SCOTTISH CENTRE FOR ADMINISTRATIVE DATA RESEARCH

DATA INSIGHTS

Automatic Coding of Occupations: Methods to create the Scottish Historic Population Database (SHPD)

Authors: Richard Tobin, Claire Grover, Beatrice Alex, Lee Williamson, Eilidh Garrett & Chris Dibben

Date: November 2021

DOI: 10.7488/era/1506

The Digitising Scotland project digitised 25.8 million Scottish civil registration vital events records, including digitising birth, marriages and deaths, from when records began in 1855 until 1973 (from 1974 records became digital). To use these pre-digital records effectively for large-scale full life history research, they must not only be made machine-readable, but also coded in a suitable research ready format – including the broad classification of some of the information gathered.

WHAT WE DID

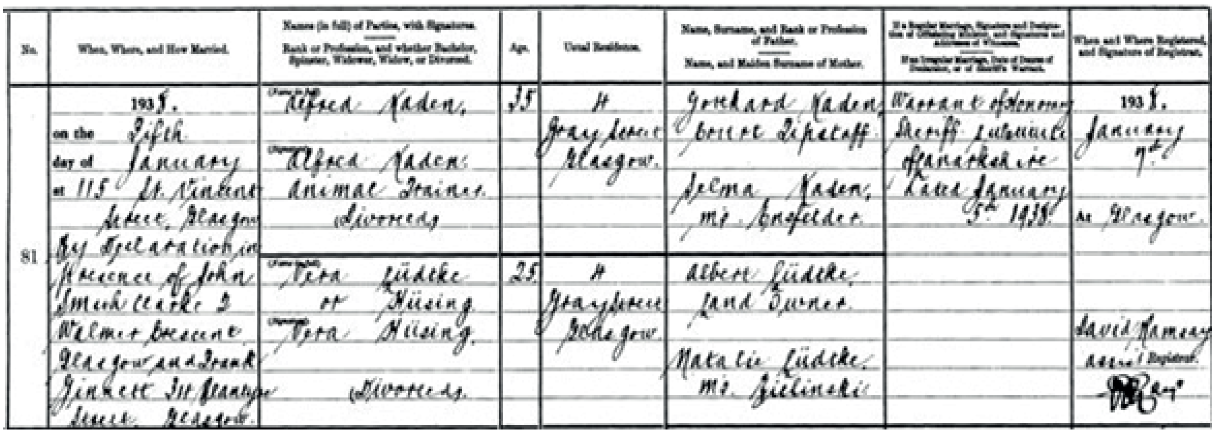
The overall aim of our SCADR work is to help create the research ready Scottish Historic Population Database (SHPD) by coding the information included on the digitised birth, marriage, and death certificates – namely the textual descriptions of occupations (see Figure 1) and the cause of death – to widely used standard coding schemes.

For SHPD, we are coding transcribed occupations to the Historical International Standard Classification of Occupations (HISCO) and the causes of death to the International Classification of Diseases, 10th revision (ICD-10).

Figure 1 shows the marriage record of Alfred Kaden and Vera Husing in 1938. The entry shows the groom, Alfred, to be an animal trainer and the bride's father to be a land owner. This methodological work focuses on converting occupation descriptions to the recognised international standard of HISCO.

It is impractical to have domain expert's hand-code all the 31 million occupations to create SHPD. To resolve this problem, we are treating coding as a text classification task, and automating the process by applying Natural Language Processing and Machine Learning techniques. To facilitate the auto-coding a proportion of the records – a random sample of 90,000 occupations unique strings – were recently manually coded and will now be used to train the system.

Figure 1: Published by National Records of Scotland (NRS) on their Twitter (@NatRecordsScot on 24/11/2017): https://twitter.com/NatRecordsScot/status/934062831426785280.



No.	When, Where, and How Married.	Names (in full) of Parties, with Signatures. Rank or Profession, and whether Bachelor, Spinster, Widower, Widow, or Divorced.	Age.	Usual Residence.	Name, Surname, and Rank or Profession of Father. Name, and Maiden Surname of Mother.	[illegible]	When and Where Registered, and Signature of Registrar.
81	1938. on the Fifth day of January at 115 St. Vincent Street, Glasgow By declaration in presence of John Smith Clarke, 2 Walmer Crescent, Glasgow and Frank Jinkett, 311 [illegible] Street, Glasgow	Alfred Naden, Alfred Naden. Animal Trainer. Divorced	35	4 Gray Street Glasgow.	Gotthard Naden, [illegible] [illegible]. Selma Naden, ms. [illegible]	Warrant of [illegible] Sheriff [illegible] of Lanarkshire dated January 5th 1938.	1938. January 7th At Glasgow.
		Vera Lüdke or Hüsing, Vera Hüsing. Divorced.	25	4 Gray Street Glasgow.	Albert Lüdke, Land Owner. Natalie Lüdke, ms. Zielinski		David Ramsay Asst. Registrar. [illegible]

A set of pilot data was hand-coded by domain experts and there was then a need to create further training data. The Centre for Data Digitisation and Analysis (CDDA) at Queen's University are creating both occupation and cause of death coded data for this purpose. The occupation strings have been combined from all the SHPD sources (births, marriages and deaths) and then a random sample generated providing a certain proportion of unique strings for CDDA to code.

Due to incorrect spellings in the original text and small transcription errors (likely resulting from unreadable historical handwriting combined with the instructions to 'key what you see' rather than trying to interpret the words) there will be a number of duplicate occupations coded. This has the useful effect of allowing us to assess the consistency of the coding. When coding occupations, our colleagues at CDDA did not have any other information other than the occupational string (ie.no year, age or sex for context). Once this coding is complete, we will conduct further experiments on full training data of 90,000 occupations coded to HISCO with the aim of auto-coding the 31 million SHPD occupations.

Ahead of the direct matching or auto-coding, initial pre-processing, cleaning and standardising was done on both the raw transcribed occupations and the training data. This included removing white space and unreadable characters identified as part of the transcription. This reduced the 31 million occupations down to 2 million uniques (or 1.6 million removing those with unreadable characters).

WHAT WE FOUND SO FAR

Preliminary experiments have been undertaken using a relatively small pilot dataset (birth records from 1900 and marriage records from 1930) and obtained reasonable results from a combination of exact matching and statistical classification.

In the pilot, by combining exact matching for texts that have been seen in the training data and the Bayes classifier for the rest, the accuracy levels achieved from cross-validation are 94-97%.

Currently, experiments using a larger section of coded training data (50,000 occupations) have uncovered a number of interesting features:

- Because some occupations are very common, a relatively small set of manually coded strings covers a very large proportion of the SHPD records. In applying these 50,000 different randomly-selected strings we have coded 22 million of the 25 million records (almost 88%). At time of writing, not all 31 million were available to process.
- A significant number of the records that are not exact matches contain spelling errors or illegible letters.

- Some records contain more than one occupation, particularly in war years when it is common for both a civilian and a military occupation to be recorded. Dividing these into the individual occupations would allow more to be matched exactly.
- The military occupations often contain details such as a regiment which is irrelevant to the rather coarse HISCO coding, and this usually prevents them from being exactly matched.
- Present work includes CDDA undertaking a review of coded occupations, part of which was to review spelling in the occupation strings coded. This in turn provides a large corpus of correctly-spelled, relevant words that can be used to detect and correct errors in the un-coded data. Examples of possible corrections are: BULLDER to BUILDER, FISHRMAN to FISHERMAN, FARMA to FARM, and PLOUGHMNA to PLOUGHMAN).

WHAT'S NEXT?

This is work in progress to create a research ready Scottish Historic Population Database (SHPD) and future work will include the results from experiments using the full training data (90,000 occupations) covering the whole period from 1855 to 1973 to auto-code the 31 million SHPD occupations. Future activities include building a tool using various techniques to split strings into individual occupations. For example, if a string consists of two strings that were coded as single occupations, it is split into those two strings. The splitter also recognises and splits off war-time military occupations where they are present, and extracts the rank. As HISCO only provides four different codes for military occupations, we expect to be able to algorithmically assign codes to these occupations when they are identified by the splitter.

WHY IT MATTERS

By completing both current and future work, it is envisaged that **this will allow well over 90% of the occupations to be coded** as exact matches. Those strings that have not been matched will be passed to a statistical classifier trained on the manually coded records.

Excitingly, once occupations are fully coded, this will be a fantastic resource for researchers to access to some 23 million individuals dating back to 1856. Finally, for the first time the UK will have a data system of similar depth and breadth to those in Scandinavia.

The Scottish Centre for Administrative Data Research (SCADR) analyses public sector data, using new data sources and novel methodologies to deliver cutting-edge applied research with real-world impact. We work with our partners to provide evidence-based insights, which inform policy and practice for public benefit.

Data Insights is produced by the Scottish Centre for Administrative Data Research
Website: www.scadr.ac.uk | Twitter: @scadr_data | Email: scadr@ed.ac.uk

The centre is a multi-institutional initiative hosted at the University of Edinburgh, which is a charitable body registered in Scotland (Registration number: SC005336).

SCOTTISH CENTRE FOR ADMINISTRATIVE DATA RESEARCH

UKRI Economic and Social Research Council

ADR SCOTLAND