Building the Addressing Health Pensions Database

Harry Smith

King's College London, Department of Geography

harry.j.smith@kcl.ac.uk

Working Paper 1

Addressing Health: Morbidity, Mortality and Occupational Health in the Victorian and Edwardian Post Office

https://addressinghealth.org.uk/

Funded by Wellcome Trust grant 217755/Z/19/Z

15 November 2023

Comments are welcome on this paper, please contact the author directly

Harry Smith, 'Building the Addressing Health Pensions Database', Addressing Health Working Paper 1 (2023) DOI:10.13140/RG.2.2.27152.17920

Keywords: mortality, health, United Kingdom, Post Office, pensions

JEL codes: J12, J18, N33

This paper describes the creation of the Pensions Database for the Wellcome-Trust funded project 'Addressing Health: Morbidity, Mortality and Occupational Health in the Victorian and Edwardian Post Office'.[1] This database contains information on all postal workers granted pensions or gratuities between 1860 and 1908. Any established postal worker who retired after ten or more years' service was eligible for a pension; those who retired having worked for the Post Office for fewer than ten years could be awarded a gratuity.[2] For ease, unless specified, in this paper 'pensioners' will be taken to refer to both those postal workers receiving pension and those who received gratuities.

The information on each pensioner is collected from three separate sources: the Index to Pension Application Forms held at The Postal Museum archives (TPM), the Pension Application Forms submitted by the Post Office to the Treasury held in the series POST 1 at TPM; and the Parliamentary Paper lists of pensions granted and ceased published in the *House of Commons Sessional Papers*.[3] For pensioners retiring in 1861, 1871, 1881, 1891 and 1901, their death certificates were also obtained and transcribed. The information contained in each of these sources had to be extracted, cleaned, linked and coded; this paper describes how this was done. The first section details the three main sources. The second covers the methods by which the data were extracted and cleaned. The same individuals appear in each source, so the information on each person needs to be linked together and the third section describes the record linkage methods used for this purpose. The fourth section details the coding schemes used to categorise the information on each pensioner and the methods used to enrich the data and account for missing variables and to enrich existing ones.

The Sources

As noted above, three sources provided information on the postal workers included in this database. Initially it was planned to just collect the information on retired postal workers from the Pension Application Forms from the POST 1 series held at TPM. However, the Covid-19 pandemic restricted access to TPM to photograph and transcribe these forms so alternative sources of data were used in the meantime: the Index to Pension Application Forms held at TPM and published lists of pensions granted and ceased. These two additional sources contain much of the information found in the pension forms and allowed some data to be gathered without access to the original forms, although there are some notable gaps in the availability of certain information and in the coverage of postal workers. This section describes the data contained within each source.

Pension Application Forms

The Pension Application Forms contain information about each postal worker that applied for a pension or gratuity.[4] The information collected comprises the following details:

---

[1] This work was funded by the Wellcome Trust under grant 217755/Z/19/Z 'Addressing Health: Morbidity, Mortality and Occupational Health in the Victorian and Edwardian Post Office'. A version of the database is available from https://addressinghealth.org.uk/data-mapper/. The full database will be made available via UKDS.

[2] David R. Green, Douglas H.L. Brown and Katheen McIlvenna, 'Addressing Ill Health: Sickness and Retirement in the Victorian Post Office', *Social History of Medicine*, 33/2 (2020), 561.

[3] The index to Pension Application Forms is held at The Postal Museum archives (hereafter TPM), POST 1/175, Index to Treasury letters, 1860-1882; POST 1/207, Index to Treasury letters, 1883-1887; POST 1/244, Index to Treasury letters, 1888-1893; POST 1/293, Index to Treasury letters, 1894-1899; POST 1/342, Index to Treasury letters, 1900-1903. The Pension Application Forms for our period are found within POST 1/106-402, volumes of Treasury letters covering 1860-1908. The published lists of Post Office pensions granted and ceased for our period are found with the *House of Commons Sessional Papers* accessed through *Proquest U.K. Parliamentary Papers*, https://parlipapers.proquest.com. A full list of the individual papers used is provided in Appendix A.

[4] Images of the pension forms can be viewed at https://data.addressinghealth.org.uk/

- name
- occupation
- location of workplace
- type of application made
- age at retirement
- length of service
- cause of retirement
- the dates at which the duties ceased and the salary stopped being paid
- the number of days off taken over the previous ten years broken down by whether they were sick days or taken for other reasons,
- the date that the form was submitted to the Treasury.[5]

These data come in the form of responses to questions on a form that is three to four pages in length. There are two changes to the form that affect the data available. First, in the 1890s the age question also asked for the applicant's date of birth to be reported. Secondly, the information on days off was provided by a table which listed initially contained three columns, one for the year, one for the number of days off owing to sickness and the one for days off for other reasons. From the 1890s an additional column and the count of days of for non-sickness reasons was split in two: one column reported the days of ordinary leave taken in each year and the other provided the number taken for any other reason.[6]

For most of the period covered by this database there were no substantial alterations to the questions and therefore the data derived from the answers to these questions should be consistent.[7] However, towards the end of this period there is a major change to the cause of retirement data. The answers to this question provide precise causes of retirement up until midway through 1900 when they stopped giving specific causes of retirement and began to only state whether a worker retired for reasons of 'ill health' or age.

The Index to the Pension Application Forms

Given the large size of the POST 1 series of material held at TPM, a series of indexes exist to allow the volumes to be navigated. One of the items indexed from the POST 1 series are the Pension Application Forms. This Index provides the name of the pensioner, their occupation, their location of work, the year they applied for a pension, and the volume and page number of the form within the POST 1 series. This source covers both workers who applied for a pension and those who applied for a gratuity.

This source has two complications. First, it contains a considerable number of duplicate entries for various reasons. The most common reason for someone to appear twice in this list was marriage as women were generally, but not always, listed twice, once under their maiden name and once under their married name. Some people applied for a pension

[5] The forms also contain information on salary at retirement, summaries of the worker's appointments and other career landmarks (such as the date of their civil service certificate if applicable), and a general statement on 'the manner in which the applicant has discharged his duties' that varies from single sentences to extensive career histories. These data have not been collected for every postal worker but have been consulted where useful.

[6] The 'other' reasons were only sometimes specified, they included enforced leave when a someone in a postal worker's household had a notifiable infectious disease, leave to attend funerals, training with reserve military units, and the generic term 'special leave'. These and other notes were sometimes found in the earlier, three-column, table but increased in frequency after the change in the form.

[7] Consistency may, however, still be affected by changes in organisational, or other, definitions of the data provided; for example, changes in the definition of 'sickness' or in the Post Office's policy towards sick leave may mean that the information extracted from the sickness table may not be comparable across time.

multiple times, having been rejected on previous occasions, or had their pension amount adjusted, necessitating an additional form be sent to the Treasury, or their original application was never dealt with, meaning they had to apply again. Some people had forms submitted in order to prove their right to a pension, but did not actually retire until some years later. Sometimes, these duplicates refer to multiple forms, sometimes just to correspondence between the Treasury and the Post Office about forms already, or about to be, submitted. In all such cases these individuals appear in the Index more than once. Such multiple entries have to be identified by hand and removed.

Secondly, the year of application was not always the year of retirement. Where this was the case, the pensioner usually ceased work in the year prior to application, presumably just reflecting administrative delays. However, in a few cases forms were submitted a decade or more after their retirement. In these cases the worker had often transferred from the Post Office to another branch of the Civil Service and presumably this form provided information to allow that individual's total length of service within any branch of the civil service to be calculated.[8] In other cases, administrative errors and disagreements prevented pension applications being submitted for some years.[9] In all cases the year of retirement provided by the Pension Application Form has been used to correct the year given by the Index.

Parliamentary Papers

Lists of all the workers granted a pension by the state were published each year by the House of Commons in papers initially entitled *Account of Retired Allowances or Superannuations in Public Office*. These were first published in 1831-2 and expanded over the course of the nineteenth century as pensions were granted to more areas of the Civil Service, including pensions granted to postal workers.[10] Between 1831-2 and 1881 these lists provide information on every pension granted in the previous calendar year.[11] In the early 1880s this reporting method changed. The final yearly account was that published in March 1882 covering pensions granted between 1st January and 31st December 1881.[12] After this date the lists of pensions granted are published in the reports entitled *Estimates for Revenue Departments* published each year. The *Estimates* had been published since 1854 and provided summaries of the expenditure by various Civil Service departments to be approved by Parliament.[13] The level of detail given in these summaries increased over time, but prior to 1884 the reports only gave the total spent in each year on superannuations, not the individual-level data on each pensioner that was given in the *Account of Retired Allowances or Superannuations*. The *Estimates* for 1884-5 was the first to contain the individual-level lists of the recipients of pensions as well as the general accounts of expenditure by the Post Office

[8] For example, TPM, POST 1/241, 90, pension form of Edward Gay, 1893, who ceased working for the Post Office on the 8th September 1856 when he was transferred to work as a supplementary clerk at the Treasury. He had a long career in various parts of the Civil Service, serving as Controller and Auditor-General, and Head Commissioner of Paper Currency in the Indian Civil Service, see *Monthly Notices of the Royal Astronomical Society*, 69 (1909), 246-7.

[9] For example, TPM, POST 1/124, 314, pension form of Francis Shute, 1866, a Postmaster in Crediton who retired in 1859 but did not apply for a pension until 1866 because the District Surveyor covering Shute's Post Office thought that a Postmaster whose net income fell below a certain threshold was not eligible for a pension. This was not the case, and Shute himself wrote to the Postmaster General asking for aid given he was in a position of 'great necessity'.

[10] *Account of Retired Allowances or Superannuations in Public Office, 1831*, *Parliamentary Papers*, 26 (1831-32).

[11] Thus, *Account of Retired Allowances or Superannuations in Public Offices, 1861*, *Parliamentary Papers*, 31 (1862) provides a list of all pensions granted between 1st January 1861 and the 31 December 1861.

[12] *Account of Retired Allowances or Superannuations in Public Office, 1881*, *Parliamentary Papers*, 37 (1882).

[13] *Estimates, revenue departments, for the year 1854, ending 31 March 1855*, *Parliamentary Papers*, 40 (1854).

and other Civil Service departments.[14] This first report containing pensioners' details covered those postal workers who were awarded a pension between 1st January and 31st October 1883. The next report, published February 1885 covered pensions granted between 1st November 1883 and 30th November 1884.[15] After that report, each subsequent report published details of the pensions granted between 1st December one year and 30th November the subsequent year.[16]

The shift from publishing the lists of pensions granted in the annual *Account of Retired Allowances or Superannuations in Public Office* to publishing them in the *Estimates for Revenue Departments* was not seamless, and no information on pensions granted in 1882 exists in the *Parliamentary Papers*.[17]

The lists of pensions granted provide the name of each pensioner, the location of their workplace, their occupation, age at retirement, length of service, final salary, cause of retirement and the value of their annual pension. The information on their place of work is only consistently given for all pensioners for pensions granted after 1st January 1883, in other words from when they start being published in the *Estimates for Revenue Departments*. The other substantial change is that the cause of retirement for pensioners is given in detail prior to the 1886-7 *Estimates*, which provided the list of pensions granted between 1st December 1884 and 30th November 1885; from that list onwards cause of retirement was only specified as owing to 'ill health', age or, occasionally, administrative reasons, such as roles being abolished.

The most notable problem with these lists is that they only list pensions granted, they do not contain details of gratuities or other kinds of awards granted.[18] This means that they contain information on fewer postal workers than the Pension Application Forms or the Index to said forms. For example, between 1861 and 1869 the Index reports 2,266 postal workers applying for pensions or gratuities, whereas the *Parliamentary Papers* lists give information on just 1,953 pensioners. The difference between these two sources increased over time as more gratuities were granted, thus in the 1890s the Index lists 7,731 pensioners and gratuitants while the pensions granted lists give only 4,115 pensioners for that decade. Gratuities were not listed because they were one-off payments, rather than continuing expenses for the state to cover, and so it was not necessary to provide Parliament with the same detailed information about them. Figure 1 shows the number of pensioners and gratuitants by year to better illustrate how the Parliamentary Paper lists cover less of the total workforce towards the turn of the century.[19] As the majority of female workers received gratuities rather than pensions, the lists of pensions granted are particularly deficient when studying female retirement.

The publications which provided lists of the pensions granted also gave lists of the pensions that ceased during the period covered by the publication. For the majority of cases, the reason for payment ceasing was death of the pensioner, but some were commuted, some ceased because the pensioner had been convicted of a crime and some pensioners returned to work for the Post Office. These lists provided the name of the pensioner, their occupation and location of work, the reason for the pension stopping, the date on which the pension ceased, and the value of the pension being paid. If the pension ceased owing to the

[14] *Estimates for revenue departments for the year ending 31 March 1885*, *Parliamentary Papers*, 52 (1884).
[15] *Estimates for revenue departments for the year ending 31 March 1886*, *Parliamentary Papers*, 50 (1884-5).
[16] The lists were published in arrears; for example, the *Estimates* for 1888-9 published in March 1888 contained information on pensions granted between December 1886 and November 1887.
[17] A full list of the Parliamentary Papers consulted is found in Appendix A.
[18] Various other types of award were made, such as allowances or grants to relatives of postal workers who died while in service.
[19] The peaks in 'compensation' awards in the early and late 1870s reflect the reorganization of the workforce following the nationalization of the telegraph companies.

death of the pensioner, the date the pension ceased was also the pensioner's date of death. As with the lists of pensions granted, the location data was only consistently given for pensions ceasing from 1883 onwards. Otherwise, the data provided by these lists was consistent across the entire period.

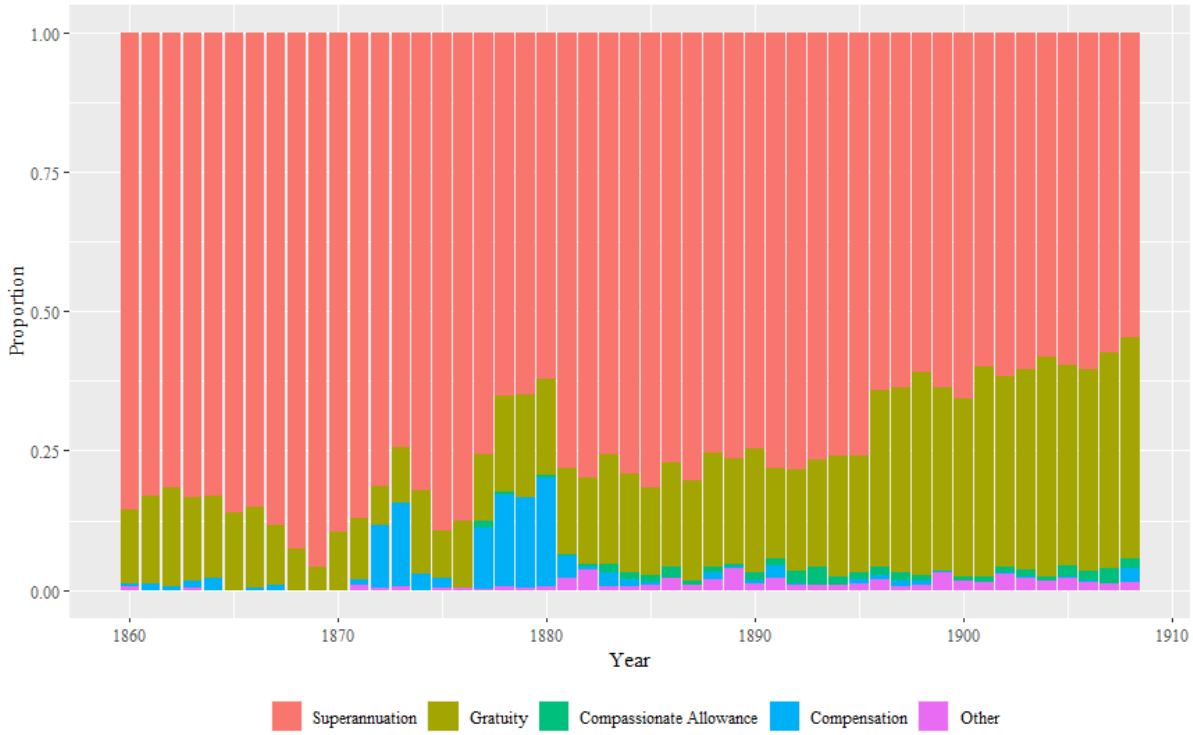

Figure 1. Category of retirement award by year, 1860-1908.
Source: Addressing Health Pensions Database.

The Index and the published lists of pensions granted provide much of the information on pensioners contained within the Pension Application Forms held in the POST 1 series at TPM. The only information missing for those postal workers who claimed a pension is the table of days off taken in the ten years prior to retirement and, from December 1884-1900, the precise information on the cause of retirement. The coverage of gratuitants is substantially less complete. The pensions granted lists contain few postal workers granted gratuities, and while the Index contains all individuals granted gratuities it only provides their name, occupation, year of application and place of work. Therefore, while much of the data contained in Pension Application Forms can be recreated from other sources, access to those forms has been necessary to cover sickness absence for all workers, detailed cause of retirement for everyone from late 1884 onwards, and for gratuitants their POST 1 form is the only source for most of the relevant information about their career and health.

Data collection

Each of the three sources was digitised, transcribed and checked in different ways. This section describes the different processes and how they affected the nature of the data collected.

Pension Application Forms

The forms completed by postal workers applying for a pension or gratuity are held at TPM, and the transcription was undertaken using the Zooniverse platform. Zooniverse is a platform

that allows research projects to engage with members of the public to undertake large-scale transcription and classification projects.[20] A project was set up on their platform with separate workflows to guide users through the transcription of each relevant part of the Pension forms.

The pension records were then photographed by a combination of volunteers, staff at TPM and a photographer employed by Ancestry.com. The images were then processed by both staff at TPM, project team members and volunteers working remotely. This photography process produced high- and lower-resolution jpgs: the high-resolution jpgs were stored for the use by TPM, the lower resolution jpgs were uploaded to the Zooniverse to be transcribed by volunteers.[21] Each piece of information was transcribed by three to seven different individuals to ensure accuracy.

The results of the transcription process were then exported as .csv files containing the transcriptions produced by the Zooniverse users as well as various metadata. These .csv files were processed to extract the individual transcriptions which were then reconciled to produce a single transcription of each part of the pension form required by the project using the python script reconcile.py.[22] The reconciled transcriptions from each workflow were then further checked and linked together to produce a complete transcription of each individual's pension record.

These complete records were then checked to ensure that the transcriptions produced through the Zooniverse platform were accurate. The checks were made in three ways. First, the Zooniverse transcriptions were checked on their internal consistency. For example, an individual's age as transcribed from the form must always be greater than the transcribed length of service. Similarly, the dates provided in the table listing the number of days off an applicant had taken needed to agree with the dates of retirement provided on the form. Wherever these consistency checks were failed the original images were checked and the transcription corrected as necessary.

Secondly, Zooniverse transcribers are able to indicate if their transcription was uncertain by flagging it as 'unclear', all of these cases were checked by hand. Similarly, users sometimes left fields blank and in those cases the original images were checked to ensure that no data had been missed.

The third set of checks involved comparing the transcriptions from Zooniverse against other data. Some of the Zooniverse data could be checked against the data produced by the pilot project. The pilot project covered 1861, 1871, 1881 and 1891 and transcribed the pension records from those years, providing a ready set of data against which the transcriptions from Zooniverse could be tested. The results of these tests were highly encouraging. In 1861, 300 transcriptions of the days off table were made and just 3 of the Zooniverse transcriptions were wrong, equivalent to an error rate of 1 per cent; in 1871 the error rate for the same task was slightly higher as 10 out of 600 transcriptions from Zooniverse were wrong (1.67 per cent). Similar checks against other data were also carried out by comparing the Zooniverse transcriptions to the information extracted from the Parliamentary Paper lists of pensions granted and the Index of Pension Application Forms.

---

[20] https://www.zooniverse.org/ For more details on Zooniverse see Chris Lintott, *The Crowd and the Cosmos: Adventures in the Zooniverse* (Oxford, 2019).

[21] Zooniverse requires images uploaded to its platform to be less than 1mb in size.

[22] The python script reconcile.py was developed by the Notes for Nature Zooniverse project (https://github.com/juliema/label_reconciliations). This script takes multiple transcriptions of the same source and determines the 'best' transcription. When matching text fields, the script normalises the transcriptions provided and then uses a fuzzy matching algorithm to establish the most common transcription provided by users. To take a simple example, if a cause of retirement was given as 'phthisis' by five transcribers, 'phthsis' by one and 'pthsis' by another, reconcile.py would return 'phthisis' as the 'best' transcription. As the degree of disagreement between the seven different transcriptions increases so reconcile.py uses more sophisticated methods of identifying the most common transcription.

Once again, these checks suggest that the transcriptions produced by the Zooniverse users were high quality. Name, occupation, location, cause of retirement, age and length of service were all checked against the relevant external data.[23] From an initial batch of images transcribed using Zooniverse (covering the years 1860-62, 1871 and 1899) only 18 incorrect transcriptions were found from a total of 10,673 transcriptions, an error rate of just 0.17 per cent. These checks against Parliamentary Paper and Index records revealed that the Pension Application Form often contained more detailed versions of the same information. For example, Patrick Sweeney was a letter carrier who retired in 1862, his entry in the Parliamentary Paper lists of pensions granted stated that the cause of his retirement was 'Phthisis' but his Pension form reports his cause of retirement as 'Phthisis and diseased heart'.[24]

The first two checks were done on the raw Zooniverse data, while the third check could only be done once the Zooniverse data was linked to the Parliamentary Paper and Index data as detailed below.

The Index to the Pension Application Forms

The Index is held at TPM on microfilm. A digital version of this was obtained and it was then transcribed by volunteers. Each volunteer was given a number of pages to transcribe into a spreadsheet provided for the purpose of transcription. These transcriptions were then checked line-by-line by hand. The Index was only transcribed up to the end of 1902. After that date, the page numbers of each form were transcribed, but the name and other information and this brief transcription was solely used to guide the Ancestry.com photographer in finding pension forms within the POST1 volumes.

Parliamentary Papers

PDFs of the relevant reports were downloaded from the online version of the *U.K. Parliamentary Papers*. The pages listing the postal pensions granted and ceased were then extracted and ABBYY FineReader was used to transcribe these lists using OCR. The resulting outputs were then checked and rectified by hand.

Data linkage

The transcription and checking processes described above produced three separate sets of data containing information on overlapping groups of postal workers. Everyone in the Index should also have a Pension Application Form. Everyone reported in the Parliamentary Paper lists of pensions granted and ceased should also appear in the Index and Pension Application Forms. Consequently, it should be possible to link all three sources together. This section describes how that was achieved.

Linking the Index and the Pension Application Forms

The pension forms and the Index should overlap completely: everyone whose application form appears in the POST 1 series should also appear in the Index as it indexes the POST 1 series. The individuals in each transcription are easy to link as the information used to generate an individual's ID (year of retirement, POST 1 volume number and page number of their pension record within that volume) appear in both sources. The IDs were generated and

[23] Checks on text fields were done using Levenshtein distance scores, any transcription where the score was greater than 5 was checked by hand. Age and length of service were simply compared and any cases in which the Zooniverse and Parliamentary Paper transcription did not match exactly were checked by hand.

[24] Patrick Sweeney, id 18621120389; *An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1862*, Parliamentary Papers, 31 (1863), 31; TPM, POST1/112, 389-91, pension form of Patrick Sweeney.

then used to link the transcribed data derived from each source. Where IDs did not match the original sources were checked and corrections made.

However, not everyone who appears in the Index has a surviving pension form, some have been lost or were not photographed. In some cases, the Index itself was wrong in the page number given. Finally, as noted above, many of the Index entries do not refer to pension forms, but to letters about pension forms. For these reasons, therefore, not every entry in the Index corresponded to an extant, photographed, transcribed pension form. There are 470 such individuals in the final database where their Index entry could not be linked to a form, 1.8 per cent of the total database. Information is still available on these retirees from the Index and, in 188 cases, from the Parliamentary Paper lists.

Linking the Pension Records to the Parliamentary Papers

In theory everyone who was reported in the Parliamentary Papers as receiving a pension should appear in the Index and have a Pension Application Form. The opposite is not true because most individuals who received a gratuity appear in the Pension records, but not in the Parliamentary Papers, as noted above. The Parliamentary Paper lists do not contain any information about the volume or page number of the pensioner's forms within the POST 1 series and so the ID cannot be used to link them. Consequently, probabilistic record linkage was used to link the majority of the Parliamentary Paper records to the Index. Those who could not be linked in that fashion were linked by hand.

The probabilistic record linkage was done in Python using the package *recordlinkage* which provides a suite of comparative functions as well as the ability to create bespoke matching functions.[25] The populations to be linked are small and should, in theory, be identical.[26] Given this, it was not necessary to use any training data or other more sophisticated methods of record linkage.

Entries from the Parliamentary Paper lists of pensions granted and the Index of Pensions were linked on surname, first initial, occupation and the year the pension was granted. The data was initially indexed on the year granted field using a sorted neighbourhood algorithm. This compares records with the same value for the indexed field but also records close to that value. In this case it compared records in both datasets where the year the pension was granted was the same, but also records where the year granted was +/- 1 year. For example, someone who appeared in the Index as retiring in 1871 was compared to entries in the Parliamentary Paper lists from 1870, 1871 and 1872.[27] The matching on surname and occupation were done using the Jaro-Winkler algorithm included in the *recordlinkage* package with a threshold of 0.85.[28] Initials were matched using a custom function in which matches between initials had to be exact unless one of the initials was a 'J' or 'T', which were frequently confused in the transcription of the Index. This enabled 'J.

[25] https://recordlinkage.readthedocs.io/en/latest/index.html (accessed 30/5/23).

[26] The Parliamentary Paper lists of pensions granted for 1861-1901 report 12,585 individual pensioners; the Index for the same period lists 19,172 people.

[27] After 1881 when the Parliamentary Paper lists begin to cover more than one calendar year, matching and indexing was done using the calendar year that accounted for most of the retirees in that report; for example, individuals listed in the report covering 1st December 1891 to 30th November 1892 were assumed to have retired in 1892. The sorted neighbourhood indexing method is usefully described here https://uwaterloo.ca/networks-lab/blog/post/sorted-neighbourhood-indexing-recordlinkage (accessed 30/5/23). Generally, the Index gave the year the pension application was submitted while the Parliamentary Paper lists provide the year the pension was granted; for most people this was the same calendar year.

[28] Jaro-Winkler is a method for comparing strings of text; it calculates a score between 0 and 1 based on the number of matching characters in the two strings. Jaro-Winkler modified the original Jaro similarity measure to give more weight to differences at the start of the string.

Smith' and 'T. Smith' to be returned as exact matches to account for any mistranscriptions and for errors in the original Parliamentary Paper and Index sources.

The *recordlinkage* package provides a score for every potential match between 0 and 1, where 1 means a definite match and 0 means the two individuals are definitely different. All match scores below 0.5 were dropped and the highest score for each Index individual was exported to a .csv file. Where an individual in the Parliamentary Paper lists or Index was matched to only one individual in the other source, these matches were considered correct. Where individuals in either sources were matched to two or more individuals with the same score the matches were checked by hand and the better match picked.

The matches produced by the first run through of this process were subject to several further checks. The first run sought to match all pensioners from the period 1861-1887. All cases where the surname of the matched individuals was different in the Index and Parliamentary Paper list were checked. For 1861-1887 there were 866 matches where the surname differed, but after checking only 16 of these were uncertain matches, the rest were correct and reflected differences in spelling by the Parliamentary and Postal clerks, or errors in transcription. Every match where the surname was the same in both sources but the overall matching score was low was then checked. For 1861-1887 a low score was any match below 0.617. There were 876 such matches, of which 55 were uncertain matches mostly due to the same individual being recorded with different occupations in each source. The final check used location data, which was always given in the Index but, as noted above, was uncommon in the Parliamentary Paper lists until the 1880s, so it was not used to generate matches. Comparing locations from the two sources identified two further uncertain matches. For all of these 73 uncertain matches, the original records were checked and 17 were found to be incorrect, the remaining 56 matches were correct. The 17 incorrect matches were added to the people requiring hand matching.

The automatic record linkage process dealt with most matching. Remaining individuals were matched by hand. Of the 13,358 individuals in the lists of pensions granted between 1861 and 1903, 12,922 have been matched to a Pension Application Form, representing 96.7 per cent coverage. In nearly all cases the remaining 436 unmatched entries were not actual pension applications. Of these, 176 (40 per cent) related to pensions taken on by the Post Office when the telegraphs were nationalised in 1870; 106 (24 per cent) were cases where the worker died and so no pension was awarded; and 47 (11 per cent) were changes to pre-existing pensions. In only 107 cases has it been impossible to link a Parliamentary Paper listing of an actual retiree to a Pension record. It is likely that the individual in question's form has been transcribed but that the information provided in the list of pensions granted was too vague to link definitively to a record in the database. Consequently, these unlinked pensions granted entries have been excluded from the database to avoid double counting of individuals.

As noted above, the Parliamentary Paper lists included only the people granted pensions, while the Index includes everyone granted a pension and also those individuals granted gratuities. This means that, while a high proportion of individuals in the Parliamentary Papers can be found in the pension records, a smaller share of the individuals in the pension records can be linked but even so, 68 per cent of individuals with a transcribed Pension Application Form have been linked to Parliamentary Paper entries.[29]

Figure 2 shows the proportion of Index entries that are currently unlinked by year. Two years stand out as particularly poor: 1860 and 1882. There is no Parliamentary Paper report on Post Office pensions covering 1882, as noted above, and the high proportion of

[29] Parliamentary Paper lists of pensions granted were not transcribed after the volume covering 1902, 12,921 of the 18,975 individuals in the database retiring between 1860 and 1902 have been linked to a list of pensions granted entry.

Index entries from that year currently unmatched reflects this fact. The Parliamentary Paper report that covers 1860 was not transcribed, which is why that year shows a very high proportion of unmatched entries. The rising proportion of unmatched individuals is because gratuities became more common. A substantial part of these were marriage gratuities paid to women leaving postal employment upon marriage. The percentage of unmatched individuals that were female rises from 2 per cent in the 1860s to 40 per cent in the 1890s. Women made up a growing share of the established workforce in this period, rising from 12 per cent in 1883 to 16 per cent in 1901, while the total number of female established workers rose by nearly 200 per cent in that period, from 5,394 to 15,216.[30]

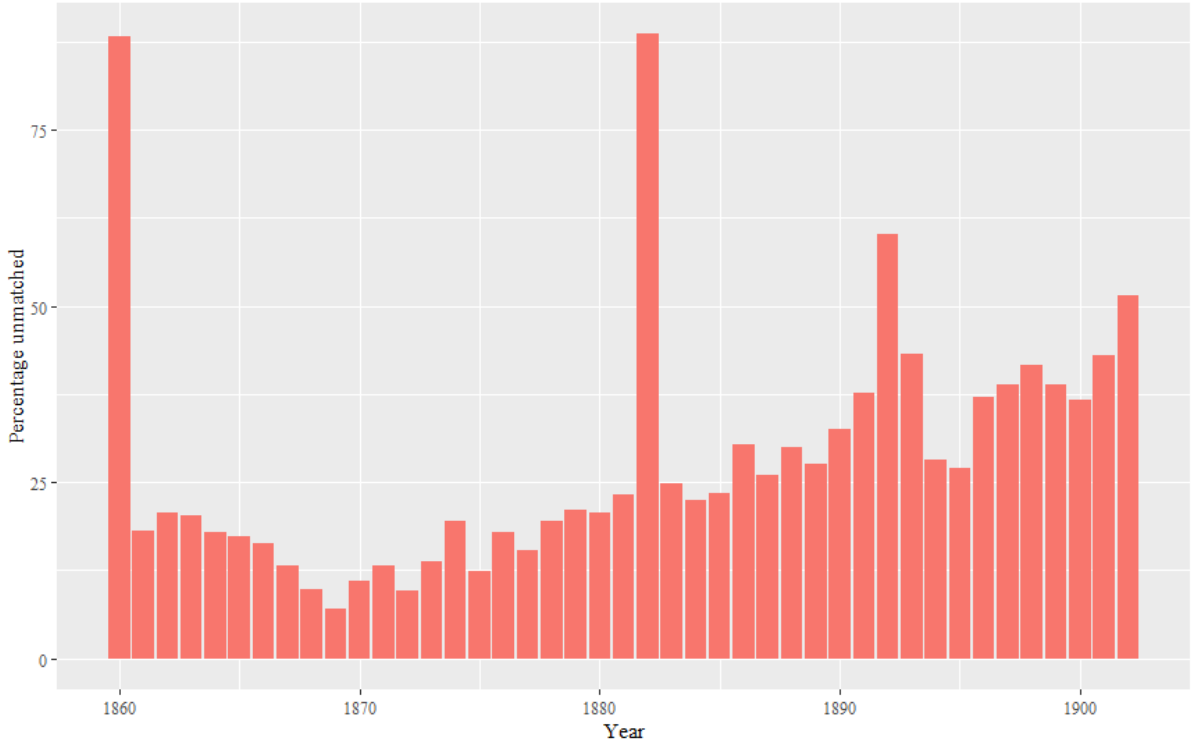

Figure 2. Percentage of Index entries not matched to the Parliamentary Paper Lists of Pensions Granted, 1860-1901.
Source: Addressing Health Pensions Database.

Linking pensions granted and pensions ceased

The lists of pensions ceased published in the Parliamentary Papers provide information on the date of death for most of the postal workers granted a pension in this period. They can be linked to the lists of pensions granted to create a population of individuals for whom we know their date of retirement and date of death and thus the length of their life after retirement.

Much the same information is provided in the lists of pensions granted and ceased and they cover the same people; consequently, like the Index and pensions granted data, they are ideal sources for probabilistic record linkage. However, this process is complicated by the time lag between retirement and death and the fact that the lists of pensions ceased were only

[30] *Twenty-Ninth Report of The Postmaster General on the Post Office*, *Parliamentary Papers*, 22 (1883), 31-2; *Forty-Seventh Report of The Postmaster General on the Post Office*, *Parliamentary Papers*, 18 (1901), 49-50. The total number of established female workers is likely somewhat smaller than these totals as some of the occupational categories in these reports are not broken down by gender and establishment status, all individuals in such categories were counted in these calculations.

published up to 1916, with the final list covering pensions that ceased between 1st December 1913 and 30th November 1914.[31] This means that a substantial proportion of individuals who appeared in the lists of pensions granted never appeared in the lists of pensions ceased. Life expectancy aged 60 for men in England and Wales was fairly stable over our period, falling somewhat between the 1850s and 1890s from 13.53 years to 12.93 before returning to 13.49 in the 1900s.[32] We have undertaken linkage between the lists of pensions granted covering 1861-1901 and the lists of pensions ceased covering the period 1861-1909. The life expectancy figures suggest that we should be able to match most pensioners that retired for reasons of old age between 1861 and 1896 but some will be missed who lived longer than the 12-13 years post retirement. After 1896 our linkage success rate for old-age retirees will steadily decline. The more serious problem, however, is that many pensioners retired before the age of 60 for reasons of ill health; these individuals could have lived substantial periods of time after retirement and thus not appear in the lists of pensions ceased that have been used for linking purposes. Someone retiring aged 30 in 1881 may have lived long enough to only appear in the lists of pensions ceased after 1909, the last year for which we have transcribed data.[33]

A related issue is that this instance of record linkage compares a complete population to an incomplete one, increasing the likelihood of false positives because they are harder to detect. For example, when matching the pension records to the lists of pensions granted all possible matches will be found and can be checked; when matching the lists of pensions ceased and granted the potential matches for any individual in the pensions granted will be incomplete because potential matches may appear in theoretical lists of pensions ceased published after 1909. Thus, a putative John Smith retiring in 1880 will be linked to any John Smiths in the pensions ceased lists, but not to any John Smiths who survived after 1909. This potentially biases the linkage against long-lived individuals and so underestimates survival. This problem is relatively small, however, because of the large number of individuals with unique combinations of initial, surname, occupation and country. Of the 13,358 individuals in the pensions granted lists between 1861 and 1902, 13,125 (98 per cent) have a unique combination and are thus unlikely to be matched to anyone other than themselves in the pensions ceased lists, where the information provided on these variables is usually the same. Nevertheless, these two issues have the tendency to bias the linked sample towards people who a) retired earlier in our period and b) who died relatively soon after retirement. Figure 3 illustrates these issues by showing the percentage of pensioners linked in each year.

The record linkage between the lists of pensions granted and ceased was undertaken in a similar fashion to the linking of the pension granted lists and the Index data described in the previous section. Once again, the *recordlinkage* package was used, but this time year of pension granted could not be used as the indexing field. Instead, surname was used and again the sorted neighbourhood algorithm was used to compare all people with the same or similar surnames.[34] The matching score was calculated on the basis of evaluating matches between

---

[31] *Estimates for revenue departments for the year ending 31 March 1916*, *Parliamentary Papers*, 41 (1914-16), 118-24.

[32] Taken from the 45th, 55th, 65th and 75th *Annual Reports of the Registrar General*. Some of the variation may arise from the methods used to produce the life tables and changing quality of the underlying data, see discussions of the various life tables in E.A. Wrigley and R.S. Schofield, *The Population History of England, 1541-1871: A Reconstruction* (Cambridge, 1981), 708-14; Robert Woods, *The Demography of Victorian England and Wales* (Cambridge, 2000), 179-90.

[33] The life table for 1881-90 produced by the Registrar General gives a life expectancy of 32.52 years for a man in England or Wales aged 30 in 1881, suggesting a death date in 1913, see *Supplement to Registrar-General's Fifty-fifth Annual Report*, *Parliamentary Papers*, 23 (1895), 10.

[34] When indexing text, the neighbourhood approach sorts the text into alphabetical order and then potentially compares individuals whose surname strings match exactly and to strings either side of the exact match in the

initial, surname, occupation and country. Initial and country had to match exactly in order to be considered a good match. Surname and occupation were, as in the Parliamentary Paper-Index matching process above, compared using the Jaro-Winkler algorithm with a threshold of 0.85.

The part of the project primarily interested in tracing the death dates of workers focusses on those who retired in the census years 1861, 1871, 1881, 1891 and 1901. However, this record linkage process was run on the lists of pensions granted and ceased for the entire period in order to match as many pensioners as possible. Of the 12,909 pensioners reported in the Parliamentary Paper lists for 1860-1902, 6,678 were linked to the lists of pensions ceased (52 per cent). As with the pension record-Parliamentary Paper linked data, various checks were undertaken to ensure the accuracy of these matches: death dates and retirement dates were compared to ensure no one was matched to someone who died before they retired. As noted above, further hand matching was undertaken for those pensioners who retired in census years to attempt to obtain complete data on the death dates of those retirees.

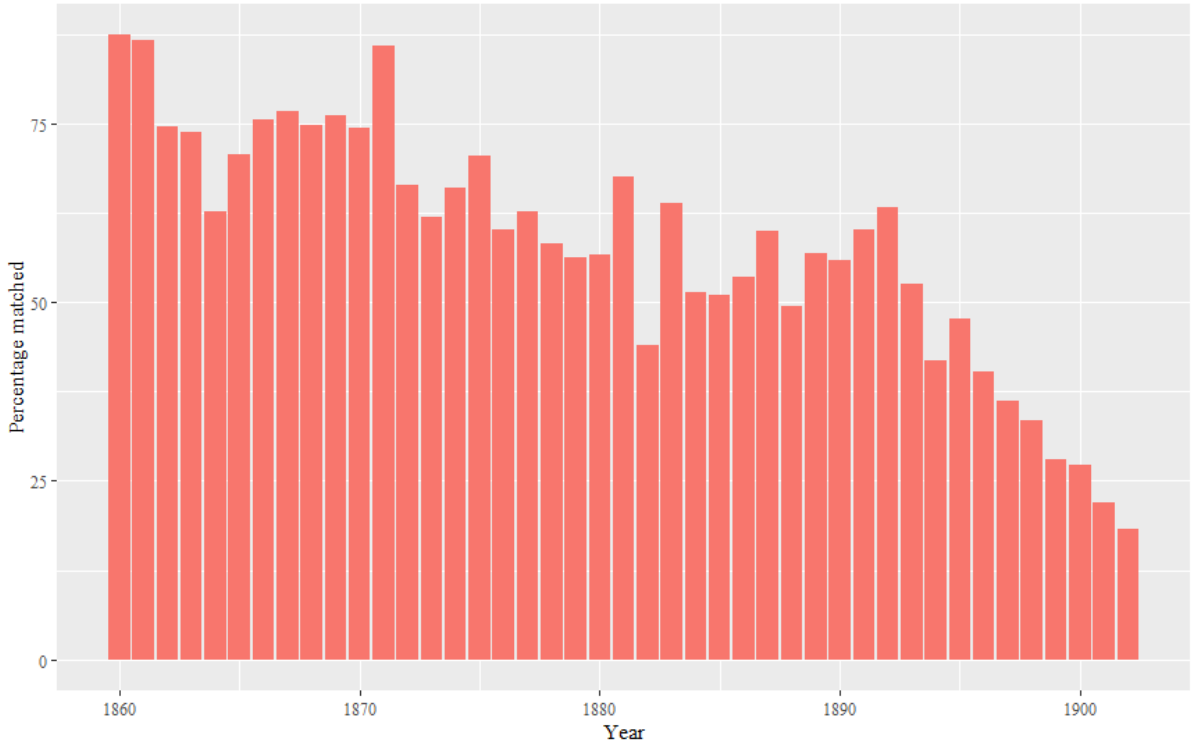

Figure 3. Percentage of individuals listed in the Parliamentary Paper Pensions Granted lists matched to the Pensions Ceased lists, 1860-1902.
Source: Addressing Health Pensions Database.

Death certificates

Death certificates were ordered for the pensioners that retired in each census year. The transcriptions of these were undertaken by volunteers before being hand checked. As the dates of death were obtained from the Parliamentary Paper lists of pensions ceased, in most cases the resultant transcribed data could be linked directly to the Parliamentary Paper data, and through the links between the pensions ceased and granted to the pension records, these data can be linked to the rest of the data obtained from the TPM records and included in the main database. The pensioners for each census year were identified using the lists of pensions

ordered list, the number checked either side depends on the window chosen, in this case a window of 9 was used so each individual could be matched to a surname up to 4 places before or after their exact name.

granted, consequently, even where a pensioner was not identifiable in the pensions ceased lists, they could still be linked to the Index and Pension Application Form data through the matches between the pensions granted and Index data described above.[35] These death certificates also provided dates of death for 521 individuals not successfully found in the lists of pensions ceased.

These four matching exercises allowed the data on a single individual that appeared in the Pension Application Forms, the Index to Pensions, the Parliamentary Paper lists of pensions granted, the Parliamentary Paper lists of pensions ceased, and the death certificates to be linked together and hence the creation of a database containing as much information as possible on each pensioner who retired between 1860 and 1908. Not every individual could be found in each source, such as gratuitants in the Parliamentary Paper lists, so some individuals have null values for the variables derived from those sources.

Some of the variables are given in more than one of the sources. For example, the Index, Parliamentary Papers and Pension Application Forms all provide the pensioner's occupation. The Pensions database includes all different versions of these variables, indicating which source each has been drawn from. It also includes a set of variables indicating which of the options we have decided is the 'primary' version, but users can use any of the other options if they disagree with our choice. In these cases, we have prioritised specificity by picking the longest version of any given variable. For example, if a pensioner's first name was given as 'A' in the Index, 'Andrew J' in the Parliamentary Papers and 'Andrew John' in the Pension Application Form, the primary first name variable in the database would use 'Andrew John'.

Coded and inferred variables

The data extracted from the sources discussed above forms the core of the Addressing Health Pensions database. However, much of the information provided in those sources requires additional coding to be usable for historical analysis. Additional variables can be calculated from the values provided in the original sources, while some need to be inferred. This section describes the various codes and categories added to organise the raw data. It also covers the inferred and calculated variables added to the transcribed information described above.

The data contain three key variables which required coding before they can be easily used: occupation, cause of retirement and cause of death, and location. In each case there are thousands of unique values in these fields and so they have to be coded and categorised before they can be easily used for analysis.

Occupation

The Post Office as an organization contained a wide range of different occupations. It employed not only workers that directly sorted and delivered the post, but also considerable numbers of clerks, engineers and others involved in construction and production, medical personnel, and people involved in transportation. Many of these occupations included workers of different grades, adding further complication to the occupational structure of the

[35] The process of obtaining death certificates is described in more detail in Laura Newman and Harry Smith, 'Linking Post Office Superannuants to Death Records: Sources and Methodology', Addressing Health Working Paper 2 (2023).

postal workforce.[36] The Pensions Database contains 2,373 unique occupation strings which had to be cleaned, standardized and then classified.

The first stage was to clean the unique strings, many of which were the same occupation but with slight variations in spelling, capitalization or punctuation. This produced a smaller number of unique strings, which were then standardized. This mainly involved removing place names and expanding the considerable number of abbreviations used in the sources. These standardized occupations were then classified into twelve categories. The categories were based on the kind of work done by the occupation rather than seniority or Civil Service grade. Thus, all clerks, regardless of their grade, were included in a single category. A function-based classification was chosen because of the project's focus on occupational health. Organising the individuals by the kind of work undertaken would allow us to examine how outdoor work differed from indoor work, how managers fared compared to other employees and so on. An alternative classification is available, however, one based on the wages paid to each pensioner. These data are available for all pensioners from the Parliamentary Paper lists of pensions granted.

Cause of retirement and cause of death[37]

The database contains over 4,349 unique causes of retirement and 1,094 unique causes of death. The process for coding them was similar to that described above for occupations. First, the original strings were cleaned and standardized and then they were coded. All causes were coded to ICD10h, an adapted version of ICD10, a version of the long-standing international classification of diseases that was in use in over 100 countries until the start of 2022 when it was replaced by ICD11.[38] The history of the ICD stretches back to 1893 when Jacques Bertillon presented a classification of the causes of death that he proposed be adopted for international use. It was revised every ten years at meetings held in Paris until its stewardship was taken over by the World Health Organisation in 1948.[39] ICD10, and now ICD11, aimed to allow countries to record mortality and morbidity data in a consistent way to enable comparisons across time and space. ICD10h aims to do much the same for historical causes of death. This scheme codes individual words and phrases rather than diseases. It attempts to group different terms for the same phenomenon under a single broad code, while also not over-interpreting a given term. This is done by expanding the existing ICD10 coding system to allow historical terms to be related to contemporary ones while remaining distinct. Thus, for example, the various conditions 'enteric fever', 'bilious fever' and so on which are thought to be historical synonyms for typhoid fever, in ICD10h they are all coded to the same block of codes as 'typhoid fever' in ICD10 with additional 2 digits added to the standard code to distinguish all these potential synonyms. This method allows all potential cases of typhoid to be grouped together, but also allows the evolution of different terms for typhoid to be

[36] For works examining the postal workforce see Alan Clinton, *Post Office Workers: A Trade Union and Social History* (London, 1984); Martin Daunton, *Royal Mail: The Post Office Since 1840* (London, 1985), 191-268; Helen Glew, 'Women's Employment in the General Post Office, 1914-1939' (University of London, Institute of Historical Research, PhD thesis, 2009). A guide to common postal occupations can be found at https://addressinghealth.org.uk/resources/occupations/ (accessed 60/5/23) our thanks to Alex Obradovic for his work in preparing this invaluable guide for the project.

[37] This process of coding and classification is described in more detail in Harry Smith, 'Classifying Cause of Retirement in Historical Pension Records', Addressing Health Working Paper 3 (2023).

[38] Angélique Janssens, 'Constructing SHiP and an International Historical Coding System for Causes of Death', *Historical Life Course Studies*, 10, 64-70.

[39] Alastair H.T. Robb-Smith, 'A History of the College's Nomenclature of Diseases: its reception', *Journal of the Royal College of Physicians*, 4/1 (1969), 16; Iwao M. Moriyama, Ruth M. Loy and Alastair H.T. Robb-Smith, *History of the Statistical Classification of Disease and Causes of Death*, eds. Harry M. Rosenburg and Donna L. Hoyert (Hyattsville, MD: 2011), 10-22.

tracked.[40] Once coded the causes of death can then be classified into categories depending on the purpose of the study. The use of ICD10h allows ready comparison between our data and other historical or contemporary studies that have used ICD10 or ICD10h.

The method used was as follows. First, the causes were checked and cleaned. Each unique cause was then split into component parts. For example, 'chronic rheumatism and general debility' contains two individual causes which need two codes. This was done by splitting all causes based on the presence of a number of conjunctions and symbols.[41] The unique bits produced by this process ('chronic rheumatism', 'general debility' and so on) were then coded by hand, following the same principles as the SHiP network in producing ICD10h: individual terms rather than 'diseases' were coded; causation was not inferred; and historical terms were placed in additional codes associated with the modern terminology.

Much of the coding was straightforward, especially given the large number of unspecific terms, such as 'disease of the heart', which can only be coded to the unspecified disease codes in ICD10h, I51.900 in the case of 'heart disease, unspecified'. Where the coding was more difficult was in cases of terms whose meaning has changed over the last 150 years, or which have dropped out of medical usage. In those cases, contemporary medical texts were used to infer the meaning of terms and map that meaning onto the ICD10h coding scheme. For example, 'chronic rheumatism' clearly referred to joint pain without heart involvement in our period, and so can be coded to M79.000 ('rheumatism, unspecified'); whereas 'acute rheumatism' was used by nineteenth-century doctors to describe joint pain with heart involvement and so was coded to I00.000 within the heart disease chapter.

The data contain a considerable number of causes of retirement and death which mention more than one medical cause. In these cases of co-morbidity/mortality, the order of the causes was kept. Thus, 'chronic rheumatism and general debility' has two codes: the first (the variable ICD10h_1 in the Pensions Database) is M79.000, the code for 'chronic rheumatism' and the second (ICD10h_2) is R53.003, the code for 'general debility'. This rule was not followed in two sets of cases. First, where an element of causation is mentioned, thus in the case of 'Ulcerated feet caused by chronic eczema', the code for 'chronic eczema' is given as the primary cause. Secondly, where the first mentioned cause was generic and unspecific (such as 'general debility') the second cause was taken as the primary cause to minimise the number of people with generic primary causes.

Having coded every medical cause to ICD10h it was necessary to classify those codes into a set of categories to allow aggregate analysis.[42] One advantage of using ICD10h is that, once coded, causes can be immediately aggregated to the ICD10 chapters. These chapters derive, ultimately, from Jacques Bertillon's 1893 classification which mainly categorised conditions by the part of the body affected, with separate categories for injuries and notable infectious diseases.[43] The categories have changed somewhat over time, as more general disease categories have been added: cancers, autoimmune diseases and psychiatric conditions for example, all of which was part of the gradual inclusion of aetiology into the classification.[44] This scheme currently has 22 categories covering infectious diseases, cancers, diseases associated with different anatomy parts (the nervous, respiratory and circulatory systems for example), conditions related to pregnancy and childbirth, mental and behavioural disorders, blood diseases, metabolic diseases, injuries and a range of categories

[40] Janssens, 'Constructing SHiP', 10, 68-9.
[41] The following were used 'and', 'also', 'plus', as well as the symbols '&', ';' and ','. Further splitting was required during the hand coding process.
[42] Janssens, 'Constructing SHiP', 69.
[43] Moriyama et al., *Statistical Classification*, 11-12.
[44] Ibid., 15-22.

for other and ill-defined conditions.[45] Two changes were made to this scheme. First, chapters 19, 20 and 21 have been combined since these all relate to external causes of ill health – injuries, accidents, encounters with health services. Second, tuberculosis, the most common non-generic cause, is such an important of cause of retirement in our data that it has been separated from the rest of infectious diseases in its own category.

Location

There are 3,377 unique place names in the database. Each Pension Application Form, Index entry and many Parliamentary Paper entries contained the name of the Post Office that the pensioner worked at prior to retirement. It is important to remember that these data are for the last place of work, not residence. As with occupations and causes of retirement, these had to be cleaned and standardized initially. They were then geo-coded. This process assigned latitude and longitude co-ordinates to each location. For England, Wales and Scotland the initial coding was done using the GB1900 Gazetteer obtained from the *Vision of Britain* project.[46] This contains co-ordinates for over 2.5 million places names in Great Britain. For locations not included in the Gazetteer or where the match between the Gazetteer and the database location name was ambiguous, the location was hand-coded. For Irish locations, the number of unique locations was relatively small, just over 300, so they were all geo-coded by hand.

Once all locations were coded to latitude and longitude co-ordinates they were then coded to a series of other geographic units. All locations were coded to a country and a county using shapefiles of English, Welsh, Irish and Scottish counties and the QGIS join attributes by location tool. Each location was also coded to the administrative unit in the used by the Registrar General in each country to publish demographic statistics. For English and Welsh locations, each place was coded to a Registration District and Registration Sub-District, for Scotland they were coded to the parish, and for Ireland, to the Poor Law Union.[47]

In order to identify urban places, the location names were compared to three sources: the lists of Parliamentary Burghs in Scotland, the lists of Irish towns published in the Irish Census reports; and the list of towns in England and Wales with populations over 10,000 used by Smith et al.[48] For these locations composite units were used when calculating

---

[45] The full list can be found here https://icd.who.int/browse10/2019/en (accessed 13/7/2022).

[46] *GB1900 Gazetteer (Abridged)*, produced by the Great Britain Historical GIS Project at the University of Portsmouth, the GB1900 partners and volunteers, https://www.visionofbritain.org.uk/data/.

[47] Again shapefiles of these respective units were used to do this coding: thanks are due to Joe Day for the Registration Sub District files for England and Wales: *Registration Sub-District Boundaries for England and Wales, 1851-1911* (2016); to Mike Anderson and Corrine Roughley for parish shapefile for Scotland: *Scotland's Parish Populations: Parish Boundaries, 1755-1891* (2019) deposited at the National Records of Scotland; the Irish Poor Law Union shapefiles were taken from Ian Gregory and Paul Ell, *Irish Poor Law Union and Barony Boundaries, 1841-1871* (2004) [data collection], UK Data Service, SN: 4999, http://doi.org/10.5255/UKDA-SN-4999-1.

[48] *Census of Scotland, 1861, Population tables and report, Parliamentary Papers*, 50 (1862), 152-3; *Census of Scotland, 1871, Eighth decennial census of the population of Scotland taken 3rd April 1871, with report*, Vol. 1, *Parliamentary Papers*, 68 (1872), 160-1; *Census of Scotland, 1881, Ninth decennial census of the population of Scotland taken 4th April 1881, with report, Vol. 1, Parliamentary Papers*, 76 (1882), 166-7; *Census of Scotland, 1891, Tenth decennial census of the population of Scotland taken 5th April 1891, with report, Vol. 1, Parliamentary Papers*, 94 (1892), 172-3; *Census of Scotland, 1901, Parliamentary Burghs, districts of burghs and counties in Scotland, Parliamentary Papers*, 129 (1902), 2-3; *Census of Ireland, 1861, Part V, General Report, Parliamentary Papers*, 61 (1863), 130-31, 256-7, 364-5, 436-7; *Census of Ireland, 1871, Part III, General Report, with illustrative maps and diagrams, summary tables, and appendix, Parliamentary Papers*, 81 (1876), 30-31; *Census of Ireland, 1881, Part II, General Report, with illustrative maps and diagrams, tables, and appendix, Parliamentary Papers*, 76 (1882), 221-3; *Census of Ireland, 1891, Part III, General Report with illustrative maps and diagrams, tables, and appendix, Parliamentary Papers*, 90 (1892), 327-9; *Census of Ireland, 1901, Part III, General Report, with illustrative maps and diagrams, tables, and appendix*,

population or other contextual variables. For example, a location in Birmingham was given the population density of the city as a whole rather than just the values for the registration sub district in which the Birmingham Post Office was located.[49] Currently, all London pensioners are assigned to London, but it will be possible in future to provide a more precise geographical reference for them once each London post office has been located and geo-coded.

Having associated every unique location with a set of co-ordinates and with the appropriate administrative geography, various contextual information can be calculated and associated with locations, which can then be directly linked to individual pensioners. For example, census data was used to calculate population density values for each location for each census year. Each pensioner was then allocated the population density derived from the closest census year. For example, someone retiring in 1875 had the 1871 population density for their location assigned to them, while a worker who retired in 1876 was assigned the 1881 population density. This process can be repeated with any other contextual data (mortality rates, rateable value and so on). An urban code was then generated based on population density. All individuals in London were assigned to 'London'; all individuals in towns with populations over 10,000 in that census year were coded as 'Urban'; those in locations with a population density greater than 0.3 people per acre were coded as 'Semi-Urban'; and the remainder as 'Rural.

Some individuals in the database did not work within the United Kingdom: some were Post Office Agents in locations such as Cairo or Buenos Aires, others were working on telegraph ships and others were postal workers of various kinds in overseas post offices in Malta and elsewhere. These individuals have no additional geographical information beyond their location.

Gender

Neither the Pension Application Form, nor the Index, nor the Parliamentary Paper lists of pensions granted or ceased contained a field listing a pensioner's gender. This information is included in the answer to question 18 on the form, which asks whether the applicant has performed their duties 'with diligence and fidelity'. The answer to this question nearly always includes the pronouns 'his' or 'her'. However, this information was not transcribed and it is not feasible to extract it by hand for all 26,500 pensioners.

As a consequence, it has been necessary to infer an individual's gender from other information provided in these sources. For some women, the various sources noted that they were either married or soon to be married. Sometimes this was indicated by the fact that their maiden name was given as well as their married surname, sometimes their honorific is given ('Miss' or 'Mrs'), and sometimes it is noted that they were 'resigned with a view to marriage'. In those cases, assigning an inferred gender is relatively straightforward. This was also possible in a small number of male cases where 'Mr' or 'Sir' was written. In the majority of cases, however, pensioners have been assigned a gender on the basis of their first name.

The use of first names to predict the gender of individuals appearing in large datasets that lack explicit self-reported gender information is well established. In order to predict pensioners' gender we adapt a method and developed by Blevins and Mullen, who use

---

Parliamentary Papers, 129 (1902), 178-81, the inclusion of towns in these Irish census reports varied over time, where a town present in other years was missing in a given year the area and population was taken from the townland lists in the relevant census volume; Harry Smith and Robert J. Bennett, 'Urban-Rural Classification using Census Data, 1851-1911', Working Paper 6, ESRC Project ES/M0010953, 'Drivers of Entrepreneurship and Small Businesses', https://doi.org/10.17863/CAM.15763; Harry Smith, Robert J. Bennett and Dragana Radicic, 'Towns in Victorian England and Wales: a new classification', *Urban History*, 45/4 (2018), 568-94.

[49] The General Post Office in Birmingham was located in the St Martin's Registration Sub District.

contemporary and historical data on gender and first names to establish the probability that any given first name was male or female.[50] This method is undoubtedly crude: it relies on state-gathered information and is reliant on a binary definition of gender, but it works well for nineteenth-century British data and provides a measure of certainty in each case.[51] We have taken their method, but used the 1881 and 1911 British censuses to generate for each unique combination of forename and birth year the probability that the name in question referred to a man or a woman. For example, 435 of 462 people called 'Jessie' born in 1851 and alive in 1881 were female in that year's census, so everyone with that name and birth year is assigned the gender 'female'. In contrast, 298 of 324 born in the same year called 'Jesse' were male, so everyone with that name is coded as male. These percentages can be used to judge the likelihood or otherwise of the inferred gender being correct.

Not every combination of forename and birth year in the Pensions Database appeared in the 1881 and 1911 census. In those cases, the original forms were checked and the answer to question 18 used to assign a gender. Once this had been done a small number of people remained without a gender, individuals who had no known first name, only an initial, or whose form was missing or incomplete. In those few cases gender has been assigned on the basis of occupation using the same method as with forename. Occupations in the Post Office were highly segregated by gender as figure 4 shows. These three methods allow all individuals to be assigned a gender, albeit with varying degrees of certainty.

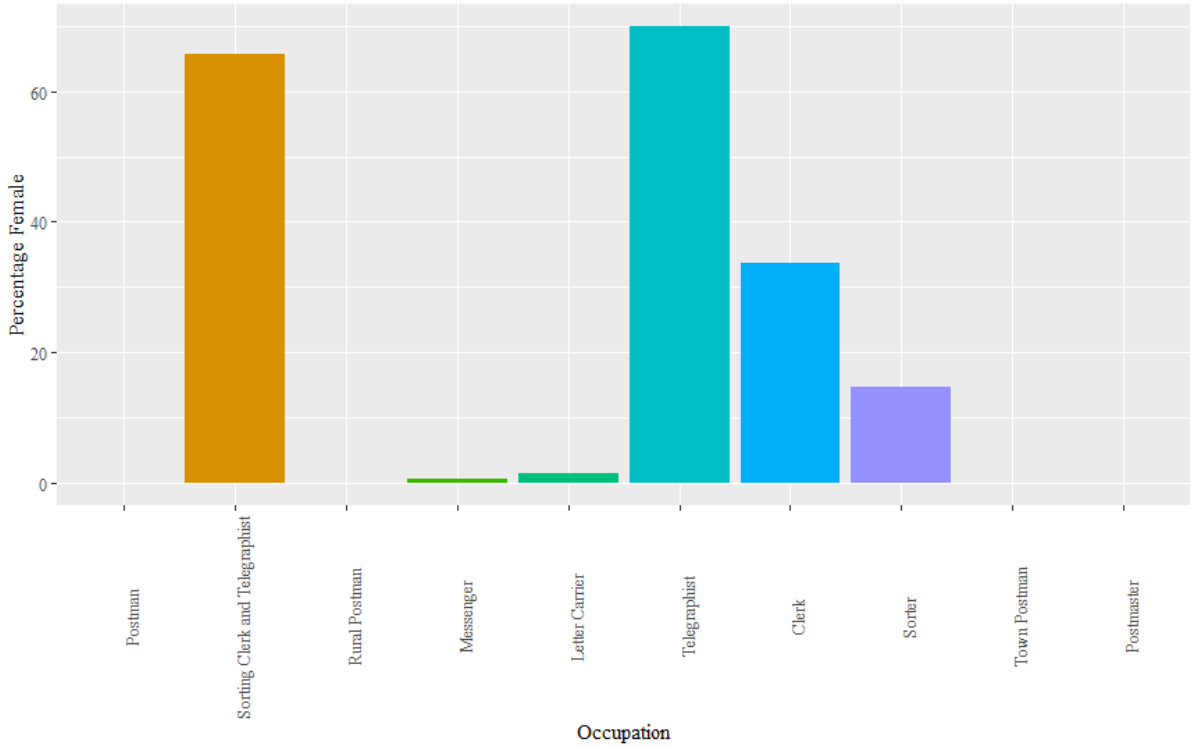

Figure 4. Ten most common postal pensioner occupations, percentage female, 1860-1908.

[50] Cameron Blevins and Lincoln Mullen, 'Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction', *Digital Humanities Quarterly*, 9/3 (2015), http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html. The package used it Lincoln Mullen, 'gender: Predict Gender from Names using Historical Data' (2021), R package version 0.5.4.1000, https://github.com/ropensci/gender.

[51] These problems are discussed in Helena Mihaljević, Marco Tullney, Lucia Santamaría and Christian Steinfeldt, 'Reflections on Gender Analyses of Bibliographic Corpora', *Frontiers in Big Data*, 2/29 (2019), https://www.frontiersin.org/articles/10.3389/fdata.2019.00029/full.

Source: Addressing Health Pensions Database.
Note: The ten most common occupations include 73 per cent of all postal pensioners in the database.

Sickness and Date of Retirement

The data on sick leave and other leave taken provided in the table included in the Pension Application Form requires some additional processing before it can be used. The data in these tables was not reported in years to the day of retirement, but instead in calendar years. For example, an individual who worked for the Post Office for ten years and who retired on the 21st January 1899 has sickness data related to 9 full years and 1 partial year (1st to 21st January 1899). In order to use the final year of data, therefore, it is necessary to use the date of retirement to calculate the share of the final year that the individual in question worked.

The Pension Application Forms do not include a simple 'retirement date' but instead provide three dates: the date at which the individual stopped performing their duties, the date when the individual stopped being paid, and the date when the form was submitted to the Treasury. It is clear from the sickness days reported in the final year of the table that the day at which an individual stopped being paid was the date at which the Post Office stopped counting sick days and so, for our purpose here, the date of retirement. For example, Frederick James Gibson retired in 1902. In that year he took 137 days of sick leave and his salary was paid until the 17th May 1902, which is 137 days into the calendar year.[52] The key date for our purposes then is the date at which salaries stopped being paid and, consequently, sick days stopped being counted.

For most of the pensioners in the database this date was readily available. However, some workers submitted forms while their salary was still being paid and thus no date of cessation of salary was given. For others, the page of the form where this information was provided was missing. In these cases, the date of the submission of the form to the Treasury was used as the retirement date. Unfortunately, in some cases the form does not specify a date of submission, but this can be estimated as each form is bound into a volume which covers a number of months' worth of correspondence between the Post Office and the Treasury. Therefore, for individuals with no specified date of submission it was assumed that they retired on the 28th day of the month in the middle of time period covered by that volume.[53] This adds a degree of imprecision to the date of retirement, but it is necessary in order to make the final year of sickness data usable by researchers. It has no effect on the data in the years prior to the year of retirement, so the final year can be ignored if the assumptions made here are felt to be unreasonable.

Having established a date of retirement for each pensioner, the share of the final year worked can be calculated and this can be used to establish two measures. First, an accurate value for the mean sick days taken by each worker that takes into account the fact that the final year was likely not complete. Second, the share of each year taken as sick leave can be calculated, to allow sickness in the year of retirement to be compared to previous years.

In some cases, there was a note on the final line of the sickness table which notes that sick leave has been taken 'since' a certain date. These days off have to be calculated and then added to the total in the final year of work. This was possible once a date of retirement was established. Where the date of retirement was after the final year in the sickness table, this added an additional year of data to the table. For example, William Bowman retired on the

[52] TPM, POST 1/323, 110, pension form for Frederick James Gibson, 1902.
[53] The volumes vary in the amount of time they cover, in the 1860s and 1870s volumes usually cover four months, in the 1880s and 1890s two months, and in the late 1890s and 1900s one month. For example, POST 1/119 covers January to April 1865 so someone with an unknown date of submission would be inferred a retirement date of 28th April 1865.

21st March 1896; however, the last year in his sickness table was 1895 which stated he had taken 37 days off sick that year and, additionally, that he had been off sick 'since 21 September'. Thus, he took 182 days off sick, from the 21st September 1895 to the 21st March the next year. A new year was, therefore, added to the sickness table, 1896, and the 182 days sick leave split between those two years, giving a total of 139 days off in 1895 and 80 days off in 1896.

Establishing the date of retirement, therefore, allowed the sickness table to be enhanced in two ways. First, the share of the final year covered by the sickness table can be calculated and therefore that year's sickness data can be compared to the rest of the table. Secondly, additional, unspecified, days off sick can be calculated and added to the existing data ensuring a complete record of each pensioner's sick leave is available.

Death Dates and longevity

The main source of information on the date of death of pensioners in the database is list of pensions ceased discussed above. However, this has been augmented with additional death dates from a digital version of the English and Welsh civil registration death registers. This was searched for any pensioners retiring 1860 and 1902.[54] Where the name, birth year and registration district of work and death matched, it was assumed that these were the correct death date for the worker in question.[55] This provided dates of death for an additional 1,278 pensioners. In total death dates are known for 8,340 pensioners and for those individuals post-retirement longevity has been calculated by subtracting their year of retirement from their year of death.

Conclusion

This working paper has detailed the creation of the Addressing Health Pensions Database. The data has been collected from four sources: the Pension Application Forms, the Index to Pension Application Forms, Parliamentary Paper lists of pensions granted and ceased, and death certificates. In each case the transcriptions have been subjected to considerable automatic and hand checking before being combined to create individual records for 26,500 pensioners and gratuitants who retired from postal work between 1860 and 1908. Information is provided on each pensioner's name, age, length of service, occupation at retirement, place of work at retirement, cause of retirement, sickness record for up to ten years before retirement and, for a substantial subset, date of death and longevity.

[54] Thanks to Neil Cummins for providing these civil registration data.

[55] These assumptions are fairly parsimonious and if the location restriction were dropped a number of additional pensioners could be found in the death registers, but the confidence in the matches would drop.

Appendix A: Parliamentary Papers

This appendix contains a list of all the Parliamentary reports used in the construction of the pensioners database.

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1861, *Parliamentary Papers*, 31 (1862).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1862, *Parliamentary Papers*, 31 (1863).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1863, *Parliamentary Papers*, 34 (1864).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1864, *Parliamentary Papers*, 31 (1865).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1865, *Parliamentary Papers*, 40 (1866).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1866, *Parliamentary Papers*, 40 (1867).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1867, *Parliamentary Papers*, 41 (1867-8).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1868, *Parliamentary Papers*, 34 (1868-9).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1869, *Parliamentary Papers*, 41 (1870).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1870, *Parliamentary Papers*, 37 (1871).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1871, *Parliamentary Papers*, 36 (1872).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1872, *Parliamentary Papers*, 39 (1873).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1873, *Parliamentary Papers*, 35 (1874).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1874, *Parliamentary Papers*, 42 (1875).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1875, *Parliamentary Papers*, 42 (1876).

An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1876, *Parliamentary Papers*, 49 (1877).
An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1877, *Parliamentary Papers*, 46 (1878).
An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1878, *Parliamentary Papers*, 42 (1878-9).
An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1879, *Parliamentary Papers*, 40 (1880).
An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1880, *Parliamentary Papers*, 57 (1881).
An Account of all allowances or compensations granted as retired allowances or superannuations in all public offices or departments, which remained payable on the 1st January 1881, *Parliamentary Papers*, 37 (1882).
Estimates for Revenue Departments for the year ending 31 March 1885, *Parliamentary Papers*, 52 (1884).
Estimates for Revenue Departments for the year ending March 1886, *Parliamentary Papers*, 50 (1884-5).
Estimates for Revenue Departments for the year ending March 1887, *Parliamentary Papers*, 43 (1886).
Estimates for Revenue Departments for the year ending March 1888, *Parliamentary Papers*, 54 (1887).
Estimates for Revenue Departments for the year ending 31 March 1889, *Parliamentary Papers*, 70 (1888).
Estimates for Revenue Departments for the year ending 31 March 1890, *Parliamentary Papers*, 52 (1889).
Estimates for Revenue Departments for the year ending 31 March 1891, *Parliamentary Papers*, 46 (1890).
Estimates for Revenue Departments for the year ending 31 March 1892, *Parliamentary Papers*, 53 (1890-91).
Estimates for Revenue Departments for the year ending 31 March 1893, *Parliamentary Papers*, 53 (1892).
Estimates for Revenue Departments for the year ending 31 March 1894, *Parliamentary Papers*, 56 (1893-4).
Estimates for Revenue Departments for the year ending 31 March 1895, *Parliamentary Papers*, 55 (1894).
Estimates for Revenue Departments for the year ending 31 March 1896, *Parliamentary Papers*, 66 (1895).
Estimates for Revenue Departments for the year ending 31 March 1897, *Parliamentary Papers*, 55 (1896).
Estimates for Revenue Departments for the year ending 31 March 1898, *Parliamentary Papers*, 57 (1897).
Estimates for Revenue Departments for the year ending 31 March 1899, *Parliamentary Papers*, 57 (1898).
Estimates for Revenue Departments for the year ending 31 March 1900, *Parliamentary Papers*, 56 (1899).

Estimates for Revenue Departments for the year ending 31 March 1901, *Parliamentary Papers*, 52 (1900).
Estimates for Revenue Departments for the year ending 31 March 1902, *Parliamentary Papers*, 43 (1901).
Estimates for Revenue Departments for the year ending 31 March 1903, *Parliamentary Papers*, 62 (1902).
Estimates for Revenue Departments for the year ending 31 March 1904, *Parliamentary Papers*, 41 (1903).
Estimates for Revenue Departments for the year ending 31 March 1905, *Parliamentary Papers*, 54 (1904).
Estimates for Revenue Departments for the year ending 31 March 1906, *Parliamentary Papers*, 49 (1905).
Estimates for Revenue Departments for the year ending 31 March 1907, *Parliamentary Papers*, 71 (1906).
Estimates for Revenue Departments for the year ending 31 March 1908, *Parliamentary Papers*, 51 (1907).
Estimates for Revenue Departments for the year ending 31 March 1909, *Parliamentary Papers*, 66 (1908).
Estimates for Revenue Departments for the year ending 31 March 1910, *Parliamentary Papers*, 55 (1909).