo 149 sposati e solo 128 fertili. Escludendo i fratelli e le sorelle, che non apportano ulteriori informazioni sugli antenati in quanto le Ahnentafeln coincidono, sono stati scelti 48 soggetti nati tra il 1865 e il 1917 e rappresentanti tutte le famiglie della nobiltà europea, salvo alcune estinte qualche generazione prima ma che rientreranno nelle genealogie dei soggetti scelti. A ciascuno di essi è stato attribuo un numero da 1 a 48 indicante l'albero di appartenenza.

Per quanto riguarda le generazioni successive alla prima, per l'analisi genealogica, il campione è composto dai genitori dei soggetti della generazione

precedente, senza escludere eventuali fratelli e sorelle, mentre per quella anagrafica, il campione è costituito dai soggetti nati all'interno di fasce di anni prestabilite. Nella selezione dei campioni successivi al primo si è ritenuto opportuno escludere alcuni soggetti spurii, non appartenenti alle famiglie nobili ma entrati nell'Ahnentafel per matrimoni con discendenti di famiglie non facenti parte del gruppo. Da un punto di vista delle dinamiche della popolazione, tali soggetti costituiscono un fenomeno di immigrazione che non verrà preso in considerazione nel presente lavoro.

Dal punto di vista informatico si è optato per la creazione di un database SQLServer. Le proprietà relazionali di tale database, sebbene non necessarie per l'analisi della base dati, sono state utili per implementare l'estrazione dei dati consentendo di creare tabelle di utilità relative a soggetti con particolari caratteristiche, soprattutto nelle fasi di disambiguazione. L'uso di indici ha permesso di velocizzare le elaborazioni e le interrogazioni sulla tabella principale dei soggetti.

La parte di scripting per l'estrazione dei dati è stata sviluppata in C#. Sebbene tale linguaggio non sia spesso considerato adatto alle *computational sciences*, preferendogli Python o R, si è invece dimostrato sufficientemente flessibile nell'analisi di grandi file di testo, nella elaborazione delle stringhe e, come ci saremmo aspettati, nell'interazione con il database. Il progetto prevede una serie di console applications dedicate a singoli task ed una libreria denominata Sosa contenente le funzioni di trasformazione dei numeri Sosa. Maggiori

dettagli sull'implementazione degli algoritmi saranno presentati nel prossimo capitolo.

Predisposti gli strumenti informatici per l'estrazione e l'elaborazione delle informazioni, si è potuto procedere con l'approccio al sito Genealogics.

Genealogics non consente la visualizzazione di alberi oltre la 12ª generazione pertanto si è dovuto operare estraendo 48 alberi fino alla 5ª generazione e successivamente, individuati per ciascun albero i 31 soggetti della 5ª generazione (in realtà alcuni meno in virtù della ripetizione degli antenati), per essi scaricare manualmente gli alberi fino alla 12ª generazione.

L'estrazione dei dati è stata realizzata mediante il parsing del file di testo e l'uso di operazioni su stringa per individuare i valori di interesse ovvero, oltre all'albero di appartenenza, il numero Sosa, il nome, le eventuali date di nascita e di morte e il solo anno di nascita e di morte utili questi ultimi, come vedremo, per la disambiguazione dei soggetti omonimi.

Mediante un processo di riassegnazione dei Sosa secondo la formula

$$\text{Sosa} = 2^{\lfloor log2(a) \rfloor} b + (a - 2^{\lfloor log2(a) \rfloor})$$

con a = Sosa da trasformare e b = Sosa del soggetto radice nell'albero di destinazione, è stato possibile unire, per ciascun soggetto iniziale, le sue 5 generazioni con le 12 dei suoi antenati alla 5ª generazione ottenendo un albero completo fino alla 16ª generazione.

Per ciascun albero sono stati quindi ricavati $2^{16} - 1$ soggetti per un totale di 3.145.680 soggetti.

Appare evidente che si è in presenza di una base dati di considerevoli dimensioni che può, a tutti gli effetti, rientrare nel concetto di big data e come tale ha richiesto elaborazioni automatiche e massive.

Ovviamente non tutti gli alberi si presentano completi, soprattutto nelle generazioni più antiche, contenendo molti soggetti anonimi. Vi è inoltre un certo numero di soggetti omonimi e come tali ambigui, in mancanza di ulteriori specificazioni. Gli anni di nascita e di morte costituiscono un'informazione disambiguante: qualora due persone abbiano lo stesso nome e le stesse date di nascita e di morte, il soggetto è ragionevolmente lo stesso.

Per i soggetti ambigui privi di data, è stato individuato in modo automatico il coniuge secondo la formula

($Sosa - 1$) se il Sosa è dispari (e quindi è una donna)

($Sosa + 1$) se il Sosa è pari (e quindi è un uomo)

e, qualora il coniuge avesse specificata una o entrambe le date, queste sono state assegnate al soggetto tra parentesi.

Ai soggetti rimasti privi di data ma ambigui perché presenti più volte si è provveduto ad aggiungere il nome del coniuge nella forma "Nome cg Nome_coniuge".

A questo punto si è ritenuto opportuno tentare l'integrazione con i dati presenti in Roglo. Questo sito consente la generazione di Ahnentafeln anche di 16 generazioni pertanto è stato sufficiente scaricare 48 file di testo da sottoporre al parsing.

Dalla base dati sinora costituita sono stati estratti i soggetti con uno o entrambi i genitori anonimi a partire dai quali tentare di ricostruire le lacune attingendo ai dati di Roglo. Questi soggetti sono stati confrontati con i corrispettivi in Roglo (stesso albero, stesso Sosa) in modo da essere certi che si trattasse della stessa persona e non vi fossero false assegnazioni.

Dei soggetti per i quali si è avuta la certezza che fossero gli stessi in Genealogics e in Roglo, si è proceduto ad estrarre gli antenati dalle tavole di Roglo e ad integrarli nella base dati, identificando così 31033 soggetti.

Infine sono stati individuati i coniugi dei soggetti rimasti completamente anonimi e, se presenti, si è assegnato al soggetto anonimo un nome composto da "NN cg" seguito dal nome della moglie per gli uomini e "NeN cg" seguito dal nome del marito per le donne. Qualora fossero presenti la data di nascita o di morte del coniuge, esse sono state assegnate, tra parentesi, anche al soggetto.

A questo punto, quella che si è ottenuta è la più completa base dati ricavabile dalle fonti prescelte. Essa presenta tuttavia ancora un certo numero di soggetti anonimi (521698 in totale, v. Tabella B in Appendice) che si è provveduto ad identificare assegnando loro la sigla "NN" per gli uomini e "NeN" per le donne, seguita dal simbolo #, dal Sosa relativo al più prossimo discendente noto e dal nome di quest'ultimo. Sebbene questo procedimento non consenta di identificare con precisione se si tratti o meno dello stesso individuo (es. padre di più figli presente nei vari alberi), si è considerato tale errore trascurabile.

Lo studio del MRCA è stato condotto su più campioni appartenenti a

generazioni sia genealogiche che anagrafiche successive. Per le generazioni genealogiche si è provveduto ad estrarre tutti i soggetti univoci della seconda generazione in modo da costituire il campione. Da esso sono stati rimossi gli spurii ed è stato assegnato a ciascun soggetto un numero progressivo indicante l'albero di appartenenza. Attraverso un procedimento ricorsivo sono stati ricostruiti gli alberi degli antenati, attingendo alla base dati della generazione precedente, ottenendo alberi sempre più corti di una generazione ma comunque sufficientemente lunghi per compiere l'analisi degli antenati comuni. Alla 6ª generazione infatti si dispone ancora di un albero di 11 generazioni.

Un procedimento analogo è stato condotto per la suddivisione in generazioni anagrafiche. Sono state individuate, negli anni tra il 1910 e il 1730, 6 generazioni della durata di 30 anni ed il campione è stato diviso in base all'anno di nascita. Eliminati i soggetti spurii, si è provveduto a generare i relativi alberi degli antenati partendo dai dati della base dati iniziale.

Infine dalla base dati è stata estratta una matrice tridimensionale avente come dimensioni gli individui, gli alberi e le generazioni. I valori di incrocio delle dimensioni rappresentano le frequenze. Dal punto di vista simbolico essa è rappresentata dalla funzione

$$R(i, a, g)$$

che indica il numero di ricorrenze dell'antenato *i*, nell'albero *a*, alla generazione *g*.

Da tale funzione si possono ricavare due funzioni ausiliarie. La prima

$$S(i,g)=\sum_{a} R(i,a,g)$$

rappresenta il numero di ricorrenze di *i* nella generazione *g*. Essa genera una matrice bidimensionale con 16 colonne, una per ciascuna generazione, da cui ricavare il MCRA.

La quantità

$$w(i,a,g)=\frac{R(i,a,g)}{2^{g-1}}$$

rappresenta il peso dell'antenato *i* nell'albero *a* alla generazione *g*. In questo modo, i valori di frequenza possono essere confrontati tra loro tra generazioni.

La quantità

$$w(i,a)=\sum_{g} w(i,a,g)$$

rappresenta il peso totale dell'antenato *i* nell'albero *a* mentre

$$w(i,g)=\sum_{a} w(i,a,g)$$

rappresenta il peso totale dell'antenato *i* alla generazione *g*. L'individuazione del MCRA per generazione farà quindi riferimento a questo ultimo valore all'interno della matrice a 16 colonne.

L'altra formula ausiliaria ricavabile

$$T(i,a)=\sum_{g} R(i,a,g)$$

rappresenta il numero di ricorrenze di *i* nell'albero *a*. Essa genera una matrice bidimensionale con 48 colonne, una per ciascun albero, e consente di ricavare il MRCA, individuando l'antenato *i* presente in tutte le colonne *a*, alla

generazione *g*.

La distribuzione in frequenza degli antenati è rappresentata da

$$P(w, a)$$

ovvero il numero degli antenati che compaiono con peso *w* nell'albero *a*.

Per ridurre il rumore nella distribuzione si è fatto ricorso alla distribuzione cumulativa data da

$$C(w, a)$$

ovvero il numero di antenati che compaiono con peso maggiore o uguale a *w* nell'albero *a* considerando la relazione

$$P(w) = C(w) - C(w + \Delta w)$$

dove Δw è la differenza tra due valori successivi di *w*.

Capitolo 3:

Algoritmi e metodi

Una delle componenti fondamentali del presente lavoro, e senza dubbio il contributo più interessante al futuro delle ricerche in ambito genealogico, è stata la creazione di un'ampia base dati informatizzata. Per l'archiviazione dei dati si è fatto uso della piattaforma SQLServer nella sua versione 2014 Express.

La natura dei dati, derivanti come già detto da file di testo riportanti le tavole degli antenati, non è in sé relazionale quanto piuttosto lineare ma la possibilità di mettere in relazione tra loro i dati ed estrarre informazione mediante il linguaggio T-SQL si è rivelata estremamente utile nelle elaborazioni delle singole generazioni sia genealogiche che anagrafiche.

Tutta la gestione e l'interrogazione della base dati è stata condotta attraverso query T-SQL mentre per la parte di analisi è stato utile anche il foglio di calcolo Calc della suite OpenOffice (sostanzialmente simile a Excel di Microsoft).

La struttura del database, se si escludono tabelle provvisorie usate per particolari operazioni, è composta da poche tabelle di base. Si è volutamente mantenuto una struttura quanto più simile a quella di un tracciato dati in formato testo in modo da mantenere l'estrazione futura dei dati quanto più lineare e adatta all'analisi anche mediante strumenti di parsing come il

linguaggio R, uno standard di fatto nell'approccio ai big data. Non è da escludere, in futuri utilizzi di questa base dati, l'opportunità di aggiungere una chiave univoca assegnando a ciascun soggetto un codice identificativo numerico impostato come chiave primaria sulla tabella.

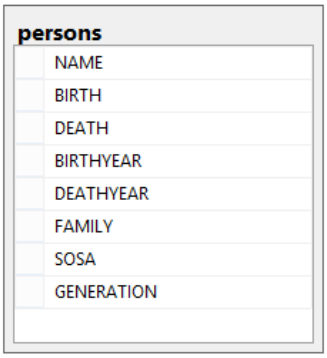

Figura 1: Tabella *persons*

La tabella *persons* (Figura 1) contiene tutti i soggetti della base dati ovvero gli antenati fino alla 16ª generazione dei 48 soggetti prescelti come campione di base per un totale di 3.145.680 records. I campi censiti sono di seguito descritti.

NAME (*varchar*): il nome completo del soggetto o quello ricavato dal coniuge o dal discendente più prossimo come spiegato al Capitolo 2.

BIRTH (*varchar*) e **DEATH** (*varchar*): data completa di nascita o di morte, ove conosciute. Tale campo non è stato mai usato nelle analisi condotte ma è stato comunque censito per usi futuri.

BIRTHYEAR (*varchar*) e **DEATHYEAR** (*varchar*): anno di nascita e di morte, ove presente. Oltre ad essere informazioni di tipo cronologico, tali campi concorrono assieme a **NAME** alla disambiguazione del soggetto. Si è optato per

il tipo *varchar* in modo da poter usare le parentesi per le date assegnate dal coniuge. Solo in fase di interrogazione, con apposita funzione scalare, si è provveduto a convertire i dati in valori numerici.

FAMILY (*int*): identificativo dell'albero. È un numero progressivo che distingue ciascun albero all'interno della base dati. Nella tabella *persons* assume valori [1-48]. Ciascun record, in questo modo, appartiene ad uno degli alberi. Non esiste una tabella di transcodifica ma per individuare il soggetto di partenza è sufficiente individuare quello con SOSA = 1.

SOSA (*int*): codice Sosa del soggetto. E' sempre relativo al soggetto iniziale dell'albero di appartenenza pertanto nelle tabelle derivate da *persons* i Sosa saranno ricalcolati, come sarà spiegato in seguito.

GENERATION (*int*): la generazione genealogica a cui il soggetto appartiene. È un campo ridondante in quanto il Sosa porta in sé l'informazione sulla generazione ma è di estrema utilità nell'individuazione del MRCA e nelle elaborazioni dei dati sia da script che da query T-SQL.

Sulla tabella è stato creato un *clustered index* sui campi FAMILY e SOSA in modo da velocizzarne l'interrogazione.

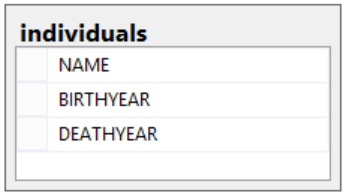

Figura 2: Tabella *individuals*

Dalla tabella *persons* deriva la tabella *individuals* (Figura 2), ottenuta raggruppando per NAME, BIRTHYEAR, DEATHYEAR. La tabella contiene tutti gli individui univoci della base dati. Si tratta quindi dell'elenco degli "abitanti del villaggio" nel periodo preso in considerazione (compresi gli "immigrati"). I campi non richiedono spiegazione in quanto derivanti da quelli della tabella *persons*.

Sempre dalla tabella *persons* sono state poi ricavate altre 5 tabelle identiche (*persons2*, *persons3*, …) una per ciascuna delle successive 5 generazioni genealogiche prese in considerazione. In ciascuna di esse sono contenuti tutti gli alberi relativi ai soggetti selezionati per le generazioni successive alla prima. Il campo FAMILY pertanto è stato riassegnato con valori da 1 a *n* con *n* = numero di soggetti selezionati e, allo stesso modo, i campi SOSA e GENERATION sono stati aggiornati con i valori relativi alla nuova posizione assunta da ciascun soggetto. Come aiuto nella elaborazione dello script, i dati relativi alla generazione precedente sono stati mantenuti in due campi aggiuntivi (ma superflui a fine elaborazione), **OLDFAMILY** e **OLDSOSA**.

Con un procedimento analogo sono state generate altre 6 tabelle (*generation1*, *generation2*, …) con i soggetti appartenenti al campione suddiviso per anno di nascita.

Infine, sempre partendo dalla tabella *persons*, sono state estratte 48 tabelle (*family1*, *family2*) con i soggetti dei singoli alberi dalle quali, attraverso l'istruzione PIVOT di T-SQL, sono state estratte le matrici di frequenza degli

antenati utili al calcolo del MCRA.

In sostanza, l'unica tabella veramente apportatrice di informazioni è la tabella *persons* da cui tutte le altre derivano e sono solamente strumentali alle analisi che devono essere condotte. Molte altre tabelle sono state create nel corso della definizione della base dati, raggruppando soggetti che necessitavano di essere disambiguati (omonimi, anonimi, "orfani") ma anche queste tabella hanno avuto solo funzione strumentale. L'intero database quindi è fondamentalmente composto da un'unica grande tabella. Per questo motivo, la scelta di mantenerlo quanto più simile ad un tracciato dati appare non priva di senso.

Prima di eseguire qualsiasi operazione di analisi, tuttavia, è stato necessario popolare la base dati, ovvero la tabella *persons*, con i dati estratti dai tracciati ricavati da Genealogics e Roglo. Questa parte di parsing, analisi delle stringhe di testo, elaborazione e scrittura su database è stata realizzata attraverso una serie di *console applications* sviluppate con il linguaggio C# attraverso il tool di sviluppo Microsoft Visual Studio e il Framework .NET 4.5.

A supporto di queste applicazioni è stata implementata un'ulteriore applicazione nominata "Sosa" e composta da una classe "Utility" contenente metodi per eseguire calcoli sui numeri Sosa. Come abbiamo avuto modo di vedere, il sistema Sosa è basato sulle potenze e sui multipli di 2, pertanto tutti i rapporti di parentela sono matematicamente ricavabili con facilità. La classe "Utility" espone i seguenti metodi principali:

- GetFather: dato un Sosa, restituisce il Sosa del padre;

- GetMother: dato un Sosa, restituisce il Sosa della madre;
- GetChild: dato un Sosa, restituisce il Sosa del figlio;
- GetGeneration: dato un Sosa, restituisce la generazione di appartenenza;
- Refactor: dato il Sosa di un soggetto e un altro Sosa da usare come radice dell'albero, restituisce il nuovo Sosa del primo soggetto relativo al secondo (usato per concatenare gli alberi);
- IsOdd: dato un valore numerico intero, restituisce TRUE se è dispari, FALSE se è pari (usato per distinguere i maschi dalle femmine).

Si riporta di seguito il codice della classe “Utility”

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;

namespace Sosa
{
    public class Utility
    {
        public static int GetFather(int child) {
            int father = child * 2;
            return father;
        }

        public static int GetMother(int child)
        {
            int mother = child * 2 + 1;
            return mother;
        }

        public static int GetChild(int parent) {
            if (IsOdd(parent)) {
                return (parent - 1) / 2;
            } else {
                return parent / 2;
            }
        }
```

```
        public static int GetGeneration(int sosa) {
            return Convert.ToInt32(Math.Truncate(Math.Log(sosa, 2))) + 1;
        }

        // trasforma il SOSA di un soggetto rispetto al sosa di un
              soggetto radice; serve per concatenare gli alberi
        // current: il SOSA del soggetto che stiamo analizzando
        // root: il sosa del soggetto cui andiamo ad aggiungere gli
              antenati
        public static int Refactor(int current, int root) {
            int result;
            // base = 2^int(log2(x))
            result = Convert.ToInt32(Math.Pow(2,

       Convert.ToInt32(Math.Truncate(Math.Log(current, 2)))));
            // sosa = base * root + (current - base)
            result = (result * root) + (current - result);
            return result;
        }

        public static bool IsOdd(int value)
        {
            return value % 2 != 0;
        }

    }
}
```

La libreria Sosa è stata poi importata nel progetto principale, denominato "TreeGenerator". Esso è un contenitore di *console applications* autonome, attivate e lanciate di volta in volta in base all'operazione da compiere.

La prima applicazione "mrca.cs" esegue il parsing degli alberi fino alla 5ª generazione secondo quanto descritto in Algoritmo 1.

Algoritmo 1: mrca.cs

1. Per ogni file contenente un albero da 1 a 48
2. Il nome del file è il valore di FAMILY
3. Per ogni riga del file
4. Se il primo carattere è numerico, i caratteri prima del primo punto (.)

sono il valore di SOSA, i restanti di NAME. Da SOSA ricava il valore di GENERATION

5. Se il primo carattere è "B" quello che segue è la data di nascita (BIRTH e BIRTHYEAR)
6. Se il primo carattere è "D" quello che segue è la data di morte (DEATH e DEATHYEAR)
7. Salva i dati come record

I file di testo estratti da Genealogics presentano la forma esemplificata in Tracciato 1.

```
1. Joachim Ernst, Herzog von Anhalt
B:   11 Jan 1901
P:   Dessau
D:   18 Feb 1948
P:   Buchenwald

2. Eduard, Duke von Anhalt
B:   18 Apr 1861
P:   Dessau
M:   6 Feb 1895
P:   Altenburg
D:   13 Sep 1918
P:   Berchtesgaden


4. Friedrich I, Duke von Anhalt
B:   29 Apr 1831
P:   Dessau
M:   22 Apr 1854
P:   Altenburg
D:   24 Jan 1904
P:   Schloss Ballensted


8. Leopold IV, Duke von Anhalt-Dessau
B:   1 Oct 1794
P:   Dessau
M:   18 Apr 1818
P:   Berlin
D:   22 May 1871
P:   Dessau
```

```
16. Friedrich, Erbprinz von Anhalt-Dessau     =>
B:   27 Dec 1769
P:   Dessau
M:   12 Jun 1792
P:   Homburg v.d.Hohe
D:   27 May 1814
P:   Dessau
```

Tracciato 1: Family 1 (estratto)

Le stringhe ricavate dal file di testo, prima di essere inserite come valori nel database sono state processate mediante splitting, parsing e regular expression in modo da convertirle nel formato adatto. Soprattutto le date, e in particolare quelle ricavate da Roglo, hanno richiesto un notevole lavoro di analisi del tracciato per individuare tutte le modalità di formattazione (non censite dal sito). In questo caso, data la mole di dati, si è spesso dovuto procedere per tentativi ed errori prima di ottenere un risultato soddisfacente.

L'inserimento dei record nel database non sempre è avvenuto in modo diretto dalla classe incaricata di estrarre i dati. Si è notato infatti che il principale "collo di bottiglia" nell'elaborazione dei dati era causato dalla scrittura su disco e che alcuni script necessitavano di tempi di esecuzione troppo lunghi. In molti casi, soprattutto in presenza di set di dati di grosse dimensioni, si è fatto ricorso a scritture massive salvando i risultati su file CSV ed eseguendo un import di tipo DTS (Data Transformation Services) dal Management Studio di SQLServer.

Dopo aver caricato le prime 5 generazioni nella tabella *persons*, si è

provveduto ad integrare le successive 12, dopo aver estratto i tracciati da Genealogics, attraverso l'applicazione “joiner.cs” descritta in Algoritmo 2.

Algoritmo 2: joiner.cs

1. Per ogni file contenente un albero
2. Il nome del file è composto dal valore di FAMILY e dal Sosa root
3. Per ogni riga del file
4. Se il primo carattere è numerico, i caratteri prima del primo punto (.) sono il Sosa, i restanti il valore di NAME
5. Se Sosa = 1: ignora il record
6. Il valore di SOSA è Refactor(Sosa, root)
7. Dal valore di SOSA calcola il valore di GENERATION
8. Se il primo carattere è “B” quello che segue è la data di nascita (BIRTH e BIRTHYEAR)
9. Se il primo carattere è “D” quello che segue è la data di morte (DEATH e DEATHYEAR)
10. Salva i dati come record

L'Algoritmo 2 è sostanzialmente simile all'Algoritmo 1 con la differenza che, dovendo unire due alberi genealogici, ignorerà il soggetto con Sosa = 1 poiché è il soggetto di 5ª generazione già presente in *persons* (il soggetto root su cui innestare l'albero) e ricalcolerà gli altri Sosa rendendoli relativi non al soggetto root ma alla radice reale dell'albero (ovvero il soggetto di partenza dell'albero estratto fino alla 5ª generazione).

La combinazione dei due algoritmi ha portato alla creazione della tabella *persons* completa, almeno dei Sosa se non dei nomi, per 16 generazioni dei 48 soggetti scelti come campione.

La fase di disambiguazione e assegnazione di nomi agli anonimi ha richiesto alcune elaborazioni sia a livello di database che di scripting.

L'assegnazione delle date del coniuge ai soggetti non anonimi ma privi di date è stata gestita con l'applicazione "spuosesdate.cs" descritta in Algoritmo 3. L'assegnazione delle date, come già ricordato in precedenza, è fondamentale per la disambiguazione in quanto l'univocità del soggetto è data dalla combinazione di nome e almeno uno tra l'anno di nascita o di morte.

Algoritmo 3: spousesdate.cs

1. Seleziona da *persons* tutti i soggetti con NAME non nullo e BIRTHYEAR e DEATHYEAR nulli
2. Per ciascun soggetto individua il Sosa del coniuge: (SOSA – 1) se IsOdd(SOSA), (SOSA + 1) altrimenti
3. Ricava da *persons* BIRTHYEAR e DEATHYEAR del coniuge
4. Se BIRTHYEAR non è nullo, assegnalo al soggetto come BIRTHYEAR tra parentesi
5. Se DEATHYEAR non è nullo, assegnalo al soggetto come DEATHYEAR tra parentesi

Per i soggetti ancora privi di data e ripetuti più volte (e quindi potenzialmente

ambigui) si è fatto seguire al nome quello del coniuge nella forma "Nome cg Nome_Coniuge". Per fare questo sono state estratte in una tabella temporanea *spouses* tutte le coppie in cui uno dei coniugi fosse privo di entrambe le date. L'estrazione è stata compiuta con l'applicazione "spouses.cs" (Algoritmo 4).

Algoritmo 4: spouses.cs

1. Seleziona da *persons* tutti i soggetti con NAME non nullo e BIRTHYEAR e DEATHYEAR nulli
2. Per ciascun soggetto individua il Sosa del coniuge: (SOSA – 1) se IsOdd(SOSA), (SOSA + 1) altrimenti
3. Ricava da *persons* NAME del coniuge
4. Inserisci in *spouses* il nome del soggetto e quello del coniuge

In questo modo sono stati individuati i soggetti presenti più volte con coniugi diversi e si è provveduto, mediante analisi manuale, ad individuare i soggetti distinti (due soggetti omonimi con due partner) rispetto a quelli che hanno avuto più partner (unico soggetto con più partner). Ai soggetti distinti è stato quindi assegnato, insieme al proprio nome, quello del partner in modo da distinguerli.

A questo punto si è ritenuto opportuno integrare la base dati con le informazioni provenienti da Roglo. Gli alberi estratti da Roglo in file di testo si presentano in modo lineare, ordinati per Sosa ed estesi per 16 generazioni. Presentano però gli antenati ripetuti in forma di rimando (nella forma Sosa y

=> Sosa x) che richiedono quindi di essere individuati esplicitamente. Le date inoltre fanno parte del nome ed è necessario separarle cercando di individuarne la forma.

Gli algoritmi che utilizzano i dati di Roglo hanno pertanto una fase di parsing per la creazione, in memoria, della hashtable {Sosa: nome} da cui attingere i dati.

I file di testo estratti da Roglo si presentano nella forma indicata in Tracciato 2.

```
Generazione 1
      1 - x x
Generazione 2
      2 - Eduard von Anhalt-Dessau, Herzog von Anhalt 1861-1918
      3 - x x
Generazione 3
      4 - Friedrich I von Anhalt-Dessau, Herzog von Anhalt
1831-1904
      5 - Antoinette, Prinzessin von Sachsen-Altenburg 1838-
1908
      6 - Moritz, Prinz von Sachsen-Altenburg 1829-1907
      7 - Augustine, Prinzessin von Sachsen-Meiningen und
Hildburghausen 1843-1919
Generazione 4
      8 - Leopold IV von Anhalt-Dessau, Herzog von Anhalt 1794-
1871
      9 - Friederike von Hohenzollern, Prinzessin von Preußen
1796-1850
      10 - Eduard, Prinz von Sachsen-Altenburg 1804-1852
      11 - Amalie, Prinzessin von Hohenzollern-Sigmaringen
1815-1841
      12 - Georg, Herzog von Sachsen-Altenburg 1796-1853
      13 - Maria, Herzogin von Mecklenburg-Schwerin 1803-1862
      14 - Bernhard II, Herzog von Sachsen-Meiningen und
Hildburghausen 1800-1882
      15 - Marie, Prinzessin von Hessen-Kassel 1804-1888
```

Tracciato 2: Family 1 (estratto)

Si noti in Tracciato 2 la presenza di “x x” per anonimizzare soggetti troppo recenti, una precauzione in termini di privacy dei redattori di Roglo

sicuramente eccessiva dato che i soggetti in questione sono personaggi di dominio pubblico. Fortunatamente si tratta di soggetti recenti e quindi sicuramente presenti in Genealogics.

L'estrazione dei dati è compiuta mediante le istruzioni di seguito riportate.

```
string filename = nomefile.Replace("INPUT_ROGLO\\", "");
string family = filename.Replace(".txt", "");

if (File.Exists("working.txt")) { File.Delete("working.txt"); }

// Rimuove gli spazi iniziali dal file
File.WriteAllLines(
       "working.txt",
       File.ReadAllLines(nomefile, Encoding.Default).Select(line =>
       line.Trim()
       ).ToArray()
       );

// Genera la prima hastable codice:nome
Hashtable ht = new Hashtable();
string[] readText = File.ReadAllLines("working.txt");
foreach (string s in readText)
{
       if (!s.ToUpper().Contains("GENERAZIONE"))
       {
              if (s.Contains(" - "))
              {
                     string[] tokens = s.Split(new string[] { " - " },
                            StringSplitOptions.None);
                     string name = tokens[1].ToString();
                     if (name == "Ne N" || name == "N N" || name == "x x"
                     || name == "? ?") {
                            name = string.Empty;
                     }
              ht[Convert.ToInt32(tokens[0].ToString().Replace(".", ""))]
```

```
                    = name;
                }
                if (s.Contains(" => "))
                {
                   string[] tokens = s.Split(new string[] { " => " },
                            StringSplitOptions.None);
                      ht[Convert.ToInt32(tokens[0].ToString().Replace(".",
                            ""))] = tokens[1].ToString().Replace(".", "");
                }
         }
}

// Genera la hashtable con i riferimenti espliciti
Hashtable ht2 = new Hashtable();
foreach (DictionaryEntry pair in ht)
{
       if (Extension.IsNumeric(pair.Value.ToString()))
       {
       ht2.Add(pair.Key, ht[Convert.ToInt32(pair.Value)].ToString());
       } else
       {
       ht2.Add(pair.Key, pair.Value);
       }
}

// Genera il dizionario ordinato per codice
IDictionary<int, string> d = HashtableToDictionary<int, string>(ht2);
var ordered = d.OrderBy(p => p.Key).ToDictionary(p => p.Key, p =>
p.Value);
```

L'integrazione dei dati di Roglo nella base dati non è stata compiuta in modo massivo per il timore che vi fossero false assegnazioni ovvero che alcuni soggetti presenti in Roglo non corrispondessero a quelli di Genealogics a causa di false genealogie o attribuzioni errate. Nella costruzione della base dati,

Genealogics costituisce la fonte più autorevole pertanto i dati di Roglo sono stati integrati solo se coerenti. Per fare ciò si è proceduto estraendo dalla base dati gli "orfani" ovvero quei soggetti di cui uno o entrambi i genitori fossero anonimi utilizzando l'applicazione "orphans.cs" (Algoritmo 5).

Algoritmo 5: orphans.cs

1. Per ogni albero, seleziona da *persons* tutti i soggetti e crea una hashtable {Sosa:nome}
2. Seleziona da *persons* i soggetti senza NAME
3. Per ciascuno di questi soggetti seleziona dalla hashtable il soggetto figlio e salva SOSA, NAME e "P" se padre, "M" se madre nella tabella *orphans*

Nella tabella *orphans* creata dall'Algoritmo 5 sono stati poi inseriti i soggetti corrispondenti (stesso albero, stesso Sosa) presenti nei file di Roglo usando l'applicazione "comparer.cs" (Algoritmo 6).

Algoritmo 6: comparer.cs

1. Per ogni albero, crea l'hashtable con i dati di Roglo
2. Per ogni albero seleziona da *orphans* tutti i soggetti
3. Se il soggetto è presente nella hashtable, inserisci il nome nel campo ROGLO della tabella *orphans*

Affiancando i nomi dei soggetti in Genealogics e in Roglo, è stato possibile individuare i soggetti corrispondenti (quelli per i quali il nome era identico o ragionevolmente lo stesso, differente magari solo per grafia). Questi soggetti sono stati marcati come "updatable" nella tabella *orphans*. A questo punto è stato possibile, partendo da ciascun "orfano", integrare i suoi alberi ascendenti con i dati di Roglo, ove presenti, mediante l'applicazione "integration.cs" (Algoritmo 7).

Algoritmo 7: integration.cs

1. Per ogni albero, crea l'hashtable con i dati di Roglo
2. Per ogni albero seleziona da *orphans* tutti i soggetti marcati "updatable"
3. Per ogni soggetto ricava il Sosa del padre o della madre in base a PARENT
4. Se il soggetto ricavato è presente nella hashtable, estrai le date dal nome e aggiorna la tabella *persons* (NAME, BIRTHYEAR, DEATHYEAR)
5. Esegui in modo ricorsivo l'estrazione del padre e della madre del soggetto ricavato
6. Per i soggetti presenti nella hashtable, estrai le date dal nome e aggiorna la tabella *persons* (NAME, BIRTHYEAR, DEATHYEAR)

In questo modo è stato possibile assegnare un nome ad oltre 30 mila soggetti.

Si è fatto uso, qui come in altri algoritmi, della ricorsione per poter risalire gli alberi degli ascendenti, ottenendo ogni volta i dati di ciascuno dei due genitori. Poiché i valori restituiti sono due, anziché utilizzare una funzione si è usato un

metodo che, dopo aver eseguito l'estrazione dei dati, richiamasse se stesso due volte, una con i dati del padre e una con quelli della madre. In questo modo, partendo da un soggetto qualsiasi, è possibile risalire il suo albero degli antenati fintanto che la base dati dispone di informazioni. La guardia di uscita dalla ricorsione, quindi, è in misura del massimo numero Sosa disponibile nella base dati di riferimento.

Dopo aver integrato i dati con Roglo, si è provveduto ad assegnare un nome ai soggetti anonimi, basandosi sul coniuge se presente. La forma prescelta è "NN cg Nome_coniuge" per gli uomini e "NeN cg Nome_coniuge" per le donne. Questo permette di distinguere i maschi dalle femmine anche in assenza di Sosa.

I nomi generati con questo procedimento, attraverso l'applicazione "spousename.cs" (Algoritmo 8) sono stati salvati nella tabella *coppieanonime* prima di procedere all'aggiornamento di *persons* mediante istruzioni T-SQL.

Algoritmo 8: spousesname.cs

1. Seleziona da *persons* tutti i soggetti con NAME nullo
2. Per ciascun soggetto individua il Sosa del coniuge: (SOSA – 1) se IsOdd(SOSA), (SOSA + 1) altrimenti
3. Se in *persons* il coniuge ha NAME non nullo, ricava NAME, BIRTHYEAR e DEATHYEAR del coniuge
4. Componi il nome con "NeN" se IsOdd(SOSA), "NN" altrimenti + " cg " + NAME
5. Se BIRTHYEAR non è nullo, assegnalo al soggetto come BIRTHYEAR tra

parentesi

6. Se DEATHYEAR non è nullo, assegnalo al soggetto come DEATHYEAR tra parentesi
7. Salva il record nella tabella *coppieanonime*

Con questo ultimo algoritmo sono stati assegnati i nomi a tutti i soggetti identificabili in qualche modo; tuttavia ne rimangono ancora molti, oltre 520000, completamente anonimi. Si è quindi ritenuto opportuno assegnare loro un nome che fosse riferibile al discendente più prossimo. La forma scelta è stata la concatenazione di “NN” o “NeN”, come visto in precedenza, del simbolo “#” (che contraddistingue quindi i soggetti nominati in questo modo), del Sosa che il soggetto assume nei confronti del discendente più prossimo (es. 4 se è il nonno paterno) espresso su 4 cifre, del simbolo “->” e del nome del discendente più prossimo.

Tale nomenclatura non permette sempre di individuare se due soggetti sono in realtà la stessa persona ma, in mancanza di altre informazioni, è stato l'unico modo possibile per completare la base dati. L'errore introdotto da eventuali soggetti ripetuti ma non individuati come tali dovrebbe essere comunque trascurabile.

Mediante l'applicazione “filler.cs” (Algoritmo 9) si è proceduto all'assegnazione dei nomi completando la base dati.

Algoritmo 9: filler.cs

1. Seleziona da *persons* tutti i soggetti con NAME nullo
2. Per ciascun soggetto individua il Sosa del figlio
3. Se il nome contiene "#", ricava il nome originario e il Sosa relativo precedente
4. Componi il nome con "NeN" se IsOdd(SOSA), "NN" altrimenti + " # " + Sosa relativo ricalcolato + "->" + NAME
5. Aggiorna la tabella *persons*

In fase di elaborazione dell'algoritmo, la presenza nel nome del segno "#" indica che tale nome è già stato oggetto di assegnazione pertanto il Sosa va calcolato non in relazione al soggetto corrente ma a quello originario in modo da avere una catena continua di Sosa per tutti gli antenati anonimi dello stesso soggetto.

Per quanto riguarda le analisi relative a MRCA e MCRA, è stato necessario estrarre dalla base dati ulteriori tabelle relative alle generazioni successive alla prima, sia in senso genealogico (tutti i padri e le madri, tutti i nonni e le nonne) sia in senso anagrafico (tutti i nati tra un anno ed un altro). È stato necessario quindi estrarre i nuovi campioni e generare i loro alberi degli antenati recuperando i dati non più dalle banche dati online ma dal database appena creato.

La procedura per la creazione delle tabelle delle generazioni genealogiche successive alla prima (tabelle *persons2*, *persons3*, ...) è di seguito descritta.

Si è proceduto a creare la nuova tabella importando i soggetti di 2ª generazione usando la seguente istruzione T-SQL

```
SELECT NAME, BIRTH, DEATH, BIRTHYEAR, DEATHYEAR, FAMILY, SOSA,
GENERATION = 1, OLDFAMILY = FAMILY, OLDSOSA = SOSA
INTO persons[x]
FROM persons[x-1]
WHERE GENERATION = 2
```

dove [x] è il numero della generazione corrente. Estraendo i dati sempre dalla generazione (x – 1) è stato possibile eseguire sempre le stesse istruzioni ovvero partire sempre dalla 2ª generazione.

Dai soggetti estratti sono stati poi rimossi quelli considerati spurii, ovvero non facenti parte del gruppo di famiglie preso in considerazione. L'analisi è stata condotta sui cognomi individuando una serie di soggetti marginali da non considerare parte del campione. Di seguito si è provveduto ad eliminare i soggetti presenti più volte, individuati attraverso la ripetizione dell'istruzione T-SQL

```
SELECT MAX(OLDFAMILY), OLDSOSA, p.NAME FROM persons[x] p
INNER JOIN
(SELECT pp.NAME FROM persons[x] pp
GROUP BY pp.NAME
HAVING COUNT(pp.NAME) > 1)  d
ON p.NAME = d. NAME
GROUP BY  OLDSOSA, p.NAME
```

Si è poi assegnato a ciascun soggetto rimasto un numero progressivo nel campo FAMILY, indicante il nuovo albero di appartenenza. In questo modo si è creato il campione alla xª generazione per cui estrarre gli alberi degli antenati. Questa operazione è stata eseguita mediante l'applicazione “shifter.cs” (Algoritmo 10)

che, facendo uso della ricorsione, ricava dalla generazione precedente i nomi e i dati degli antenati di ciascun soggetto.

Algoritmo 10: shifter.cs

1. Per ciascun soggetto nella tabella *persons*[*x*] ricava i dati dalla tabella *persons*[*x* - 1]
2. Esegui in modo ricorsivo l'estrazione del padre e della madre del soggetto ricavato e collocali in una tabella temporanea
3. Ordina la tabella temporanea per Sosa e riassegna SOSA progressivamente
4. Salva i dati in *persons*[*x*]

La ricorsione consente di estrarre l'intero albero partendo dal soggetto iniziale. Risulta però necessario riassegnare i Sosa in base al nuovo ruolo assunto dal soggetto. Ordinando i Sosa in modo crescente è possibile sostituirli con un una progressione di interi da 2 a $2^n - 1$, con n = numero di generazioni.

L'inserimento dei dati nella tabella, in questo come in molti altri casi, è stato effettuato in maniera massiva da file CSV, dopo aver assegnato SOSA = 1 a tutti i soggetti originari in tabella.

Un procedimento analogo è stato utilizzato per generare le tabelle con le genealogie anagrafiche. Le tabelle *generation1*, *generation2*, … sono state create con l'istruzione T-SQL

```
SELECT * INTO generation[x] FROM persons WHERE
IntYear(BIRTHYEAR) >= 1880
AND IntYear(BIRTHYEAR) < 1910
```

dove [*x*] è il numero della generazione corrente e gli anni sono di volta in volta modificati in base alla fascia di interesse. I soggetti per fascia di età sono stati estratti partendo sempre dalla tabella *persons*. La funzione "IntYear" è definita all'interno del database come

```
FUNCTION [IntYear]
(
    @Y VARCHAR(6)
)
RETURNS int
AS
BEGIN
    DECLARE @YEAR int
    SET @YEAR = CAST(
       replace(
              replace(@Y, '(', ''), ')', '') AS int)
    RETURN @YEAR;
END
```

ed esegue la rimozione delle parentesi, aggiunte alle date quando derivano dal coniuge. In questo modo il dato è trasformato in valore numerico e può essere usato per filtrare le fasce di anno di nascita.

Eliminati spurii e doppioni con le solite metodologie viste in precedenza, vengono assegnati i valori di FAMILY e SOSA alle colonne OLDFAMILY e OLDSOSA mentre a FAMILY viene assegnato un numero progressivo come nuovo identificativo dell'albero e a SOSA e GENERATION il valore 1.

L'applicazione "binfiller.cs" (Algoritmo 11) infine estrae i dati da *persons* generando un file CSV da utilizzarsi per il popolamento della tabella.

Algoritmo 11: binfiller.cs

1. Per ciascun soggetto nella tabella *generation*[*x*] ricava i dati dalla tabella *persons*
2. Esegui in modo ricorsivo l'estrazione del padre e della madre del soggetto ricavato e collocali in una tabella temporanea
3. Ordina la tabella temporanea per Sosa e riassegna SOSA progressivamente e GENERATION di conseguenza
4. Salva i dati in *generation*[*x*]

Algoritmo 10 e Algoritmo 11 procedono sostanzialmente in modo simile pur attingendo i dati da fonti diverse. Il procedimento ricorsivo e la riassegnazione dei Sosa fanno sì che, al termine dell'elaborazione, si disponga di 6 popolazioni divise per generazione genealogica e 6 divise per anno di nascita. Vedremo in seguito come estrarre da esse gli antenati comuni.

Un'ulteriore procedura di estrazione riguarda la generazione delle matrici di frequenza degli antenati. Dalla tabella *persons* sono state estratte 48 tabelle (*family1*, *family2*, ...), una per ciascun albero, contenenti tutti i soggetti appartenenti a quello specifico albero.

Attraverso l'istruzione PIVOT di T-SQL sono state generate 48 matrici in formato CSV con gli individui sulle righe e 16 colonne, una per ciascuna generazione.

L'istruzione T-SQL per la generazione delle matrici è la seguente

```
SELECT name,
ISNULL([1],0) AS [1], ISNULL([2],0) AS [2], ISNULL([3],0) AS [3],
ISNULL([4],0) AS [4], ISNULL([5],0) AS [5], ISNULL([6],0) AS [6],
ISNULL([7],0) AS [7], ISNULL([8],0) AS [8], ISNULL([9],0) AS [9],
ISNULL([10],0) AS [10], ISNULL([11],0) AS [11], ISNULL([12],0) AS [12],
ISNULL([13],0) AS [13], ISNULL([14],0) AS [14], ISNULL([15],0) AS [15],
ISNULL([16],0) AS [16]
FROM (SELECT name + ' ' + BIRTHYEAR + '-' + DEATHYEAR AS name,
     generation,
         COUNT(1) totale
      FROM family[x]
      GROUP BY name,
         generation, BIRTHYEAR, DEATHYEAR
              ) AS dati
PIVOT( SUM(totale)  FOR generation IN ([1], [2], [3], [4], [5], [6], [7],
[8], [9], [10], [11], [12], [13], [14], [15], [16] )) AS QPivot
ORDER BY name
```

dove [x] è il numero dell'albero corrente.

Analogamente sono state estratte anche la matrice con la frequenza di tutti i soggetti indipendentemente dall'albero di appartenenza e quella con la frequenza di appartenenza a ciascun albero indipendentemente dalla generazione.

L'analisi delle matrici è stata condotta mediante spreadsheet adoperando le formule sulla distribuzione e i grafici forniti dal software.

Per quanto riguarda l'analisi degli antenati comuni e del MRCA, si è provveduto ad individuare, per ciascun campione, gli individui presenti in tutti gli alberi mediante l'istruzione T-SQL

```
SELECT individuals.NAME, individuals.BIRTHYEAR, individuals.DEATHYEAR,
MIN(GENERATION) AS MINGEN,
SUM(GENERATION) / CAST(COUNT(p.NAME) AS DECIMAL) AS MEDGEN,
COUNT(DISTINCT FAMILY) AS NFAM
FROM individuals
INNER JOIN [table] p ON individuals.NAME = p.NAME
AND individuals.BIRTHYEAR = p.BIRTHYEAR
AND individuals.DEATHYEAR = p.DEATHYEAR
GROUP BY individuals.NAME, individuals.BIRTHYEAR, individuals.DEATHYEAR
HAVING COUNT(DISTINCT FAMILY) = [N]
ORDER BY individuals.BIRTHYEAR DESC, MIN(GENERATION)
```

dove [table] è il nome della tabella contenente i dati e [N] il numero di alberi presente nella tabella.

Ordinando successivamente i risultati per BIRTHYEAR, MINGEN o MEDGEN è possibile individuare il MRCA secondo le tre metriche usate nel presente lavoro.

Per identificare gli antenati quasi comuni, ovvero quelli presenti in tutti meno uno o in tutti meno *n* alberi, è sufficiente sostituire il valore di [N] con il numero minimo di alberi accettato e cambiare il confronto da uguale a maggiore o uguale.

Capitolo 4:

MRCA, la ricerca dell'antenato comune

MRCA (Most Recent Common Ancestor) ovvero il più recente antenato comune di un gruppo di organismi è l'individuo più recente dal quale tutti gli organismi del gruppo sono discendenti.

Lo studio del MRCA trova la sua applicazione nel campo sia della genetica che della genealogia. Gli studi svolti fino ad adesso sono prevalentemente basati su modelli matematici tesi a fare una stima di quando potrebbe essere vissuto l'antenato comune di gruppi particolarmente vasti se non dell'intero genere umano. La datazione dell'antenato comune a tutti gli esseri viventi ha impegnato a lungo i ricercatori: attraverso modelli abbastanza sofisticati da tener conto non solo della diffusione geografica della specie umana ma anche delle dinamiche migratorie, si presume che tale antenato sia vissuto tra il terzo e il primo millennio a. C. Si tratta, comunque di modelli statistici la cui attendibilità potrebbe essere valutata con certezza solo con test generici applicati all'intera popolazione mondiale.

In ambito genetico si deve tenere in considerazione che i meccanismi di riproduzione sessuata portano ad una rapida diluizione del patrimonio genetico di un singolo antenato. Fanno eccezione il cromosoma Y, che si trasmette in via esclusivamente maschile, o il DNA mitocondriale, trasmesso per via puramente femminile, che rimangono più o meno inalterati in tutti i discendenti di un

antenato. Su questa base molti studi hanno cercato di individuare, sempre con l'utilizzo di modelli matematici, l'Adamo Y cromosomico e l'Eva mitocondriale ovvero gli individui di cui tutti gli esseri attualmente viventi portano il cromosoma Y o il DNA mitocondriale.

Adamo Y cromosomico avrà sicuramente avuto un certo numero di contemporanei dei cui discendenti tuttavia dopo un certo periodo si saranno riprodotte solo le femmine mentre altri possono avere ancora discendenti vivi ma non sono antenati di tutti gli esseri attualmente viventi. Un discorso analogo può essere fatto per Eva mitocondriale.

Lo studio di entrambi i casi prevede un modello, definito modello Wright-Fisher, nel quale ciascun individuo ha un solo genitore in quanto l'apporto dell'altro, nella trasmissione del gene neutro, è irrilevante.

Estensioni del modello ad una popolazione biparentale sono state proposte da Chang [Chang 1999] evidenziando che, data una popolazione di dimensione n, se n è sufficientemente grande, il numero di generazioni che separano la generazione corrente da quella del MRCA tende a $\log_2(n)$ - dove $\log_2$ è il logaritmo base 2 - mentre la generazione in cui si verifica l'IAP (Identical Ancestor Point) è distante $1{,}77 \log_2(n)$.

Per IAP si intende il momento temporale in cui tutti coloro che erano a quel tempo viventi possono essere divisi in due soli gruppi: quelli che ad oggi non hanno alcun discendente e quelli che sono antenati comuni di tutti i viventi attuali.

Quelli analizzati da Chang sono comunque sempre modelli matematici che prevedono vincoli strutturali (generazioni non sovrapposte, random mating, ecc.) e per i quali fino ad adesso non è stato possibile offrire una verifica sperimentale.

Un ulteriore passo avanti è stato compiuto dal lavoro di Rohde, Olson e Chang [Rohde et al. 2004] che prevede un modello in grado di cogliere in modo più realistico le dinamiche storiche della popolazione attraverso una simulazione che fa uso del metodo Monte Carlo. In questo modo, gli studiosi evidenziano il ruolo fondamentale della sovrapposizione delle generazioni nell'individuazione del MRCA e soprattutto nell'accorciamento delle stime del Time to MRCA (la distanza tra la popolazione attuale e l'antenato comune a tutti). Anche in questo caso, comunque, per quanto si tratti di un modello altamente sofisticato, manca una verifica sperimentale su un campione, anche ristretto, ma reale.

L'analisi condotta nel presente lavoro mira ad individuare non solo il MRCA del campione preso in considerazione ma anche a generalizzare eventuali proprietà di questo fenomeno come ad esempio la regolarità tra generazioni o la distanza temporale tra il campione e il relativo MRCA.

Dalle singole tabelle del database suddivise per generazione, è stato possibile estrarre gli antenati comuni, ovvero quelli presenti almeno una volta in tutti gli alberi genealogici dei soggetti di 1ª generazione, e tra questi individuare il più recente.

Tale analisi è stata condotta considerando tre metriche: l'antenato

anagraficamente più recente (quello nato dopo tutti gli altri), l'antenato che compare nella generazione più recente e quello che ha la media delle generazioni in cui è presente minore rispetto a quella degli altri.

La prima analisi è stata condotta considerando le generazioni in termini genealogici: la prima generazione è composta dal campione iniziale, la seconda dai genitori del campione, la terza dai nonni, ecc., fatta salva la rimozione degli spurii come precedentemente illustrato.

Dall'analisi condotta sulle generazioni genealogiche sono emersi i dati in Tabella 1.

	I	**II**	**III**	**IV**	**V**	**VI**
Antenati comuni	570	332	43	3	1	[2]
% su popolazione	51,8%	41,3%	8,3%	1,5%	0,5%	-
Anno nascita MRCA	1612	1587	1521	1485	1485	[1441]
Distanza in anni	-	25	66	36	0	[44]
Generazione minima	9	8	9	9	8	[7]
Generazione media	11,1	11,2	12	12	11,1	[10,4]

Tabella 1: Metriche per MRCA in generazioni genealogiche

La prima doverosa precisazione riguarda la generazione VI per la quale non è stato individuato alcun antenato comune a tutti gli alberi: l'antenato più comune infatti è presente solamente in 289 su 294 alberi. I dati presentati tra parentesi quadre sono quindi riferiti a questo soggetto (o meglio a questa coppia, come vedremo in seguito).

Il fenomeno può essere imputato al limitato numero di generazioni disponibili.

Per la generazione VI si dispone infatti di alberi estesi solo per 11 generazioni e, sebbene gli antenati citati compaiano alla 7ª, la media attestata sul valore 10 fa pensare che per la totale coalescenza degli alberi sia necessaria la presenza di antenati presenti oltre la 11ª generazione. Solamente estendendo gli alberi di qualche altra generazione sarebbe possibile individuare l'antenato comune. Tale operazione comunque introdurrebbe una quantità notevole di rumore nei dati a causa della scarsa disponibilità di informazione nei database online: la maggior parte dei soggetti infatti sarebbe composta da anonimi.

Alla generazione IV e alla generazione V si verifica un fenomeno interessante, caratterizzato dalla ripetizione del MRCA. Questo denota che per un certo periodo, il gruppo delle famiglie è rimasto separato in due blocchi ed è necessario risalire di una ulteriore generazione per individuare il MRCA. Anche la distanza in anni tra i MRCA delle varie generazioni indica un andamento non regolare, con una generazione doppia delle altre. Questo potrebbe essere determinato da una forte sovrapposizione delle generazioni genealogiche rispetto a quelle anagrafiche.

Si può supporre che questo fenomeno, oltre a motivazioni storiche e contingenti, sia dovuto anche alla relativa rigidità del modello usato per l'analisi, basato su generazioni genealogiche e pertanto troppo trasversali rispetto agli anni. Come avremo modo di vedere, l'analisi condotta considerando le genealogie anagrafiche presenta una maggiore regolarità di risultati.

Il numero degli antenati comuni individuati per ciascuna generazione cala con il procedere delle generazioni e questo è dovuto al sempre minor numero assoluto di soggetti presi in considerazione per ciascuna suddivisione in generazioni. Per tale motivo, il valore degli antenati comuni (considerando le loro ripetizioni) rispetto all'intero albero degli antenati è riportato anche in percentuale.

La generazione minima e la generazione media appaiono piuttosto costanti, tenendo in considerazione quanto precedentemente detto per la generazione VI.

Da un punto di vista storico, trattandosi di un campione reale, i MRCA possono essere individuati in persone realmente esistite.

Per la generazione I, Joachim Ernst I, Graf zu Oettingen-Oettingen (1612 – 1659) è il più giovane antenato e, assieme ad altri 14 antenati comuni, compare per la prima volta alla 9ª generazione. Si tratta di un antenato singolo e non di una coppia perché ha avuto due mogli che quindi saranno in due linee discendenti diverse. Rispetto alla media delle generazioni, MRCA della generazione I sono Wilhelm Ludwig, Graf von Nassau-Saarbrücken (1590 – 1640) e sua moglie Anna Amalia, Markgräfin von Baden-Durlach (1595 – 1651).

Alla generazione II, Magdalene Sibylle, Herzogin von Preussen (1587 – 1659) è la più giovane degli antenati comuni e, assieme al marito Johann Georg I, Kurfürst von Sachsen (1585 – 1656), ha anche la media delle generazioni più bassa.

Karl I, Pfalzgraf von Birkenfeld (1560 – 1600) e la moglie Dorothea, Herzogin von Braunschweig-Lüneburg (1570 – 1649) sono l'unica coppia comune presente fin dalla 8ª generazione.

Friedrich Magnus, Graf zu Solms-Laubach (1521 – 1561) e la moglie Agnes, Gräfin zu Wied (1520 – 1588) sono i MRCA della generazione III sia per età che per media delle generazioni e compaiono a partire dalla 9ª generazione assieme ad altre 4 persone.

Alla generazione IV, Anna, Herzogin von Mecklenburg-Schwerin (1485 - 1525) è MRCA secondo le tre metriche e compare per la prima volta alla 9ª generazione assieme ad altre due persone, i suoi genitori; ella è anche l'unico antenato comune individuato alla generazione V. Avendo avuto due consorti, non compare mai in coppia.

Per la generazione VI, come già detto, non è stato possibile individuare antenati comuni a tutti i soggetti di partenza. I più comuni, presenti in 289 su 294 alberi, sono Magnus II, Herzog von Mecklenburg-Schwerin und Güstrow (1441 - 1503) e Sophie von Pommern-Wolgast (1460ca - 1504), genitori della già citata Anna.

L'analisi è stata condotta, inoltre, considerando i soggetti raggruppati per generazioni anagrafiche, come precedentemente descritto, ottenendo i dati in Tabella 2.

	I	II	III	IV	V	VI
Antenati comuni	537	338	115	10	2	[1]
% su popolazione	49,6%	42,2%	23,2%	2,8%	0,5%	-
Anno nascita MRCA	1612	1587	1554	1521	1506	[1485]
tMRCA	283	278	281	284	269	[260]
Generazione minima	9	8	8	8	7	[8]
Generazione media	11	11,1	10,9	10,9	10	[9,9]

Tabella 2: Metriche per MRCA in generazioni anagrafiche

La suddivisione in generazioni di 30 anni, riportata in Tabella 3, ha permesso di avere gruppi più omogenei, senza sovrapposizioni di generazione.

I	II	III	IV	V	VI
1880 - 1909	1850 - 1879	1820 - 1849	1790 -1819	1760 - 1789	1730 - 1759

Tabella 3: Suddivisione in generazioni per anni di nascita

Anche in questo caso si assiste al calo progressivo degli antenati comuni nel corso delle generazioni, dovuto alla riduzione progressiva del campione, anche se in modo più regolare. Ed anche in questo caso, alla generazione VI si ha un unico antenato considerabile più comune ma presente solamente in 73 su 94 alberi degli antenati.

In questa analisi si è introdotto il concetto di Time to MRCA (tMRCA), la distanza in anni tra il punto medio del campione e l'anno di nascita del MRCA. Questo appare sorprendentemente costante, con una oscillazione ampiamente entro i limiti della generazione. Questo risultato sta ad indicare che,

indipendentemente dalla generazione presa in considerazione, all'interno di questo gruppo chiuso, è possibile individuare un antenato comune che dista da tutti i discendenti circa 280 anni ovvero poco più di 9 generazioni anagrafiche, considerando 30 anni per generazione.

Generazione minima e generazione media, intese anche in questo contesto in senso genealogico, risultano piuttosto costanti.

Gli antenati comuni per la generazione I coincidono con quelli individuati dall'analisi delle generazioni genealogiche.

Alla generazione II, Magdalene Sibylle, Herzogin von Preussen (1587 - 1659) è ancora l'antenato comune più giovane assieme al marito Johann Georg I, Kurfürst von Sachsen (1585 - 1656) con il quale ha anche la minor media delle generazioni, comparendo per la prima volta alla 9ª.

Friedrich I, Graf von Salm, Wild- und Rheingraf in Neufville (1547 - 1608) compare invece per la prima volta alla 8ª pur avendo una media più alta.

Margarethe von Schönburg-Glauchau (1554 - 1606) è il MRCA per anno di nascita della generazione III. Pur avendo avuto due mariti, ha avuto figli (ben 16) solamente dal secondo, Johann Georg I, Graf zu Solms-Laubach (1547 - 1600). Entrambi presentano la minor media delle generazioni.

Bernhard VIII, Graf und Edler Herr zur Lippe (1527 - 1563) e sua moglie Katharina, Gräfin zu Waldeck-Eisenberg (1524 - 1583) sono coloro che compaiono nella generazione inferiore ovvero l'8ª.

Alla generazione IV, Friedrich Magnus, Graf zu Solms-Laubach (1521 – 1561)

e sua moglie Agnes, Gräfin zu Wied (1520 – 1588) sono i MRCA sotto tutte le metriche così come alla generazione V lo sono Juliana, Gräfin zu Stolberg-Wernigerode (1506 - 1580) assieme ad uno dei suoi due mariti, Wilhelm 'the Rich', Graf von Nassau-Dillenburg (1487 - 1559).

Infine, alla generazione VI, solo escludendo un numero significativo di alberi troviamo un unico MRCA nella persona di Anna, Herzogin von Mecklenburg-Schwerin (1485 – 1525), già individuata come MRCA alla generazione IV e V dell'analisi condotta sulle generazioni genealogiche.

Confrontando i nomi dei soggetti evidenziati dalle due analisi, si nota ancora una volta la deformazione della generazione genealogica, con l'allungamento della distanza tra generazione II e generazione III come se fosse venuto a mancare un MRCA intermedio. Probabilmente sarà esistito un quasi-MRCA, un soggetto presente in tutti meno alcuni degli alberi, indicante ancora una volta una parziale e temporanea disgregazione del gruppo. La coppia quasi-MRCA potrebbe essere proprio quella composta da Johann Georg I, Graf zu Solms-Laubach e Margarethe von Schönburg-Glauchau, MRCA alla generazione anagrafica III ma presente solo in 138 dei 140 alberi nella generazione genealogica. Questo ha fatto sì che si creasse una distanza maggiore tra i MRCA delle generazioni genealogiche e un disallineamento tra i soggetti MRCA genealogici e quelli cronologici. Le ragioni di questo fenomeno sono da ricercarsi in motivazioni storiche o in variazioni delle politiche matrimoniali che hanno fatto sì che il gruppo rimanesse

temporaneamente diviso in due cluster distinti.

Uno dei risultati più significativi dell'analisi del MRCA riguarda la linearità di spostamento nel tempo, soprattutto se analizzata dal punto di vista delle genealogie anagrafiche che sono quelle più aderenti alla realtà e non soggette a sovrapposizioni.

Abbiamo già evidenziato come la distanza tra il campione prescelto e il MRCA sia più o meno costante assumendo un valore, definito Time to MRCA (tMRCA) intorno a 280 anni ovvero 9,3 generazioni.

È interessante osservare che, sebbene il campione sia di piccole dimensioni, può essere considerata valida la legge individuata da Chang [Chang 1999] secondo la quale, in un modello biparentale, la probabilità che il tMRCA, espresso in generazioni, sia dato dal logaritmo in base 2 del numero di individui N tende a 1 quando N tende ad infinito. Nel caso quindi di un campione così piccolo, circa un centinaio di persone per generazione, è ovvio che la probabilità che tMRCA sia uguale a $\log_2(N)$ sia bassa ed infatti i valori del tMRCA sono superiori a quelli predetti dal modello.

Vi è tuttavia una importante considerazione da fare sul confronto di questi valori: il modello di Chang prevede il random mating pertanto non tiene conto di alcuni vincoli culturali molto forti nelle società occidentali come il tabù di matrimoni tra fratello e sorella o tra cugino e cugina. Questi ultimi, per quanto presenti, sono in realtà abbastanza rari. Il divieto di matrimonio tra parenti così prossimi, oltre che dovuto a proibizione religiosa, veniva anche rispettato per

motivi politici: un matrimonio interno alla famiglia era un'occasione persa per creare o rafforzare un'alleanza tra famiglie. Inoltre, sebbene non si fosse ancora a conoscenza del fatto, i matrimoni tra consanguinei così prossimi erano motivo di trasmissione di malattie genetiche che portavano in breve all'estinzione del ramo. Alla luce di queste considerazioni, se aggiungiamo due generazioni al valore teorico di Chang, il risultato sperimentale non si discosta in modo significativo dal modello.

Inoltre l'oscillazione, sia pur minima, dei valori sperimentali del tMRCA, dovuta alla variazione di dimensione del campione per ciascuna generazione, segue in modo preciso l'andamento di quelli teorici come appare evidente in Grafico 1.

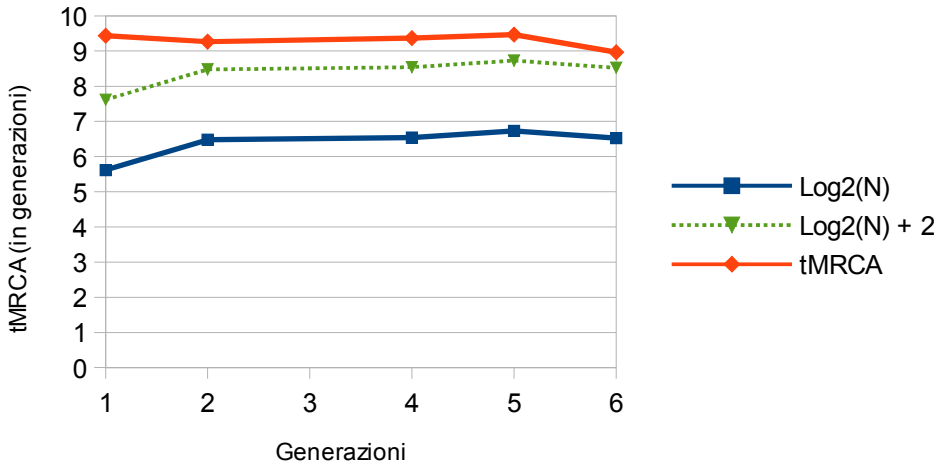

Grafico 1: Confronto tra valore teorico secondo Chang, valore compensato e valore sperimentale

Il dato di generazione 1 non risulta allineato in quanto il campione di tale generazione è sottostimato non tenendo conto di eventuali fratelli o sorelle appartenenti alla stessa generazione ma non presi in considerazione nella selezione iniziale (mentre potrebbero esserlo quelli degli individui delle generazioni successive). Il valore teorico quindi dovrebbe essere calcolato su un numero più alto di individui e quindi, con molta probabilità, dovrebbe mantenere un andamento conforme a quello dei valori sperimentali. Appare comunque evidente una correlazione tra le dimensioni del campione e il tMRCA in generazioni.

Si può quindi concludere che uno spostamento orizzontale, inteso come allargamento o restringimento della popolazione iniziale, provoca un allontanamento o un avvicinamento ad essa del MRCA mentre uno spostamento verticale, inteso come scorrimento di generazioni, mantiene comunque la medesima distanza tra popolazione presa in considerazione come iniziale e MRCA. Crediamo che la generalizzazione di questa proprietà, quasi geometrica, del MRCA, per quanto ancora suscettibile di verifiche, possa costituire un interessante filone di ricerca applicabile a tutte quelle discipline che analizzano il fenomeno della coalescenza.

Proseguendo nella verifica sperimentale del modello di Chang, ci aspetteremmo di trovare dopo 1,77 $\log_2(N)$ generazioni, l'Identical Ancestor Point. Nel nostro campione, esso dovrebbe verificarsi intorno alla 16ª generazione.

La base dati non consente l'analisi di tutti gli antenati di tutti i soggetti presenti negli alberi ma è possibile restringere il "villaggio" alle famiglie dell'alta nobiltà tedesca (Hochadel) lasciando fuori le altre, quelle del resto d'Europa e quelle di minore nobiltà.

Tenendo in considerazione questo campione è possibile compiere una stima che è la migliore approssimazione di fatto dell'IAP.

Analizzando la matrice *T(i, a)*, raggruppata dunque per generazione con gli alberi sulle colonne, sono stati estratti tutti gli individui presenti in tutti gli alberi ovvero tutti gli antenati comuni. Identificando quelli delle famiglie Hochadel (422 individui) e raggruppandoli per generazione anagrafica, si può osservare una frequenza che ha un picco proprio nelle generazioni 15ª (89 individui) e 16ª (90 individui). Si può quindi affermare che, per ogni famiglia Hochadel c'è almeno un rappresentante che è antenato di tutti e che tutti i suoi antenati sono antenati di tutti. Questo indica, almeno per il campione ristretto, una situazione di IAP per una di queste due generazioni, in un periodo collocabile intorno al 1450.

Osservando la distribuzione degli antenati comuni emerge anche un altro fenomeno interessante che ha oltretutto una forte valenza storiografica: l'individuazione di un "collo di bottiglia" nelle genealogie delle famiglie nobiliari. Si è infatti osservato che, intorno alla 16ª generazione anagrafica, ciascuna famiglia ha un solo capostipite. L'antenato comune in linea agnatizia

di ciascun albero si trova quindi più o meno nello stesso periodo. Considerando i capostipiti delle famiglie facenti parte della Hochadel e confrontandoli con l'elenco degli antenati comuni delle stesse famiglie, precedentemente estratto, si nota che i soggetti, salvo alcune eccezioni nelle famiglie minori, coincidono. Tutti i capostipiti delle famiglie dell'alta nobiltà tedesca sono quindi antenati di tutti e questo si verifica tra la 15ª e la 16ª generazione anagrafica. Possiamo quindi affermare che "collo di bottiglia" genealogico e IAP coincidono.

Il fenomeno inoltre è interessante anche dal punto di vista genetico in quanto la linea agnatizia segue lo stesso percorso del cromosoma Y [Rossi, 2013].

I dati su capostipiti e antenati comuni sono riassunti in Tabella 4. La prima parte della tabella, Gruppo I, elenca i capostipiti delle famiglie della Hochadel. Essi sono anche antenati di tutti in quanto presenti in tutti gli alberi. E' un dato significativo perché dimostra l'esistenza dell'IAP per questo gruppo e l'importanza che l'alta nobiltà tedesca assume nel panorama nobiliare europeo. È in parte possibile estendere la considerazione ad alcune delle più importanti casate europee, elencate nel Gruppo II, i cui capostipiti, pur non essendo antenati di tutti, sono comunque presenti nella maggior parte degli alberi. Di fatto anche essi fanno parte dell'IAP.

Nel Gruppo III si hanno le famiglie della nobiltà tedesca non facenti parte della Hochadel ma pur sempre importanti e presenti in modo significativo in molti degli alberi. Esse potrebbero collocarsi, con una certa approssimazione, intorno

all'IAP. Infine, nel Gruppo IV sono riportate le famiglie europee poco presenti.

Famiglia	Capostipite	Anno nascita	Gen. Ana.	Gen. Gen.	N. Alberi
Gruppo I: Hochadel					
Waldeck	Wolrad von Waldeck	1400	17	14	48
Schwarzburg	Heinrich XXVI Graf von Schwarzburg	1418	17	14	48
Isenburg	Ludwig II von Isenburg	1422	17	13	48
Oldenburg	Christian I von Oldenburg	1426	17	13	48
Sachsen	Ernst, Kurfürst von Sachsen 1464-1486 1441-1486	1441	16	13	48
Mecklenburg	Magnus II von Mecklenburg	1441	16	13	48
Mansfeld H.	Ernst von Mansfeld	1445	16	14	48
Wittelsbach B.	Albrecht IV 'der Weise', Herzog von Bayern 1469-1508 1447-1508	1447	16	14	48
Wurttemberg	Heinrich von Wurttemberg	1448	16	13	48
Hohenzollern G.	Eitel Friedrich II, Graf von Hohenzollern 1488-1512 1452-1512	1452	16	14	48
Pommern	Bogislaw X von Pommern	1454	16	13	48
Kleve	Johann II der Kindermacher	1458	15	13	48
Hohenzollern	Friedrich V der Alte von Hohenzollern	1460	15	13	48
Salm	Johann VI zu Salm	1460	15	13	48
Wittelsbach	Alexander von Wittelsbach	1462	15	13	48
Reuss G.	Heinrich XIII. Reuß, Herr zu Greiz	1464	15	12	48
Wied	Johann III, Graf von Wied 1465-1533	1465	15	13	48
Stolberg	Bodo III zu Stolberg	1467	15	13	48
Solms L.	Philipp zu Solms-Lich	1468	15	13	48
Sachsen L.	Magnus I, Herzog von Sachsen-Lauenburg 1507-1543 1470-1543	1470	15	13	48
Lippe	Simon V, Graf zur Lippe	1471	15	12	48
Sachsen	Heinrich der Fromme	1473	15	13	48

Famiglia	Capostipite	Anno nascita	Gen. Ana.	Gen. Gen.	N. Alberi
Baden	Christoph I von Baden	1475	15	14	48
Habsburg	Philipp I von Habsburg	1478	15	13	48
Mansfeld	Ernst von Mansfeld	1479	15	13	48
Oettingen	Ludwig XV, Graf zu Oettingen-Oettingen	1486	15	12	48
Schonburg	Ernst von Schoenburg	1486	15	13	48
Nassau D.	Wilhelm VIII der Reiche, Graf von Nassau-Dillenburg	1487	15	12	48
Sayn W.	Wilhelm I, Graf von Sayn-Wittgenstein	1488	15	12	48
Hohenlohe	Georg Graf zu Hohenlohe	1488	15	13	48
Solms B.	Philipp zu Solms-Braunfels	1494	14	12	48
Braunschweig	Ernst I der Bekenner, Herzog von Braunschweig-Lüneburg	1497	14	12	48
Hanau M.	Philipp II von Hanau-Munzenberg	1501	14	13	48
Barby	Wolfgang I, Graf von Barby	1502	14	12	48
Jagiello	Anna Jagiellonka, Königin von Ungarn	1503	14	12	48
Anhalt	Johann II, Fürst von Anhalt-Zerbst	1504	14	12	48
Hessen	Philipp I der Grossmütige, Landgraf von Hessen	1504	14	12	48
Nassau W.	Philipp III, Graf von Nassau-Weilburg	1504	14	12	48
Hohenzollern	Joachim II Hector von Hohenzollern, Kurfürst von Brandenburg	1505	14	12	48
Erbach	Eberhard XIV von Erbach, Graf zu Erbach	1511	14	12	48
Wittelsbach P.	Friedrich III, Kurfürst von der Pfalz	1515	14	12	48
Hanau L.	Philipp V, Graf von Hanau-Lichtenberg	1541	13	12	48
Gruppo II: Casate europee importanti					
Vasa	Gustav I Vasa, konung av Sverige	1496	14	12	47
Albret	Jean d'Albret	1469	15	13	46
Stuart	James I Stuart, King of England	1566	13	12	46

Famiglia	Capostipite	Anno nascita	Gen. Ana.	Gen. Gen.	N. Alberi
Lorraine	Rene II duc de Lorraine	1451	16	14	44
Bourbon	François de Bourbon, Comte de Vendôme, Comte de St.Pol 1470-1495	1470	15	14	44
Gruppo III: Nobiltà tedesca non Hochadel					
Schlesien	Friedrich II von Schlesien-Liegnitz	1480	15	13	47
Sayn	Gerhard II von Sayn	1417	17	14	46
Limpurg S	Gottfried II von Limpurg-Speckfeld	1474	15	14	45
Castell	Georg, Graf von Castell	1527	13	12	45
Bentheim	Everwin III, Graf von Bentheim	1536	13	12	44
Leiningen	Emich X, Graf von Leiningen	1498	14	12	43
Limpurg G	Wilhelm III von Limpurg-Gaildorf	1498	14	13	43
Oldenburg G	Anton I, Graf von Oldenburg	1505	14	12	41
Leiningen W	Kuno II Graf von Leiningen-Westerburg	1487	15	13	39
Ostfriesland	Enno II, Graf von Ostfriesland	1505	14	12	39
Lowenstein	Ludwig III, Graf zu Löwenstein-Wertheim	1530	13	12	35
Gruppo IV: Nobiltà europea poco presente					
Gonzaga	Gianfrancesco II Gonzaga	1466	15	14	22
Savoia	Philippe II Sans Terre	1438	16	15	19
De' Medici	Cosimo I de' Medici	1519	14	13	18
Este	Cesare d'Este, duca di Modena	1552	13	12	18
Valois	Claude d'Angoulême, duchesse de Lorraine	1547	13	12	18
Aviz	Duarte de Aviz, Duque de Guimarães	1515	14	12	17
Braganca	D. Jaime, duque de Braganca	1479	15	13	17
Farnese	Ranuccio I Farnese, duca di Parma	1569	13	12	17
Romanov	Nikolaï Romanov	1529	13	12	15

Tabella 4: Capostipiti delle famiglie della nobiltà europea

Ulteriori analisi di questi dati potranno essere condotte in seguito, portando ad

una verifica più precisa del modello identificato da Chang ma ci sembra che già questa prima osservazione lasci intravedere una conferma della teoria sull'IAP.

Capitolo 5:

MCRA, una misura della ripetizione degli antenati

Se il concetto di MRCA è ormai entrato nella tradizione degli studi genealogici, vogliamo adesso introdurre il MCRA ovvero il Most Common Recent Ancestor. Con questo acronimo vogliamo indicare il più frequente antenato per ciascuna generazione. È un'analisi che ha a che fare con le distribuzioni in frequenza già studiate da Deridda e colleghi [Derrida et al. 1999] ma che pone l'accento su quegli individui che si ripetono più volte all'interno della stessa generazione. Partendo dal principio della ripetizione degli antenati, andremo ad analizzare quale sia il soggetto più ripetuto e con che frequenza. L'MCRA diviene quindi una misura del grado di ripetizione degli antenati, in una comunità chiusa, nel succedersi delle generazioni. Ma è anche una indicazione della dominanza, o per lo meno della significativa presenza, di una certa famiglia in questo contesto.

Anche l'analisi delle frequenze e quindi la ricerca del MCRA, è stata condotta sia considerando le generazioni genealogiche che quelle anagrafiche.

Per le generazioni genealogiche, l'analisi delle frequenze è stata effettuata su una matrice, estratta dalla base dati, di tipo *S(i, g)* con gli individui sulle righe e le generazioni sulle colonne. Il valore di frequenza è stato pesato rispetto agli individui totali della generazione secondo la formula

$$w = \frac{N(i)}{2^{g-1}}$$

dove *N(i)* è il numero di occorrenze di *i* nella generazione *g*.

Escludendo le prime 6 generazioni, dove il numero di ripetizioni degli antenati non è significativo poiché il campione è troppo piccolo perché anche una sola ripetizione non conduca ad una sovrastima del peso, sono stati individuati, per ciascuna generazione, i soggetti più frequenti con il relativo peso assoluto, il peso in percentuale (dividendo per il numero totale degli alberi) e il numero di alberi in cui sono presenti alla generazione indicata. I dati sono riportati in Tabella 5.

Gen.	**MCRA**	**w**	**w%**	**N. alberi**
7	Carlos III, King of Spain e Maria Amalia, Prinzessin von Sachsen	0,953	2,00	14
8	Felipe V, King of Spain e Elisabeth Maria Farnese, Princess of Parma	0,742	1,55	14
9	Louis 'le Grand Dauphin', Dauphin of France e Maria Anna Victoria, Herzogin von Bayern, Pfalzgräfin bei Rhein	0,395	0,82	16
10	Albrecht, Markgraf von Brandenburg-Ansbach e Sophie Margarete, Gräfin zu Oettingen-Oettingen	0,355	0,74	42

Gen.	MCRA	w	w%	N. alberi
11	Georg II 'der Gelehrte', Landgraf von Hessen-Darmstadt e Sophie Eleonore, Herzogin von Sachsen	0,317	0,66	45
12	Johann Georg I, Kurfürst von Sachsen e Magdalene Sibylle, Herzogin von Preussen	0,360	0,75	48
13	Johann Georg, Kurfürst von Brandenburg	0,360	0,75	48
14	Ferdinand I, Emperor of the Holy Roman Empire e Anna Jagiello, Princess of Hungary	0,380	0,79	48
15	Anna, Herzogin von Mecklenburg-Schwerin	0,347	0,72	48
16	Kazimierz IV Jagiello, King of Poland, Grand Duke of Lithuania e Elisabeth of Austria	0,375	0,78	48

Tabella 5: I MCRA delle generazioni genealogiche da 7 a 16

Come si può notare, nella maggior parte dei casi, il MCRA, così come il MRCA, è formato da una coppia di sposi. Negli altri casi siamo in presenza di nozze multiple pertanto solo uno dei coniugi presenta la massima frequenza di ripetizione.

A partire dalla 10ª generazione si osserva un brusco incremento del numero di alberi del quale l'individuo fa parte e già dalla 12ª generazione, il MCRA è anche un antenato comune. Tuttavia si nota che non necessariamente vi è una corrispondenza tra MCRA e MRCA ovvero il MRCA non necessariamente è l'antenato più frequente così come l'antenato più frequente non necessariamente è il più comune. Solamente la coppia Johann Georg I, Kurfürst von Sachsen e

Magdalene Sibylle, Herzogin von Preussen, MCRA della 12ª generazione, e Anna, Herzogin von Mecklenburg-Schwerin MCRA della 15ª generazione sono anche MRCA rispettivamente della 2ª e della 4ª e 5ª generazione. In questo caso, inoltre, la distanza in generazioni tra popolazione di partenza e MRCA rimane immutata (9 generazioni, come evidenziato dall'analisi del MRCA). Nel caso di Anna, Herzogin von Mecklenburg-Schwerin, infatti, la generazione da considerare è la 5ª poiché la sua presenza come MRCA alla 4ª è dovuta, come già evidenziato, alla parziale separazione del gruppo di riferimento per un certo periodo di tempo.

L'evoluzione nel tempo del peso in percentuale del MCRA è riportata nel Grafico 2.

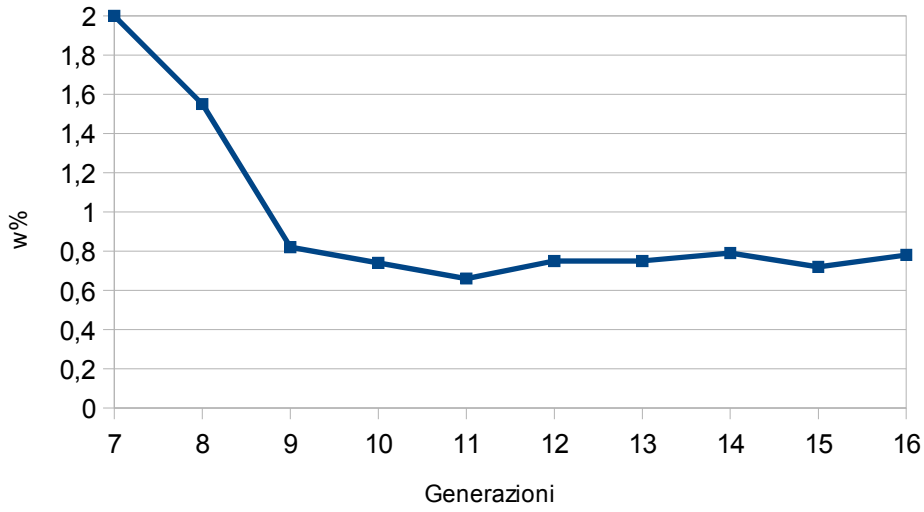

Grafico 2: Peso (*w*) del MCRA per generazione genealogica

Osservando il grafico, si nota che, dopo un fase decrescente, a partire dalla 9ª generazione il valore del peso percentuale del MCRA comincia a stabilizzarsi assumendo valori di poco inferiori a 0,8%. Questo accade in corrispondenza dell'aumento del numero di alberi in cui l'individuo è presente. Si può quindi affermare che si verifica un processo di saturazione delle ripetizioni che è identificabile come una forma di universalità nella distribuzione.

Per quanto riguarda le generazioni anagrafiche, dalla matrice precedentemente usata per le genealogiche si è ricavato, per ciascun individuo

$$\sum_{g=1}^{16} w(i)$$

ovvero la sommatoria dei pesi di ciascun individuo, ottenendo il peso dell'individuo nell'intera popolazione. Si è poi provveduto a raggruppare gli individui forniti di data di nascita in generazioni anagrafiche di 30 anni, partendo dal 1910 (come per l'analisi del MRCA) e risalendo fino al 1250. Anche in questo caso si è scelto di iniziare l'osservazione dalla 7ª generazione in modo da evitare la sovrastima del peso dovuta alle dimensioni del campione. Sebbene sia stato possibile individuare antenati comuni fino alla 21ª generazione, si è deciso di limitare l'osservazione alla 16ª generazione per evitare una sottostima del peso dovuta alla scarsità di dati. Il numero di alberi in cui l'individuo è presente è da intendersi, in questa analisi, indipendente dalla generazione genealogica. I risultati sono riportati in Tabella 6.

Gen.	MCRA	w	w%	N. alberi
7 (1700 -1729)	Carlos III, King of Spain e Maria Amalia, Prinzessin von Sachsen	1,629	3,39	17
8 (1670 -1699)	Felipe V, King of Spain e Elisabeth Maria Farnese, Princess of Parma	1,338	2,79	17
9 (1640 -1669)	Louis 'le Grand Dauphin', Dauphin of France e Maria Anna Victoria, Herzogin von Bayern, Gräfin bei Rhein	0,764	1,59	17
10 (1610 -1639)	Philipp Wilhelm, Kurfürst von der Pfalz e Elisabeth Amalie, Landgräfin von Hessen-Darmstadt	0,783	1,63	21
11 (1580 -1609)	Johann Georg I, Kurfürst von Sachsen e Magdalene Sibylle, Herzogin von Preussen	0,789	1,64	48
12 (1550 -1579)	Margarethe von Schönburg-Glauchau	0,728	1,52	48
13 (1520 -1549)	Joachim Ernst, Fürst von Anhalt-Zerbst und Dessau	1,022	2,13	48
14 (1490 -1519)	Juliana, Gräfin zu Stolberg-Wernigerode	1,041	2,17	48
15 (1460 -1489)	Anna, Herzogin von Mecklenburg-Schwerin	1,034	2,15	48
16 (1430 -1459)	Magnus II, Herzog von Mecklenburg-Schwerin und Güstrow	0,808	1,68	48

Tabella 6: I MCRA delle generazioni anagrafiche da 7 a 16

Anche in questo caso si hanno sia coppie che soggetti singoli ma la motivazione del fenomeno è differente: in alcuni casi infatti i coniugi non appartengono alla stessa generazione pertanto non possono essere entrambi

MCRA né necessariamente il coniuge sarà MCRA della propria generazione.

Ad esempio, nel caso di Margarethe von Schönburg-Glauchau, MCRA della 12ª generazione, nata nel 1554, il coniuge fertile (come già detto ebbe due mariti ma ha avuto discendenza solamente dal secondo), Johann Georg I, Graf zu Solms-Laubach è nato nel 1547 quindi, secondo la suddivisione adottata, in una generazione precedente, all'interno della quale, quanto a frequenza, è solamente sesto.

È interessante inoltre notare che, fino alla 9ª generazione (così come nelle generazioni precedenti non riportate), i MCRA genealogici e quelli cronologici coincidono mentre nelle generazioni successive si ha uno scollamento con solo alcuni punti di contatto, anche in generazioni diverse (ad esempio, la 12ª genealogica con la 11ª anagrafica). Questo sta ad indicare che, se per le prime generazioni la durata della generazione genealogica e di quella anagrafica sono più o meno paragonabili, le generazioni genealogiche più remote tendono ad avere durate meno regolari e presentano maggiore sovrapposizione.

Nel caso delle genealogie anagrafiche, la corrispondenza tra MRCA e MCRA sembra essere più presente. Notiamo infatti che i MCRA della 11ª, 12ª, 14ª e 15ª generazione sono rispettivamente i MRCA della 2ª, 3ª, 5ª e 6ª generazione, dimostrando ancora una volta la distanza costante di 9 generazioni tra campione iniziale e MRCA.

Per quanto riguarda la 4ª generazione, pur non essendovi corrispondenza, i MRCA sono comunque al terzo posto come valore di frequenza.

L'evoluzione nel tempo del peso percentuale del MCRA è riportata nel Grafico 2.

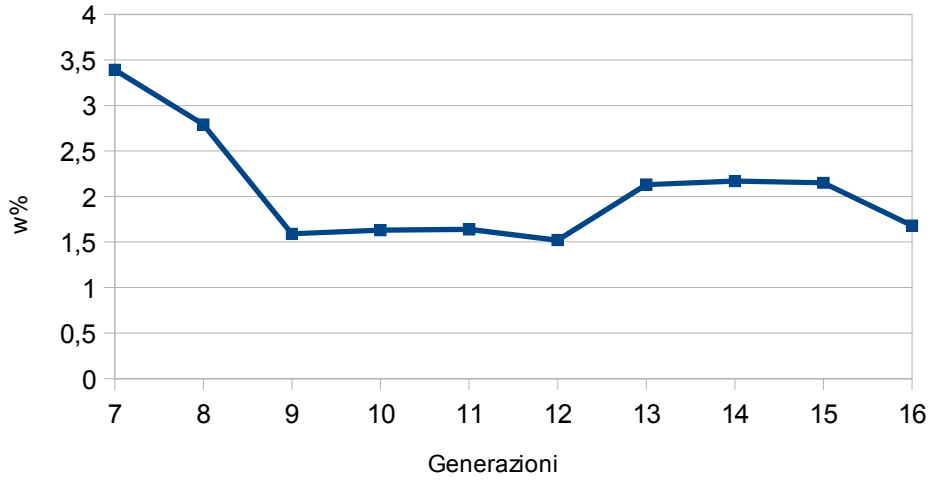

Grafico 2: Peso (*w*) del MCRA per generazione anagrafica

Anche in questo caso, dopo un iniziale andamento decrescente, i valori divengono piuttosto stazionari, con una oscillazione di circa mezzo punto percentuale, indice di una parziale saturazione. Non sembra invece esserci relazione tra l'andamento del grafico e il punto in cui il MCRA diviene antenato comune a tutti gli alberi (11ª generazione).

Un'ulteriore analisi ha riguardato la distribuzione in frequenza degli antenati che compaiono con peso *w* nell'albero *a,* rappresentata da $P(w,a)$.

Essendo le frequenze valori pesati e quindi non discreti, la notevole presenza di rumore data da valori di frequenza assenti ha suggerito l'uso di una distribuzione cumulativa. Essa è definita come la probabilità che un antenato

compaia con peso maggiore o uguale a *w* nell'albero *a*.

Per gli antenati presenti in più di una generazione, *w* è la somma dei pesi parziali che equivale a una media pesata della frequenza delle ripetizioni.

Si è calcolato poi C(*w*) come il $\log_2$ del numero di individui con peso maggiore o uguale a *w* (distribuzione cumulativa).

Riportate su assi cartesiani, le frequenze mostrano l'andamento raffigurato in Grafico 3.

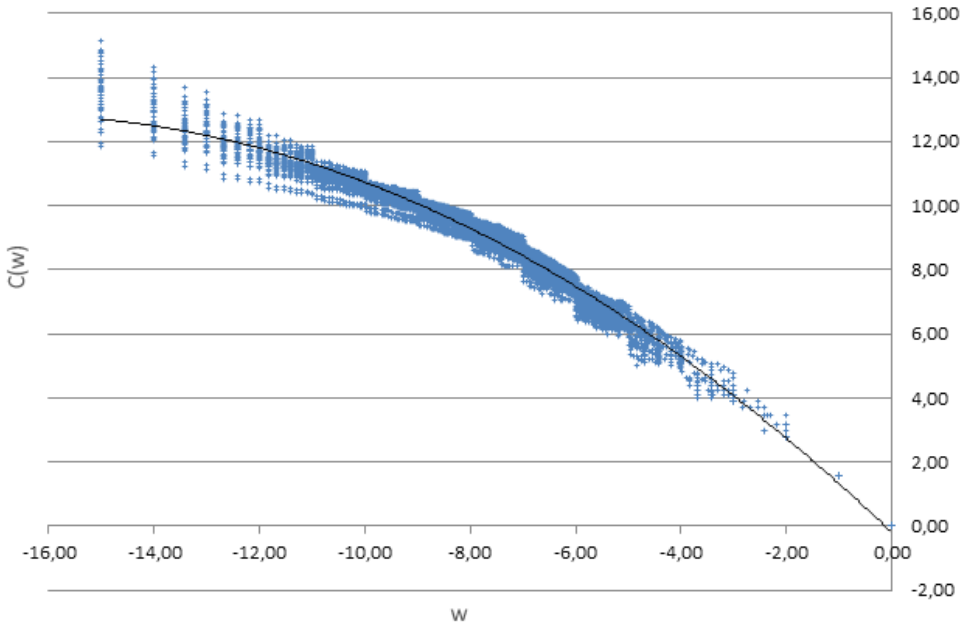

Grafico 3: Distribuzioni cumulative delle frequenze pesate

Facendo poi la media delle distribuzioni tra tutti gli alberi e riportando i valori su assi cartesiani emerge la curva raffigurata in Grafico 4.

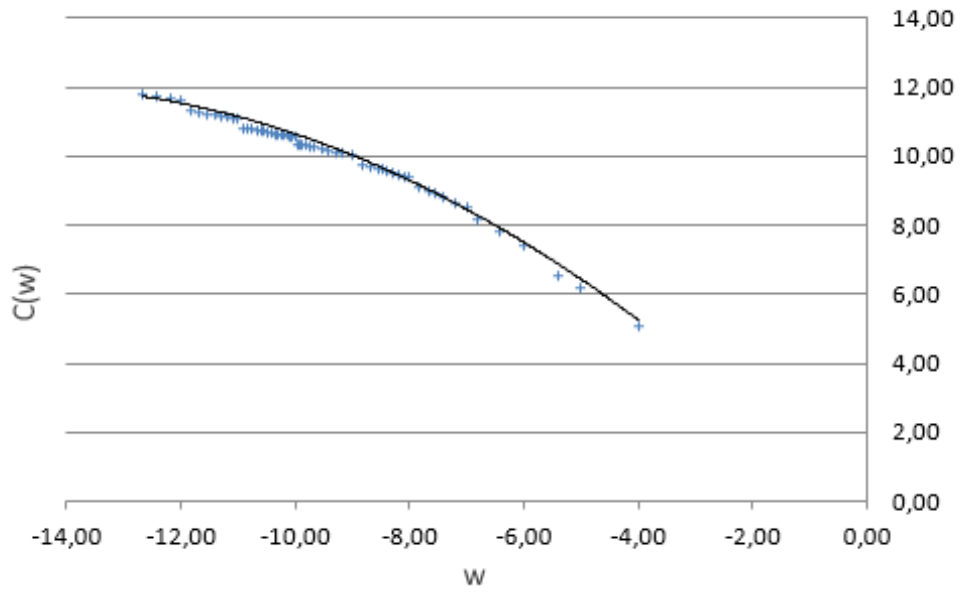

Grafico 4: Media delle distribuzioni cumulative delle frequenze pesate

Malgrado alcune, trascurabili, interruzioni si nota l'andamento di una curva crescente con un massimo (individuabile tramite la formula ricavata dallo spreadsheet) intorno al valore -15. L'andamento continuo delle medie denota quindi una certa universalità e non si discosta sostanzialmente dai risultati individuati da Derrida e colleghi [Derrida et al. 1999 e 2000b].

Nella loro analisi infatti, definiscono *M(r)* il numero di antenati che compaiono *r* volte in un dato albero genealogico, ricavando la funzione di frequenza come $F(r) = M(r) / N_A$ dove N_A è il numero di antenati distinti. Nella loro simulazione numerica, si assume che la popolazione sia fissa ed uguale a *N* per ciascuna generazione e si calcola *F(r)* in funzione di N_A e del numero di generazioni *G*.

Lo studio procede con la misurazione della probabilità delle ripetizioni *H(r, n_g)* ad ogni generazione $n_g \leq G$ facendo riferimento all'intera popolazione *N* alla generazione n_g. Dopo un sufficiente numero di generazioni, la distribuzione della ripetizione degli antenati assume una forma universale, simile a quella di Grafico 4, nella curva $P(w) \equiv 2^{n_g} H(r, n_g)/N$ come funzione di $w \equiv r N / 2^{n_g}$.

La coda di *P(w)*, per valori di *r* piccoli, è una power law con esponente $\beta \simeq 0{,}3$

La distribuzione tende ad un valore stazionario $P\infty(w)$ per un numero molto alto di *G*, oltre quello di cui disponiamo nell'analisi corrente.

Le distribuzioni calcolate sul campione oggetto di studio sembrano seguire questo tipo di comportamento (power law per valori bassi e tendenza ad un valore stazionario), evidente anche dalle curve di andamento di Grafico 3 e Grafico 4. Ulteriori approfondimenti sulla corrispondenza tra le teorie di Derrida e colleghi e i valori sperimentali ottenuti esulano dall'oggetto del presente lavoro ma potranno essere svolti in futuro partendo dalla base dati e dalle matrici da essa estratte come suggerisce Rossi in [Rossi, 2013].

Conclusioni e prospettive di ricerca

Questo lavoro di tesi, che ha avuto per oggetto l'analisi delle genealogie della nobiltà europea dell'età moderna ed in particolare delle proprietà statistiche nella ripetizione degli antenati e nell'individuazione dell'antenato comune più recente, ha condotto ad una serie di risultati che, lungi dall'essere definitivi ed assodati, gettano una luce nuova sullo studio delle genealogie se non altro per la loro natura sperimentale basata su dati concreti.

Per quanto riguarda il MRCA, è stata offerta una prima conferma delle teorie di Chang [Chang 1999] su tMRCA e, in parte, su IAP ma soprattutto è stata messa in evidenza la linearità di spostamento nel tempo del MRCA, in particolare se analizzato dal punto di vista delle genealogie anagrafiche che sono quelle più aderenti alla realtà e non soggette a sovrapposizioni.

Crediamo che quelle che ci piace definire le proprietà “geometriche” dello spostamento del MRCA, che evidenziano la autosimilarità del fenomeno, siano intuizioni che richiederanno in futuro un approfondimento non solo in ambito genealogico ma in tutti quei contesti caratterizzati dalla coalescenza nei processi di branching.

Un ulteriore contributo derivante dall'analisi del MRCA condotta in questo lavoro di tesi potrebbe essere costituito dalla opportunità di utilizzare il tMRCA come indice di coesione di una popolazione. Per quanto il tMRCA sia connesso alle dimensioni del campione, a parità di dimensioni, soprattutto per

popolazioni non troppo grandi, esso assumerà valori diversi per campioni diversi in base a quanto il gruppo risulta coeso ovvero offre meno resistenza alla coalescenza. Supponiamo di analizzare un campione di *N* individui, con *N* molto simile a quello analizzato in questo studio e con caratteristiche simili di gruppo chiuso, ad esempio la popolazione di una piccola valle montana.

In mancanza di dati genealogici potremmo supporre di trovare il MRCA nella 9ª generazione (ovvero dopo circa 280 anni se consideriamo generazioni di 30 anni) ed usare lo studio qui compiuto come un modello da cui estrapolare dati non disponibili.

Ma supponiamo di avere le genealogie degli abitanti della valle e da esse ricavare, con procedimenti analoghi a quelli messi in opera in questo lavoro, il tMRCA scoprendo che differisce, anche notevolmente, per valore da quello qui trovato. Potremmo allora dedurre da ciò un diverso grado di coesione all'interno del gruppo e magari scoprire che esso è stato per certi periodi diviso in sottogruppi isolati (dal punto di vista delle politiche matrimoniali) oppure, al contrario, particolarmente incline all'imbreeding. Il tMRCA può quindi costituire una misura importante del contesto relazionale di un gruppo ed anche sulla formalizzazione di questo concetto potrebbero aprirsi interessanti percorsi di ricerca.

Molto rimane ancora da scoprire e verificare anche sul MCRA. Pur essendo lo studio della distribuzione in frequenza della ripetizione degli antenati un ambito scientifico che vanta una notevole tradizione, poco è stato analizzato

mettendo in relazione questo fenomeno con quello degli antenati comuni. Non a caso è stato scelto l'acronimo MCRA a sottolineare la stretta dipendenza tra antenati comuni per generazione e antenato comune più recente. Il risultato emerso nel confronto tra capostipiti e IAP, pur con tutte le limitazioni dovute al campione estremamente ristretto, mette in luce come sia necessario un approccio più ampio e integrato che non può prescindere anche da conoscenze trasversali, soprattutto se i dati derivano da una popolazione di individui realmente esistiti nella quale entrano in gioco fattori culturali, religiosi, politici, sociali di cui difficilmente si tiene conto nei modelli ma che devono essere presi in considerazione se vogliamo che il risultato a cui si mira abbia un valore anche pratico, di più approfondita comprensione del mondo che ci circonda e della sua storia, antica o recente che sia.

Abbiamo volutamente tenuto fuori da questo lavoro considerazioni storiche o sociali, concentrando l'attenzione sull'analisi quantitativa dei dati ma non possiamo tralasciare quella qualitativa e neanche quella genuinamente etica che spinge alla comprensione delle dinamiche umane. Cercare l'antenato di tutti, sia pure nel nostro caso di un gruppo chiuso, significa capire che gli esseri umani sono più vicini tra loro di quanto vogliano, o vogliano far, credere [Rohde et al. 2004].

Il più tangibile risultato di questa tesi, tuttavia, è la costituzione di un'ampia base dati genealogica, composta da quasi 200 mila individui rappresentanti nello specifico l'alta nobiltà europea ma più in generale un gruppo sociale

fortemente coeso, sufficientemente chiuso, omogeneo ma anche geograficamente distribuito dalla Norvegia alla Grecia e dal Portogallo alla Russia. Una sorta di enorme “villaggio”, come già più volte detto, nel quale valgono le regole delle comunità chiuse, soprattutto in termini di politiche matrimoniali, di linee di successione, di accoglienza o rifiuto degli “immigrati” (basti pensare alle famiglie annesse agli alberi all'epoca della Rivoluzione Francese).

Poter quindi disporre di una tale base dati è già di per sé un enorme passo avanti nella ricerca di tutte quelle proprietà dei gruppi sociali intuite, modellate matematicamente, studiate in teoria ma mai verificate nella pratica. La natura nobiliare degli alberi ha permesso infatti di risalire indietro nel tempo per diverse generazioni, cosa non sempre fattibile o agevole in comunità differenti e soprattutto non sempre verificabile tramite fonti storiche attendibili.

È risaputo, ed è stato ben evidenziato nel testo, che esistono basi dati genealogiche anche molto ampie, alcune delle quali hanno fatto da fonte alla base dati usata in questo lavoro, ma nessuna di esse consente un accesso diretto, strutturato e finalizzato ad un'analisi quantitativa su grandi numeri. Se la proprietà intellettuale dei dati spetta di diritto ai redattori di tali banche dati, a noi spetta il merito di aver saputo trasformare dati in informazione aggregata e quindi fruibile che è uno degli obiettivi del data mining soprattutto se applicato ai big data. Potremmo sostenere che in questo contesto non è il dato in sé ad avere valore ma l'insieme dei dati e la possibilità di accedervi in modo

mirato.

A questo punto, disponendo dei dati, sarà possibile indagare fenomeni genealogici di varia natura, dalle politiche matrimoniali, alla distribuzione dei cognomi, dalla cladistica delle famiglie ai fenomeni di isonimia e consanguineità. Sarà inoltre possibile fornire verifiche sperimentali a quelle teorie, più volte citate, che descrivono in modo statistico questi ed altri fenomeni, alcune delle quali hanno avuto, in questo lavoro di tesi, una prima sommaria trattazione che tuttavia ha lasciato intuire risultati promettenti.

Volendo fare un paragone, possiamo dire di aver aperto una miniera e di aver estratto solamente alcune pepite per mostrare la presenza di un filone ma la miniera è ancora tutta da esplorare.

Tabelle

Tabella A: Campione iniziale

N. Albero	Nome	Anno Nascita	Famiglia
1	Joachim Ernst, Herzog von Anhalt	1901	Anhalt
2	Berthold, Prinz von Baden,	1906	Baden
3	Alfonso, Duque de Calabria	1901	Borbone-Due Sicilie
4	Alfonso XIII de Borbón, Rey de España	1886	Borbón-España
5	Felix, principe di Borbone-Parma	1893	Borbone-Parma
6	Henri d'Orléans, comte de Paris	1908	Bourbon-Orléans
7	D. Duarte Nuno, duque de Bragança	1907	Bragança
8	Ernst August III, Prinz von Hannover	1887	Braunschweig
9	Ludwig, Prinz von Hessen und bei Rhein	1908	Hessen-Darmstadt
10	Philipp, Landgraf von Hessen	1896	Hessen-Kassel
11	Wilhelm, Prinz und Landgraf von Hessen	1905	Hessen-Philippsthal
12	August, Furst zu Hohenlohe-Oehringen	1890	Hohenlohe-Oehringen
13	Friedrich Karl III, Fürst zu Hohenlohe-Waldenburg-Schillingsfürst	1908	Hohenlohe-Waldenburg
14	Gottfried, Furst zu Hohenlohe-Langenburg	1897	Hohenlohe-Langenburg

N. Albero	Nome	Anno Nascita	Famiglia
15	Karol II von Hohenzollern-Sigmaringen, roi de Roumanie	1893	Hohenzollern-Sigmaringen
16	Wilhelm von Hohenzollern, Prinz von Preußen	1906	Hohenzollern-Preussen
17	Wilhelm Karl Hermann Prinz von Isenburg	1903	Isenburg
18	Ernst, Erbprinz zur Lippe	192	Lippe
19	Wolrad, Furst zu Schaumburg-Lippe	1887	Lippe-Schaumburg
20	Gottfried, Prince of Tuscany	1902	Lorraine-Toscana
21	Otto von Habsburg-Lothringen, Erzherzog von Österreich	1912	Lorraine-Österreich
22	Adolf Friedrich VI, Grand Duke of Mecklenburg-Strelitz	1882	Mecklenburg-Strelitz
23	Christian Ludwig, Herzog von Mecklenburg-Schwerin	1912	Mecklenburg-Schwerin
24	Charlotte, Grand Duchess of Luxemburg	1896	Nassau-Weilburg
25	Wilhelmina van Oranje-Nassau, koningin der Nederlanden	1880	Nassau-Oranje
26	Frederik IX von Schleswig-Holstein-Sonderburg-Glücksburg, konge af Danmark	1899	Oldenburg-Danmark
27	Olav V von Schleswig-Holstein-Sonderburg-Glücksburg, kong av Norge	1903	Oldenburg-Norge
28	Paul Ier von Schleswig-Holstein-Sonderburg-Glücksburg, roi des Hellènes	1901	Oldenburg-Hellenes

N. Albero	Nome	Anno Nascita	Famiglia
29	Vladimir Romanov, Grand-duc de Russie	1917	Oldenburg-Russia
30	Nikolaus, Hereditary Grand Duke von Oldenburg	1897	Oldenburg
31	Heinrich XLV Erbprinz Reuss	1895	Reuss-Gera
32	Heinrich XXIV, Furst Reuss zu Greiz	1878	Reuss-Greiz
33	Heinrich XXXIX, Prinz Reuss	1891	Reuss-Köstritz
34	Ernst Heinrich Prinz von Sachsen	1896	Sachsen
35	Boris III von Sachsen-Coburg und Gotha, tsar des Bulgares	1894	Sachsen-Bulgaria
36	Carl August, Erbgrossherzog von Sachsen-Weimar-Eisenach	1912	Sachsen-Weimar-Eisenach
37	Carl Eduard I, Herzog von Sachsen-Coburg und Gotha	1884	Sachsen-Coburg-Gotha
38	D. Manuel II de Bragança Saxe Cobourg Gotha, rei de Portugal	1889	Sachsen-Portugal
39	Erbprinz Georg Moritz von Sachsen-Altenburg	1900	Sachsen-Altenburg
40	Georg, Herzog von Sachsen-Meiningen	1892	Sachsen-Meiningen
41	Léopold III de Belgique, roi des Belges	1901	Sachsen-Belgique
42	Vittorio Emanuele III di Savoia-Carignano, re di Italia	1869	Savoia
43	Friedrich Gunther Furst zu Schwarzburg	1901	Schwarzburg
44	Georg Friedrich Graf zu Solms-Laubach	1899	Solms
45	Friedrich Furst zu Waldeck und Pyrmont	1865	Waldeck

N. Albero	Nome	Anno Nascita	Famiglia
46	Hermann Erbprinz zu Wied	1899	Wied
47	Albrecht Herzog von Bayern	1905	Wittelsbach
48	Philipp, Herzog von Wurttemberg	1893	Wurttemberg

Tabella B: Anonimi per albero

N. Albero	% anonimi
1	12,0
2	15,8
3	4,4
4	6,6
5	11,2
6	6,7
7	24,4
8	7,7
9	23,2
10	3,6
11	30,9
12	39,5
13	44,6
14	18,4
15	22,2
16	21,1
17	6,6
18	41,3
19	6,7
20	4,0
21	6,1
22	12,2

N. Albero	% anonimi
23	13,8
24	11,9
25	8,8
26	22,2
27	13,9
28	5,1
29	20,4
30	7,2
31	20,8
32	15,8
33	56,0
34	6,4
35	13,0
36	20,4
37	4,3
38	11,3
39	8,1
40	38,6
41	20,1
42	18,0
43	26,7
44	39,9

N. Albero	% anonimi
45	7,8
46	3,8

N. Albero	% anonimi
47	8,2
48	4,1

Tabella C: Dimensione del campione iniziale nello studio del MRCA

	I	II	III	IV	V	VI
Generazioni genealogiche	48	96	140	208	240	294
Generazioni anagrafiche	49	89	93	106	92	94

Bibliografia

ATHREYA, K.B., NEY, P.E. (2004). *Branching Processes*, Dover Publications.

CHANG, J.T. (1999). Recent Common Ancestors of All Present-day Individuals, *Adv. Appl. Prob.* **31**, 1002–1026.

CROW J.F., MANGE A.P. (1965). Measurement of inbreeding from the frequency of marriages between persons of the same surname, *Eugenics Quarterly* **12**, 199-203 .

DERRIDA B., MANRUBIA, S.C., ZANETTE, D.H. (1999). Statistical Properties of Genealogical Trees, *Phys. Rev. Lett.* **82**, 1987 – Published 1 March 1999.

DERRIDA B., MANRUBIA, S.C., ZANETTE, D.H. (2000a). On the genealogy of a population of biparental individuals, *J Theor Biol.* Apr 7, **203(3),** 303-15.

DERRIDA B., MANRUBIA, S.C., ZANETTE, D.H. (2000b). Distribution of repetitions of ancestors in genealogical trees, *Physica A* Volume **281**, Issues 1–4, 15 June 2000, pages 1–16.

EVANS S.N., RALPH P.L. (2010). Dynamics of the Time to the Most Recent Common Ancestor in Large Branching Population, *Annals of Applied Probability*, Vol. **20**, No. 1, 1-25.

FOX W.R., LASKER G.W. (1983). The Distribution of Surname Frequencies, *Int. Stat. Review* **51**, 81-87.

LASKER G.W. (1977). A Coefficient of Relationship by Isonymy: A Method for Estimating the genetic Relationship between Populations, *Human Biology* **49**, 489-493

LOTKA A.J. (1931). The extinction of families I-II, *J. Wash. Acad. Sci.* **21**, 377-380, 453-459.

NEWMAN M.E.J. (2005). Power laws, Pareto distributions and Zipf's law, *Contemporary Physics* **46**, 323-351.

RALPH P., COOP G. (2013). The Geography of Recent Genetic Ancestry across Europe, *PLoS Biol* **11**(5): e1001555. doi:10.1371/journal.pbio.1001555.

ROHDE D.L.T., OLSON S., CHANG J.T. (2004). Modelling the recent common ancestry of all living humans, *Nature* **431**, 562-566.

ROSSI P. (2013). Surname distribution in population genetics and in statistical physics, *Physics of Life Reviews* **10,** 395-415.

ROSSI P. (2015). Self-similarity in Population Dynamics: Surname Distributions and Genealogical Trees, *Entropy* **17**, 1-13.

SERVA M., PELITI L. (1991). A statistical model of an evolving population with sexual reproduction, *Journal of Physics A* **24**, 705-709.

Genealogie cartacee

AA. VV. (1911). *Cambridge Modern History - Genealogical Tables*, Cambridge.

GEORGE H.B. (1904). *Genealogical Tables illustrative of Modern History*, Clarendon, Oxford.

ISENBURG W.K. (1975). *Europaeische Stammtafeln Band I-II*, Stargardt, Marburg.

LORENZ O. (1908). *Genealogisches handbuch der Europaeischen Staatengeschichte*, Gottasche V., Stuttgart.

LOUDA J., MACLAGAN M. (1981). *Lines of succession*, Orbis Pub.. London.

SCHWENNICKE D. (1978) *Europaeische Stammtafeln Band VI-VII*, Stargardt, Marburg.

SCHWENNICKE D. (1998). *Europaeische Stammtafeln Neue Folge Band I.1-3*, Klostermann, Frankfurt.

STOKVIS A.M.H.J. (1888). *Manuel d'histoire, de généalogie et de chronologie I-III* , Leiden (reprint).

THIELE A. (1991). *Erzaehlende genealogische Stammtafeln zur europaeischen Geschichte*, R.G. Fischer V., Frankfurt.

Sitografia

(Tra parentesi, accanto al nome del sito, è indicato il curatore)

Ancestry of Charles II King of England (Neil Thompson e Charles Hansen)

http://fmg.ac/projects/charles2

(fa parte di FMG - Foundation for Medieval Genealogy)

Charles-Quint's ancestors (JoseVerheecke)

http://users.telenet.be/JoseVerheecke/vorstenhuis/charles5/

Descendance de Louis IX roi de France (André Decloitre)

http://gw.geneanet.org/genroy

FamilySearch (Chiesa di Gesù Cristo dei Santi degli Ultimi Giorni - Mormoni)

https://familysearch.org/

Genealogics (Leo van de Pas)

http://www.genealogics.org/

Genealogie delle famiglie nobili italiane (Davide Shamà)

http://www.sardimpex.com/

Geni (MyHeritage Ltd.)

http://www.geni.com/genealogy-resources

Henry project - The ancestors of king Henry II of England (Stewart Baldwin e Todd Farmerie)

http://sbaldw.home.mindspring.com/hproject/henry.htm

Medieval Lands (Charles Cawley)

http://fmg.ac/Projects/MedLands/

(fa parte di FMG - Foundation for Medieval Genealogy)

Roglo (Daniel de Rauglaudre)

http://roglo.eu/roglo

WW-Person (Herbert Stoyan)

http://wwperson.informatik.uni-erlangen.de/ww-person.html

(non più raggiungibile, disponibile su CD-Rom)