Using Data Visualization to Uncover Demographic Trends in Early 1900s Irish Census Records

Corey Emery, Kayla Frisoli

Carnegie Mellon University
Statistics & Data Science

Overview

Using recently available 1901 and 1911 Irish national census records, we study changing demographic factors over this time interval.

One focus of our analysis is to better understand changes at the household level, so that researchers can utilize this information to inform record linkage models.

These models will be employed to statistically merge the two databases and find potential 1911 matching counterparts for the 1901 records.

Our work to understand and visualize the data across the two years will be instrumental to these models as well as the study of Irish history.

Analysis and Computation

The data we use was made available via the National Archives of Ireland, which can be found at http://www.census.nationalarchives.ie/, and is recorded for approximately 3600 divisions (called townlands in rural areas and streets in urban areas) across Ireland's 32 counties.

Common First Name Visualization

We created a unique identifier by combining the townland and county name to address the issue of repetitive townland names.

We were then able to aggregate a count for each name and isolate the most common for males and females.

Currently name standardization includes ignoring case, but has not yet been augmented to incorporate common nicknames.

We used two different regular expression matching schemes to map our aggregated data to geographic coordinates from OpenStreetMap, since apostrophes are inconsistently replaced with spaces and empty strings.

Religion Response Standardization

To address highly inaccurate spellings for various religions, we relied instead on phonetic encodings to assess the similarity of responses.

We used the soundex algorithm to determine similarity, which is widely used for English names and words.

We inspected the ten most common encodings, as these accounted for 96% of townlands, and plotted those that clearly mapped to one religion.

Raw and Aggregated Data

Year	County	DED	TownStreet	Number	ID	Surname	Forename	Age	Sex	RelationHead	Religion	Birthplace	Occupation	Literacy
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Wetherall	Ellen	14	Female	Daughter	Church of Ireland	Co Antrim	Nurse	Read and write
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Beckett	Annie	22	Female	Daughter	Church of Ireland	Co Antrim	House Keeper	Cannot read
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Beckett	William	*NA*	Male	Grand Son	Church of Ireland	Co Antrim	–	Cannot read
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Wetherall	John	52	Male	Head of Family	Church of Ireland	Co Armagh	Labourer	Read and write
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Wetheral	Henretta	17	Female	Niece	Church of Ireland	Co Antrim	Agricultural Labourer	Cannot read
1901	Antrim	Aghagallon	Aghadrumglosney	1	1002268	Wetherall	Mark	11	Male	Son	Church of Ireland	Co Antrim	Scholar	Read and write

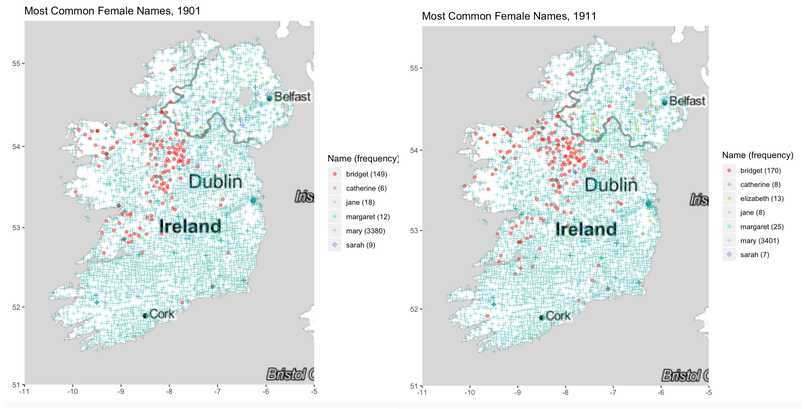
Religion	phoCodes		
Presbyterian	P621	Irish Church	I622
Catholic	C342	Roman Catholic	R552
Roman Catholic	R552	Roman Catholic	R552
Roman Catholic	R552	Roman Catholic	R552
Roman Catholic	R552	R C	R200

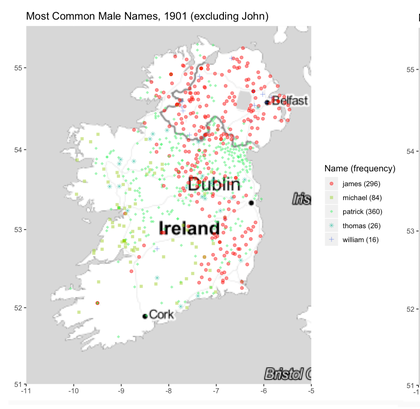
mapping_code	R552	R234	P621
mapping_religion	Roman Catholic	Roman Catholic	Presbyterian

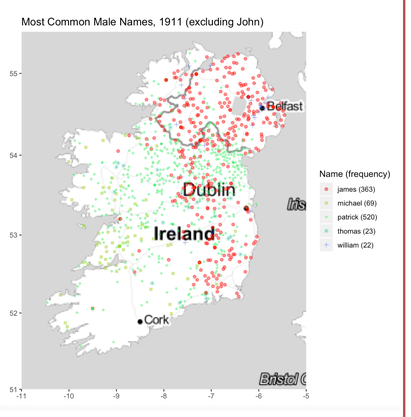
C342	R200	I622	M332	E121
Roman Catholic	Roman Catholic	Irish Church	Methodist	Episcopalian

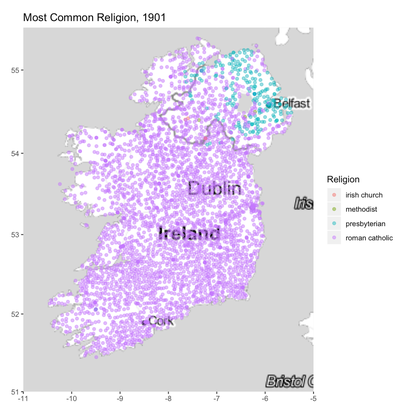
Here we show a sample of our original dataset from 1901 (top), responses for the religion question in the 1901 census data paired with the corresponding soundex encoding (bottom left) and our mappings between religious denomination and phonetic code based off our manual examination (bottom right).

Name and Religion Visualizations









Here we see the most common female first names in 1901 and 1911 (top left), the most common male first names in 1901 and 1911 (top right), as well as the most common religions in 1901, based on our phonetic encodings (bottom). In each graph, one point represents one street/townland. Especially in the graphs for male first names and religion, we can begin to see the underlying demographic factors that would lead to the split between Ireland and Northern Ireland.

Impact and Future Work

Highly common first names pose limitations for record linkage techniques that rely on string similarity. We can attempt to weight the similarities using frequencies to emphasize matchings for rarer names.

We can incorporate phonetic encodings as a way to standardize the religion data and gain more meaningful and useful metrics from them, which might also prove to be an important factor in linking individuals.