Background

This research aims at analysing relationships of families and households from historical census data. Historical census data is collected in different time, and it contains valuable information on the social status in a period of time. Such information includes family structure, household, and so on. A historical census sample is shown in Figure 1, which is a scanned hand-filled census sheet.

Datasets

The datasets to be used contain historical census data collected from the district of Rawtenstall in North-east Lancashire, the United Kingdom (Figure 2). The datasets originate from six censuses that took place every ten years since 1851, which generated around 160,000 records. Important information embedded in the datasets includes name, address, age, sex, relationship to the head of households, occupation, and the birth place.

Goals

We expect to develop efficient and effective methods to
- minimise manual cleaning and analysis which are traditionally done by the social scientists,
- link families and households data across time to reconstruct the changes or shifts of families in an era,
- build data linkage methods to improve the data quality and enrich social, economic and environmental research.

Challenges

This research includes following challenges:
- The information collected by the census is limited.
- The quality of the historical census data is poor, it contains lots of errors and incomplete data.
- Historical censuses done in different years do not have a common set of characteristics. Only limited attributes are comparable between datasets.
- Historical census was collected across time. The household information has likely changed dramatically in the span of 60 year.

Methods

The proposed research will focus on four technical aspects, which include data cleaning, data linkage, iterative data preprocessing and linking, as well as visualization.
- Data cleaning: Develop data cleaning methods to clean "whole of household" records as a group.
- Data linkage: Develop reliable probabilistic data matching algorithms to ensure accurate census data linking. This technique is built upon the above results.
- Iterative data preprocessing and linking: We will investigate combined iterative data cleaning and linking methods. This will be a two-step data cleaning method. In the first step, traditional data cleaning and matching are performed. Then in the second step, linked records with high probabilities will be iteratively cleaned and linked. We expect to improve the linking accuracy and efficiency.
- Visualization: Represent data linking results using graphical methods.

Outcome

The expected outcome includes
- new data cleaning and matching methods that are based on probabilistic modelling of temporal relational data,
- methods to allow social scientists to produce better quality historical data more efficiently and conveniently.

Figure 1: A sample from historical census.

Figure 2: Rawtenstall, Lancashire.

Figure 3: Electronic data sample from the 1891 UK Historical Census Dataset.

Figure 4: Temporal matching across time.

Figure 5: Resulting Family Tree