Statistics New Zealand
Tatauranga Aotearoa

Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project

New Zealand Government

Liability
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Citation
Statistics New Zealand (2014). *Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project.* Available from www.stats.govt.nz.

ISBN 978-0-478-42903-9 (online)

Published in June 2014 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4610
www.stats.govt.nz

Contents

List of tables and figures

Tables by chapter

Figures by chapter

Abstract

Integrated Data Infrastructure (IDI) allows for statistical outputs and research on the transitions and outcomes of people through education, the labour market, benefits, justice, health and safety, migration, and business data.

The IDI is primarily based on administrative data and also contains a number of surveys undertaken by Statistics NZ and other agencies. Some of the datasets can be merged directly on common unique identifiers, and this is straightforward to do. However, other datasets do not have common unique identifiers and these can be linked by creating links using demographic information. These datasets are linked using record linkage techniques, and uses the software IBM QualityStage v8.5.

This report outlines the probabilistic record linkage used within the Integrated Data Infrastructure system and demonstrates with specific examples.

Key words

Integrated Data Infrastructure, record linkage

1 Introduction

Purpose

This report describes the method of record linkage used in the Statistics New Zealand Integrated Data Infrastructure, gives practical examples, and outlines some of the wider issues of using the linking information in analysis.

Integrated Data Infrastructure

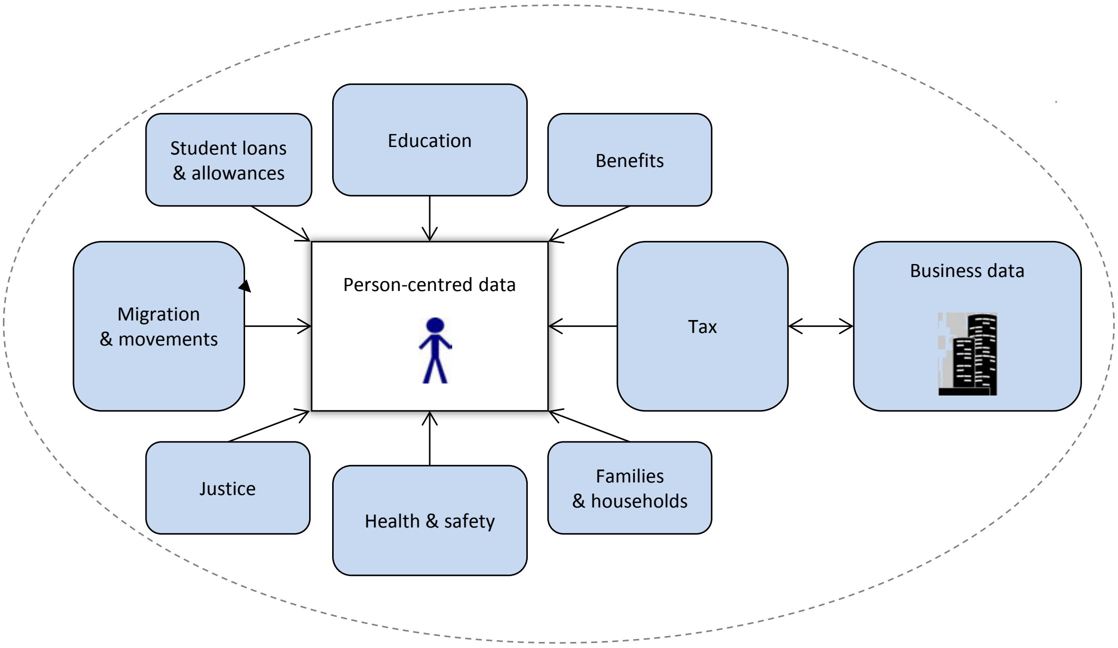
Integrated Data Infrastructure (IDI) allows for statistical outputs and research on the transitions and outcomes of people through education, the labour market, benefits, justice, health and safety, migration, and business data. The longitudinal aspect of the IDI covers an extended range of pathways and transitions information, to allow for policy evaluation and research analysis, and the production of statistical outputs.

The IDI is primarily based on administrative data and also contains a number of surveys undertaken by Statistics NZ and other agencies, as shown in figure 1.

While figure 1 shows the current state of the IDI, the data sourced from agencies will be expanded, and more datasets from other agencies will be included.

Figure 1

Integrated Data Infrastructure subjects



Data included in the IDI:

- person and business tax data, Student Loans and Allowances data – Inland Revenue (IR)
- benefit data, Student Loans and Allowances data – Ministry of Social Development (MSD)

- secondary school achievement data, tertiary education data – Ministry of Education (MOE)
- sentencing data – Department of Corrections
- injury data – Accident Compensation Corporation (ACC)
- migration and movements data – Ministry of Business, Innovation and Employment (MBIE)
- departure and arrival cards – New Zealand Customs Service / Statistics NZ
- Household Labour Force Survey (HLFS) data – Statistics NZ
- New Zealand Income Survey (NZIS) data – Statistics NZ
- Survey of Family Income and Employment (SoFIE) data – Statistics NZ
- Longitudinal Immigration Survey of New Zealand (LISNZ) data – Statistics NZ
- Longitudinal Business Database (LBD) data – Statistics NZ.

See appendix 1 for more detail on the contents of these datasets.

Benefits

The IDI:

- enables analysts and researchers to create additional statistics and research, to tell a richer story of people and businesses in New Zealand with no extra respondent burden
- offers enhanced research opportunities for developing and costing social and economic policy, as well as programme evaluation, particularly in relation to migrant outcomes, the justice sector, the education system, the benefit system, and the injury sector
- improves access to additional information sources related to justice, injury, and migration, encouraging wider use and reuse of data across government
- meets current user demand for integrated research data to allow finer-grained statistical analyses and the investigation of currently unanswered questions.

Challenges

While the IDI has all of these datasets, there is no common personal identifier across them. This is by legislation, where the Privacy Act 1993 ensures that no identifier used by one agency will be used by another agency as their identifier. This means that each agency has its own unique identifier, and the coverage and quality of this identifier can vary. Some agencies can record another agency’s identifier, mainly the Inland Revenue tax file number, for administrative purposes, and this can be used where available, but for internal purposes these agencies use their own identifiers.

Some agencies have their data at a person level, although most have data at an event level. For example, the Ministry of Education has a record for every course a student takes, and the ACC data has a record for every injury and treatment a person receives.

With more datasets being added over time, what is needed is a simple general approach that can be applied each time. When a new dataset is added, typically it is linked to the spine of people created by the customer and client lists within the Inland Revenue’s tax data. The datasets are added on one at a time to the spine, or can be linked to other datasets (see chapter 4 for details on the linking situation).

Each of these pairwise linkings is done as a separate project, and hence the word ‘project’ will be used to refer to a specific instance of linking. For example, the IR-MOE project is linking the Inland Revenue data with the Ministry of Education data.

In some cases, the link is direct. For example, the business information in the Longitudinal Business Database can be linked directly to the Inland Revenue tax data via the employer’s IRD number.

In other cases, there is no direct link, although there are people in common. For example, people who studied and are now working are in the Ministry of Education and Inland Revenue datasets.

This linking is done by record linkage. The general record linkage methodology is described in chapter 2. The specific implementation of how record linkage is used within IDI is in chapter 3. Chapter 4 provides examples of the record linkage for specific projects. Chapter 5 provides a wrap up discussion.

All rates given in the report refer to the state of the IDI dataset after the December 2013 update.

2 Overview of record linkage methodology

The aim in record linkage is to link together two records from different datasets that belong to the same person or entity.

In probabilistic record linkage, records from one dataset are compared with records from another dataset. Each variable that is common to both datasets is compared to provide evidence about whether the two records belong to the same person. An overall weight is assigned to the record pair that reflects the probabilities that the records are from the same person. The higher the weight, the higher the probability.

Weight calculation

Consider two datasets, *A* and *B*. For a record *a* from *A* and *b* from *B*, compare them on *K* characteristics. For each characteristic k, $m_k = Pr(a_k = b_k | a\ and\ b\ are\ a\ match)$ and $u_k = Pr(a_k = b_k | a\ and\ b\ are\ not\ a\ match)$. The m_k is a measure of how reliable the variable is, usually measured as 1 – error rate. The u_k takes into account how rare the value is. With these values, the probability methodology takes into account error rate and rarity.

In practice, m_k is provided by the user, and is not calculated by the software. This can be measured either from data knowledge, from linking a sample of the data, or linking the data with an approximate value and iterating to a more refined value. The u_k value is observed directly from the data, and some linkage programs calculate this for the user.

A distance function is used to compare a_k with b_k. Usually, where a_k and b_k are strings, this is a function similar to the Jaro-Winkler distance measure (Winkler, 1999). Where the distance function has these values as close enough, they are said to be in agreement. For some variables, such as sex, day, month, or year of birth, this function is simplified to being in agreement when the values are the same. The functions available are those available in the software used (see below). Other distance functions are possible, although cannot currently be implemented.

When the variables are in agreement $w_k = \log_2 \frac{m_k}{u_k}$ is assigned (any base can be used, base 2 is used due to origins of the work using information theory), with $\frac{m_k}{u_k}$ representing the odds ratio of the agreement. Otherwise $w_k = \log_2 \frac{1-m_k}{1-u_k}$ is assigned, with $\frac{1-m_k}{1-u_k}$ representing the odds ratio of the disagreement. w_k is the weight for that specific characteristic. Note that when there is disagreement, the value of w_k is negative.

The overall weight for record pair (*a*,*b*) is then $w = \sum_{k \in K} w_k$ (this is follows Fellegi and Sunter (1969)). As the ratios have been converted by $\log_2$, adding the weights is the same as multiplying the odds ratios for all the variables k. The record pairs with overall weights above a threshold are assigned as links.

For a full technical description, refer to Fellegi and Sunter (1969). See also Herzog et al (2007) and Christen (2012) for recent coverage and developments.

Link comparison

Comparing every record to every other record would be time consuming. Sectioning data into blocks, and comparing records if they are in the same block is more efficient.

A *blocking* variable is one in which every record must have the same values before they will be compared. That is if you block on date of birth then two records will only be compared if their dates of birth are identical. A *block* is a subsection of data where the blocking variables take on specific values. For example, date of birth and sex might be blocking variables, and males born on 13/12/1959 would be one block, and females born on 03/07/1987 would be another.

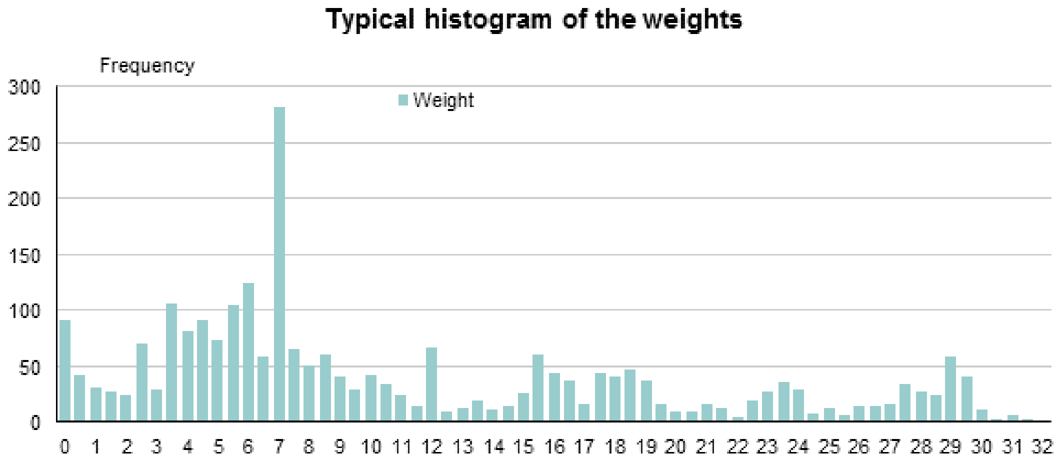
Records from the two datasets are compared using *linking* variables. The values of linking variables for a pair of records are compared to see the level of agreement between them.

As defined above, the *weight* is calculated from two pieces of information: how reliable the data is (m_k), and how common the value is (u_k). Once the weights have been calculated, a histogram of the weights is obtained to observe the frequency of the weights. This helps to determine the threshold or *cut-off* as shown in figure 2. Note that the cut-off in this case was chosen after a brief clerical review of the links.

Every record pair with a weight above the threshold is designated as a link. Every record pair with a weight below the threshold is designated as a non-link. Statistics NZ uses the term ‘link’ instead of ‘match’, as ‘match’ refers to the true nature of a record pair, whereas ‘link’ refers to the assessed nature of the record pair.

A *pass* is a combination of blocking and linking variables and a cut-off. Each linking project is a series of passes that produces a collection of linked records.

Figure 2



Source: Statistics New Zealand

Link quality

The resulting set of links may still contain errors. The *recall rate* measures how successful the linking project is at finding all the matches. This is measured by the number of matches identified, over the number of matches available. Getting a precise estimate is hard, as determining the number of possible matches is not a simple process. If a record didn’t find a link, is that because the process failed, or because there was no link to find? While we may know in principle what the dataset covers, there will be records that are difficult to classify.

The *precision rate* measures how accurately the linking process separated correct links from incorrect links. This is measured by measuring the number of correct links over all links made. The number of incorrect links is usually measured by clerical reviews of samples of the links.

There is a trade-off between the two rates, whereby increasing one results in decreasing the other. In the IDI projects, the priority is on obtaining a high precision rate. This is to allow varied analysis that will use as little erroneously linked data as possible.

Each project is ideally a one-to-one link, where each record on one side links to at most one record on the other side. *Duplicates* are records which link to more than one record and the links are above the cut-off. How these are handled in the IDI depends on the projects, see chapter 6 for more details.

Linkage bias examines variables where the linking was less successful. For example bias in year of birth is expected. Older people would be recorded earlier in a dataset's history, and generally data quality improves over time. Hence older people would have lower quality data, creating a linking bias. However it is not easy to look at linked records versus records that didn't link. The problem is that the populations are not well defined and when looking at records that did not link, there is a mixture of records that were true non-matches and records that were not linked. This mix hampers looking at the characteristics of the records that do not link. This bias is not examined in the examples given here.

Software used

The program used in the linkage is IBM QualityStage v8.5.

This program provides the software needed to perform probabilistic record linkage, according to the Fellegi and Sunter methodology. Each project can be built pass by pass, and the software generates its own precise u_k values. The comparison functions are based on Jaro-Winkler and allow for a lot of flexibility. Also, the software is easily integrated into our existing IT infrastructure.

Other packages were tested at the beginning of the IDI development, and QualityStage was found to have the most flexibility and speed for our requirements.

3 Details of record linkage methodology

The general form of the passes used in the IDI is a mixture of unique identifier linking and then probabilistic linking.

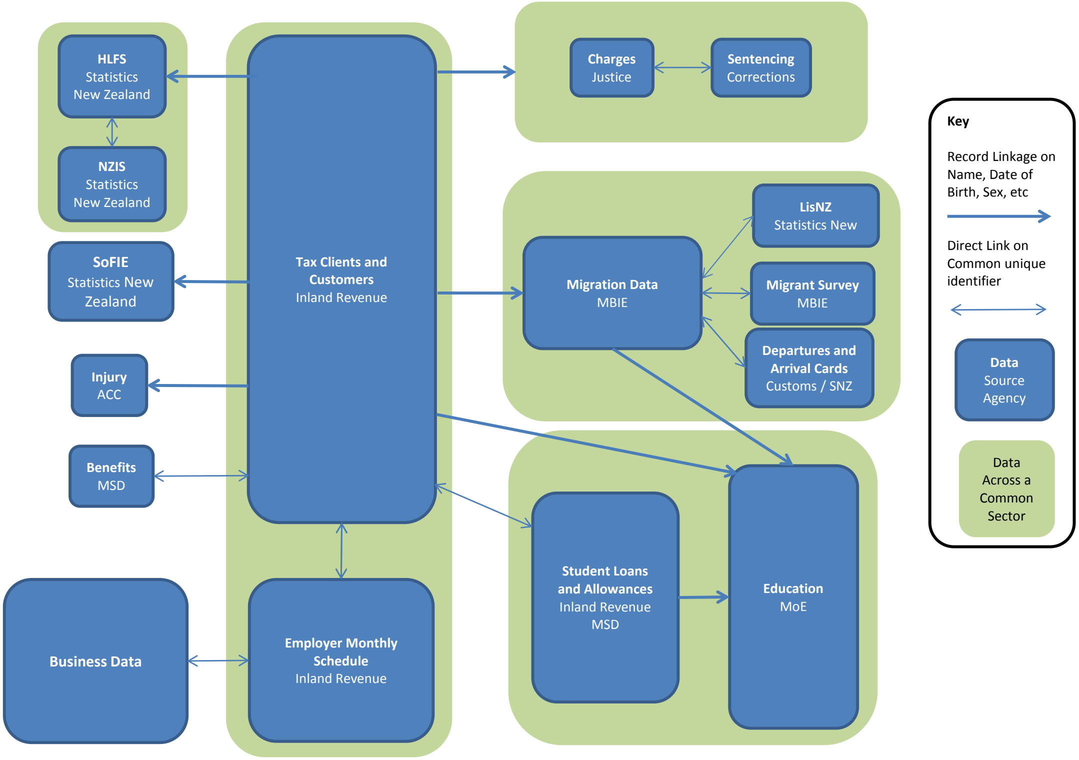
When creating probabilistic linking, there is often some type of research analysis the linking will aid, usually having a particular population to ensure is linked or some property of the data to preserve. In the case of the IDI, analysis could be quite varied, and the general aim is to ensure the created links are as correct, within reason, as possible, that is to say, to obtain a high precision rate.

Linking projects in the IDI

Figure 3 shows the complete links between all current projects, with the heavy lines indicating which projects are linked using probabilistic linking. The light lines are where the datasets have been merged exactly using the unique identifiers. See appendix 1 for details on these datasets.

Figure 3

Linking between datasets in the IDI



Note that the main linking is to the spine formed by the client and customer details in the Inland Revenue data. This is because this has the largest coverage of the population of interest, namely New Zealand residents. Some other linking is done between datasets for specific sub-populations of interest.

All the pairwise linkings done are referred to as projects. Specific examples are given in section 4. The datasets are converted to person level first where possible, aside from the Student Loans and Allowances to Education link, which is explained in chapter 4.

Recall that each project is a series of passes, with blocking and linking variables. Record pairs not formed in the first pass are available for linking in the second pass, and so on. The passes are set up so that the smallest blocks are linked first, then other passes allow for missing or incorrect information.

The variables available across all projects are:

- first names
- last names
- sex
- year of birth, month of birth, day of birth.

Beyond that, many of the projects have some form of unique identifier available which is specific to the source agency. Some projects have only a limited version of first and last names available.

Unique identifier linking

Typically when there is a unique identifier available on both datasets, these datasets are linked using that unique identifier. In these projects, the unique identifier is not always present for all records in the dataset, or can have quality issues. However, it is available for at least a part of some of the datasets in the IDI, and this forms the first pass in the linking.

The unique identifier can be used to link directly, however the general quality is such that more evidence is required before the link is established. The unique identifier is used as a first pass blocking variable, with the other variables as the linking variables.

Probabilistic linking

For the other probabilistic passes, the blocks are formed from basic demographic information, with all other variables being used as linking variables.

As date of birth is highly specific, this forms the first block. As date of birth may be missing or incorrect, a second pass is used. The Soundex of the name forms the second block (Soundex is a phonetic encoding algorithm that converts names to strings to deal with different spellings (National Archives 2007). For example, ‘Smith’ and ‘Smythe’ both encode to S530).

There are potentially other passes, depending on the data and how some sub-populations work.

Outline of the passes

The outline of the passes is given in table 1. This is the approach to the passes used as a starting point for all the linking projects. In many of the projects, the actual nature of the data means that these passes need to be modified, but this is the standard starting point.

Table 1

Outline of the passes

Pass number	Blocking	Linking
1	Unique identifier	Year of birth, month of birth, day of birth, first name, last name, sex.
2	Date of birth	First name, last name, sex, unique identifier.
3	Soundex of first name, Soundex of last name.	First name, last name, sex, year of birth, month of birth, day of birth, unique identifier.

All possible variables are used in each pass, with different blocking variables. A unique identifier can still help create links by providing evidence when some information is missing, or providing negative evidence when other variables are close.

When variables are missing, the variable weight for the linking comparison is zero. When a blocking variable is missing, then the record is not in any block for comparison.

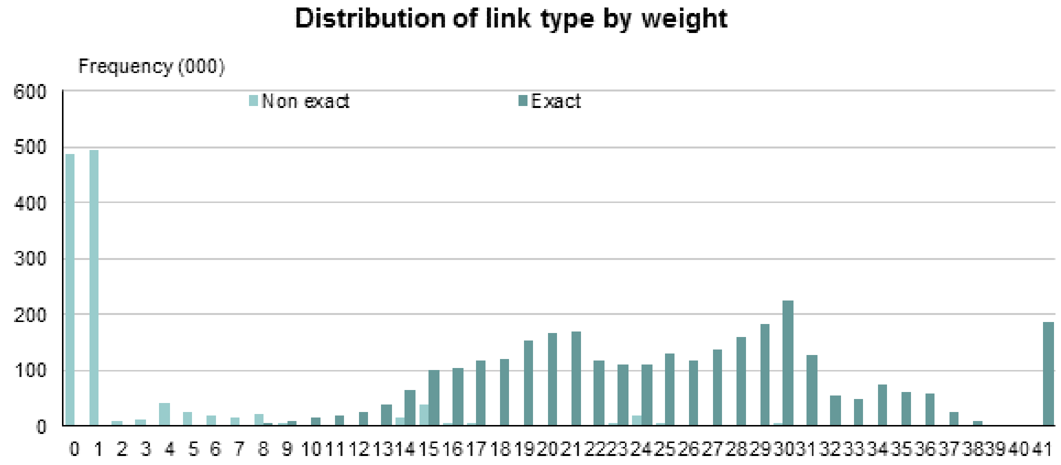
Selecting the cut-off

In each pass, a single cut-off is chosen so that all record pairs with that weight or higher are taken to be links, and all other record pairs are non-links. Cut-off levels are determined through experimentation.

Each record pair can be assessed as a ‘near exact’ or ‘non-exact’ link. Near exact links are ones in which the demographic information is in agreement, within some small tolerance for errors. Non-exact links are all other links. For example, a record pair in which first and last name were in agreement, but date of birth was incorrect (or missing), would be a ‘non-exact’ link.

After each pass is run, the graph of near exact and non-exact links is produced. See figure 4 for an example. This is from the IR-MOE project (see the next section). The weights are rounded down to the nearest whole number for convenience.

Figure 4



Source: Statistics New Zealand

From this graph, possible cut-off points can be examined to determine where to put the cut-off, usually after a spike of non-exact links. The non-exacts are informally reviewed to check if they are allowable or not. For this example, the cut-off was placed at 16.

Quality of links

The two measures of quality are the precision rate (the rate at which correct links are found from all links), and recall rate (the rate at which correct links are found out of all potential correct links). In all the projects, the focus is on getting a high precision rate, ie minimising creating erroneous links with the trade-off that more correct links may be missed. As mentioned, there is no specific analysis to design the linking for, so the general approach of minimising false positives, ie a high precision rate, was used.

This means that the created links can be trusted, although there is the possibility that some sub-populations might be under-represented.

The precision rate is estimated by sampling the links. As mentioned above, the links are divided into ‘near exact’ and ‘non-exact’ links. Near exact links are ones in which the demographic information is in agreement, within some small tolerance for errors. Non-exact links are all other links. The near exact links are assumed to have no incorrect links, which has been verified by clerical reviews of samples. The non-exact links are used to estimate the precision rate, and as this is a smaller population this is more amenable to manual clerical review of appropriately sized samples, where the largest sample (200 links) is from the largest pass, and smaller samples are taken from the other passes (100 links from other probabilistic passes, and 50 from unique identifier passes). Given the size of the non-exact links, this is enough for a suitable estimate.

The recall rate is harder to estimate, and all projects worked to get as many records as possible while keeping with a probabilistic linking methodology approach. Several extra passes were tried in each project; however no successful extra sets of links were created (without incurring many incorrect links). Investigation of the unlinked records did not find useful sets of records to link. This is not the same as a bias analysis (which is not covered in this paper), as it is possible that some sub-populations are under-represented.

Measuring the recall rate is difficult, as it is unclear if an unlinked record does have a companion record. It is possible that the person isn’t on the other dataset, because they either do not fit the population of the other dataset or they weren’t captured correctly. And some records are missing so much information that they cannot be assessed at all.

For this paper, the recall rate will not be given. Instead a different measure, the ‘overlap rate’ will be reported, and it will be made clear how this is defined and what assumptions are made. In some cases, this overlap rate will be representative of the recall rate. In all cases, the overlap rate can be no more than the actual unknown recall rate. Measuring the recall rate properly and easily is an area of open research.

In the examples in the next section, the precision rate and the overlap rate are reported.

Duplicates, transitivity, and multiple linkages

The methodology focuses on pairwise linking of datasets. However, dealing with the links created is more complicated.

For some of the projects within IDI, it is important to not have any duplicates, as the linking is assumed to be one-to-one. While some links might be valid duplicates of people who are listed more than once, in other cases they could easily be erroneous links. In those projects, duplicates are resolved by the rule that earliest pass and highest weight wins. This eliminates the duplicates in the final dataset. For example, this is used in the MBIE-IR and IR-HLFS projects.

It would be good to eliminate duplicates in the original datasets first, but either this is not a priority for the agencies or not possible for them to carry out without linking their data to another source, or it doesn’t impact on their administrative functions. In the IDI projects, the duplicate issue is small; around 5 percent of links might be duplicates, and only the MOE-SLA and IR-MOE projects deal with allowing duplicates. In those projects, once all the links have been combined, duplicates are handled by combining multiple MOE student records under one unique identifier.

With three datasets, there is the possibility of linking record a from dataset A to record b from dataset B, and record a to record c from dataset C. By the way the links are created this means that record b is linked with record c. However, there is nothing enforcing this in the overall linking process, as there may be no linking directly between dataset B and dataset C. Any linking is inferred via dataset A.

In some cases, datasets B and C are directly linked as well as through dataset A. This can lead to situations where record b now links to record c2. For example, the MBIE-MOE project provides links that could potentially contradict the IR-MOE or MBIE-IR linking. This is dealt with by prioritising the datasets, so that if the links between B and C (MBIE-MOE) contradict previously established links, then those new links are ignored. This is how consistency is maintained. The linking of international students in MBIE-MOE is considered an add-on to the links, and has lower priority. All linking is done before final links are processed, so the MBIE-MOE cannot just be the unlinked students.

The other situation is when one project might link dataset A with dataset B, and another project might link a sub-population of dataset A with dataset B. For example, the MOE-SLA project links a sub-population of the IR-MOE project. In this case, a priority is again established. As the MOE-SLA links have better quality information, any link in IR-MOE that contradicts these links is ignored.

Central linking concordance

After the links have been processed for duplicates and other issues, they are loaded into a new concordance table. This table contains one row per person (note that it is possible for people to have more than one row where their records have not been linked properly, see below), and in each row are the unique identifiers for each dataset they are located in. This table forms the basis for enabling analysis in the IDI.

For example, one row might contain the person's IR identifier, their MOE identifier, and their MBIE identifier. Another row might contain only their MOE identifier if that person was not linked to any other dataset.

Every person in all datasets in the IDI is listed in the central linking concordance table, but only their identifiers are listed. No other information is stored in this table.

Incorporating linkage errors in analysis

Linkage errors can cause problems in analysing the relationships between variables across the linked files. Given the high precision rates and expected high recall rates in the IDI projects, it is expected that these sources of errors would be small. Currently, information around the pass structures and weights are used in some analyses to check results using only links with high weights.

The question of how to incorporate linkage errors in analysis is an open area for research.

For tabular output, this is generally ignored. However, if a cell level estimate of the precision and recall rates could be estimated, then a quality measure could be provided.

For more advanced analysis (eg linear regression), this problem has been more extensively examined. Lahiri and Larsen (2005) show that not correcting errors can lead to regressions having understated relationships, and present an alternative method for correcting for this. More recently, using synthetic data, Kim and Chambers (2012) presented another method for correcting for linkage bias, particularly for the case of linking survey data to administrative datasets. Redoing the Kim and Chambers work, using the IDI data, is in development. The first part of this is to determine if the IDI needs to have this correction given the link rates.

Another approach is to create weights based from the linking information, and produce weighted output that corrects for the linkage errors. This is used, for example, in the New Zealand Census Mortality Study (see Tan et al (2010)).

4 Examples of record linkage

These are specific cases of using the approach for particular projects, which demonstrates how the approach is used and how it is modified where knowledge of the data is relevant.

The MBIE-IR and IR-MOE projects are both examples of linking to the tax data spine. The MOE-SLA project is an example of a particular subpopulation of interest having better quality linking information available. The MBIE-MOE project focuses on getting international students, and completes a triangle of linking between MBIE, MOE, and IR (see figure 3). The final project, IR-HLFS, demonstrates linking survey data to the tax spine.

Almost all of the datasets have been converted to person level datasets, with event data removed, for use in the linking. The MOE-SLA project has the datasets at the level of enrolment year, where a person can have more than one record, one for each year they were enrolled and/or had a student allowance or loan.

In each project below, the passes are described, first the unique identifier passes then the general probabilistic passes. After each table, the quality of the links are summarised with the precision and overlap rates. See the chapter 3 for the definitions of these rates.

Example 1: MBIE-IR

This project integrates Ministry of Business, Innovation and Employment migration and movements data with Inland Revenue tax data.

Unique identifier pass: The MBIE data does not contain an IRD tax file number. A unique identifier on MBIE is passport number. Some people can have more than one passport number. As the dataset is designed to be, as far as possible, one row per person, the various passport numbers are combined into one field.

On the Inland Revenue data, only a sub-population has passport number, available for people who use a passport to validate their identity on the Inland Revenue dataset. This means that passport number is only available for 4.7 percent of the people who have a name and/or a birth date.

As mentioned above, there can be more than one passport number per person, and the passport number is not used to block. Instead, two unique identifier passes are used.

The first pass blocks on date of birth and matches on passport number, looking for an exact match in any of the available passport numbers. The second pass uses Soundex of first and last names, and again matches on any available passport number. There are not many records without date of birth, however the second pass picks up some records and these are easy links to form.

Probabilistic pass: the standard passes are used here, as defined above.

Table 2

Outline of the passes for MBIE-IR

Pass number	Blocking	Linking
1	Date of birth	Passport number
2	Soundex of last and first names, sex.	Passport number
3	Date of birth	First name, last name, sex, passport number.
4	Soundex of first name, Soundex of last name.	First name, last name, sex, year of birth, month of birth, day of birth, passport number.

The precision rate is 99.7 percent and the overlap rate is 45 percent. The overlap is measured simply as the number of links over the total number of potential MBIE records. It is likely that many of those records will not have a companion record on IR because many temporary visitors captured in the migration dataset will not interact with the tax system, but this exact population cannot be defined.

Example 2: IR-MOE

This project integrates Inland Revenue tax data with Ministry of Education data.

Unique identifier pass: Both datasets have the common identifier of tax file number. However, only around 27 percent of the student population has a tax file number recorded (this is voluntary information and not verified at source).

Probabilistic passes: The two general passes are included. However, there is a large section of student information (from an earlier time period when less was recorded) where full name is not recorded. Instead only the last four letters of their surname, and sometimes the first letter of their first name, is noted. As they would likely not be in the previous passes, a separate pass was added to pick them up.

Table 3

Outline of the passes for IR-MOE

Pass number	Blocking	Linking
1	IRD identifier	Year of birth, month of birth, day of birth, first name, last name, sex, truncated last name, first name initials.
2	Date of birth	First name, last name, sex, IRD identifier.
3	Soundex of first name, Soundex of last name.	First name, last name, sex, year of birth, month of birth, day of birth, IRD identifier.
4	Truncated last name, date of birth.	Sex, first name initials, IRD identifier.

The precision rate is 98.4 percent and the overlap rate is 78 percent. The overlap is measured as the number of students registered on MOE that are linked. It is possible for some students, such as international students, not to be registered with IR.

Example 3: MOE-SLA

This project integrates Ministry of Education student data with Ministry of Social Development loans and allowances data. There is more data available, in particular provider and student level information, to enable a better linking than with just the basic demographic information, but only for students who have either an allowance or a loan.

Unique identifier pass: The basic unique identifier is a combination of provider code and student identifier number. Students can have more than one provider code and student identifier over time. The data is at the enrolment year level as students are allowed one borrowing a year, which may be to different providers over time.

Around 27 percent of the students have a tax file number on both datasets, and this can also be used as a unique identifier. The tax file number is not used as a blocking variable because the links that would be created are picked up in other passes, and this field has low coverage.

Probabilistic pass: the standard passes are used here, as defined above. Sex is added to help reduce the block sizes.

Due to the nature of the data, another pass is added because only truncated last name and first name initials are available for some students.

Table 4

Outline of the passes for MOE-SLA

Pass number	Blocking	Linking
1	Provider code, student identifier, year of birth.	Day of birth, month of birth, first name, last name, IRD identifier.
2	Date of birth, sex.	First name, last name, IRD identifier, provider code, student identifier
3	Soundex of first name, Soundex of last name, sex.	Year of birth, month of birth, day of birth, first name, last name, IRD identifier, provider code, student identifier.
4	Date of birth, sex.	First name, last name, sex, truncated last name, first name initials, IRD identifier, provider code, student identifier.

The precision rate is 99.1 percent and the overlap rate is 92 percent. The overlap is measured as the number of students registered on SLA that are linked. All students on SLA should be on MOE, although manual review of a sample found records that could not be found on MOE.

Example 4: MBIE-MOE

This project integrates Ministry of Business, Innovation and Employment migration and movements data with Ministry of Education student data. This project is targeted at linking international students, who might not be registered with Inland Revenue. All students are included after that specific linking for a chance to pick up any extra links missed in the previous projects.

Unique identifier pass: There is no common unique identifier, so this pass is not used.

Probabilistic pass: This linkage is split into two groups. The first group reduces the data on each side to international students, which can be identified in each dataset. The second group is all people. The two linking approaches are the same, with different populations. Sex was added to the blocking variables in the later passes to improve blocking size.

The two probabilistic passes are used, and a third is added because the only name information people might have is a truncated last name, and maybe a first initial.

Table 5

Outline of the passes for MBIE-MOE

Pass number	Blocking	Linking
1	Date of birth, International Student Indicator.	First name, last name, sex.
2	Soundex of first name, Soundex of last name, International Student Indicator.	First name, last name, year of birth, month of birth, day of birth, sex
3	Date of birth, last name truncated, International Student Indicator.	First initials, sex.
4	Date of birth, sex.	First name, last name.
5	Soundex of first name, Soundex of last name	First name, last name, year of birth, month of birth, day of birth, sex.
6	Date of birth, sex.	First initials, last name truncated.

The precision rate is 99.6 percent and the overlap rate is 78 percent. The overlap is measured as the number of international students registered on MOE that are linked. While it is likely that other students, not indicated as international, linked, this is the population of interest.

Example 5: IR-HLFS

This project integrates Inland Revenue tax data with the Statistics NZ Household Labour Force Survey (HLFS).

In the case of this data, the HLFS has no common unique identifier with the Inland Revenue data, so the unique identifier pass could not be used at all. The standard probabilistic passes were used.

One important note is that name on the HLFS is provided in one field, which is standardise into first names and last names. The quality of this variable is quite variable, and around 10 percent or more of adults can be missing a name entirely, and many only have a first name. This makes linking problematic for many records.

In the future, other variables, such as address, will be considered for incorporating to improve the link.

Table 6

Outline of the passes for IR-HLFS

Pass number	Blocking	Linking
1	Date of birth	First name, last name, sex.
2	Soundex of first name, Soundex of last name.	First name, last name, sex, year of birth, month of birth, day of birth.

The precision rate is 96 percent and the overlap rate is 75 percent. The overlap is measured as the number of adults on HLFS that are linked. It is possible that some adults will not have an Inland Revenue record. While children may also have linked, the adults are the population of interest.

5 Conclusion

The Integrated Data Infrastructure contains administrative and survey sourced data. The intention is to link the data at a person level. For the purposes of linking, the data is restructured to be at a person level where possible.

Where the data cannot be directly linked on common identifiers, the linking is based on probabilistic linking. This is a widely used international methodology, and provides robust results. The aim of the linking in the IDI is to minimise creating erroneous links, that is, to maximise the precision rate.

The implementation we use utilises the fact that there are unique identifiers for parts of the data. This information is used first, and then demographic information is used to form the passes. A general approach is given as the first step for any linkage project.

There are only a few variables available; first and last names, sex, and date of birth. However, the linking is still of good quality, perhaps because New Zealand is a small country (roughly 4.5 million people) meaning that name and date of birth is more frequently unique than in other countries.

In the five examples presented, the general approach is modified in each case because of the nature of the data used. While the general approach is followed, knowledge of the data means that the passes can be more precisely targeted to increase the ability of the linkage project to find links.

The quality of the linking depends a lot on the input quality. While the administrative data sources are of good quality, the survey data is less so meaning that the linking involving the Household Labour Force Survey is of lower quality.

The result of these projects are high precision rates. The recall rates are less well measured, but there are no areas of significant bias found. How the issue of duplicates is dealt with is tailored for each project, and care is maintained to ensure consistent links.

The IDI has a good foundation on which we can build robust analysis of government data to answer a wide range of important research, policy, and evaluation questions.

References

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.

Kim, G & Chambers, R (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis, 56 (9)*, 2756-2770.

Lahiri, P & Larsen, M (2005). Regression analysis with linked data. *Journal of the American Statistical Association 100 (469)*: 222-230.

Fellegi, I & Sunter, A (1969). A theory for record linkage. *Journal of the American Statistical Association 64 (328)*: 1183–1210.

Herzog, T, Scheuren, F & Winkler, W (2007). Data quality and record linkage techniques. Springer.

National Archives (2007). The Soundex indexing system. Available from: http://www.archives.gov/research/census/soundex.html

Statistics New Zealand (2013). Integrated Data Infrastructure. Available from: www.stats.govt.nz

Tan, L, Blakely, T, & Atkinson, J (2010). Record linkage of census and mortality 2001–06 records: Unlock Ratios 2005–06 and linkage weights 2001–06. *New Zealand Census Mortality Study Technical Report No.7*.

Winkler, W (1999). The state of record linkage and current research problems. Bureau of Census Research Report, RR99/04.

Appendix 1: Source agencies' datasets

Inland Revenue

Inland Revenue plays a critical role in improving the economic and social wellbeing of New Zealanders. Inland Revenue collects 85 percent of the Crown's revenue as well as collecting and disbursing social support programme payments and providing the government with policy advice.

Website: http://www.ird.govt.nz/

Inland Revenue tax data

Employment and earning data; tax able earnings and deductions; customers' file and client names.

Data includes:

- personal identifiers and details; ie IR number, name, title, date of birth, resident indicator (New Zealand or overseas), address
- income and tax details; ie taxable earnings, earnings not liable, tax deductions (including PAYE, withholding tax, family support tax credit, student loan amount)
- employer and employment details; ie employer IRD number, employment start and end dates.

Inland Revenue Data for Student Loans and Allowances

This covers all student loan borrowers plus allowance recipients since the start of the Student Loan Scheme in 1992.

Data includes:

- personal identifiers and details; ie IRD number, name, title, date of birth, address, residency status
- enrolment details (for where a loan or allowance was received); ie education provider
- student loan details; ie loan registration start and end dates, principal transferred from StudyLink, interest transferred, interest eligible for write-off, loan balance, interest compounded, interest written off
- income and loan repayment details; ie gross earnings (from each employer if more than one) including salary/wages, benefits, New Zealand Superannuation, and withholding payments, as well as tax paid, employer industry code, employer's IRD number, and employment stop date.

Student Loans Account Manager

This covers all student loan borrowers plus allowance recipients since the start of the Student Loan Scheme in 1992.

Data includes:

- personal identifiers and details; ie IRD number, name, date of birth, sex, ethnicity, address postcode
- enrolment details (for where a loan was received); ie provider, student ID number, attendance type (ie full time/part time)

- student loan details; ie student loan payments (including the amounts borrowed for fees, course-related costs, and living costs), interest and repayments, and the loan start and end date.

Ministry of Social Development

This Ministry helps create successful individuals, which in turn builds strong, healthy families and communities. It provides care and protection of vulnerable children and young people, employment, income support and superannuation services, funding to community service providers, social policy and advice to government, and student allowances and loans.

Website: http://www.msd.govt.nz/

Benefit dynamics data

This data includes information about the counts and time periods of people receiving government transfers (benefits), and relationship information where individuals have a dependent partner or children. Information about events leading to the beginning and end of receiving a benefit is also provided.

Details of benefit spells; including event leading to spell; benefit type.

Data includes:

- personal identifiers and details; ie social welfare number, IRD number, date of birth, sex, ethnicity
- details of the partner and/or child included in the benefit; ie social welfare number, date of birth
- benefit spell details; ie the event that led to the spell, incapacity reason codes, benefit type, spell start and end dates, and the occupation code, last date worked, and last weekly earnings.

Student Loans and Allowances

This data includes eligible students who apply for, and receive, financial support under these schemes.

Data includes:

- personal identifiers and details; ie IRD number, name, date of birth, sex, ethnicity, address
- enrolment details (for where a loan or allowance was received); education provider, student ID number
- student loan details; ie type and amount of loan payments, sum of repayments and refunds, loan balance and interest transferred to Inland Revenue
- student allowance details; ie type and amount of allowance paid, number of weeks of allowance, student's income, number of partners (eg spouse, civil union, etc), number of studying partners, total partner income, each parent's income (if parent-income tested).

Ministry of Education

The Ministry of Education is the Government's lead advisor on the New Zealand education system, shaping direction for sector agencies and providers.

Website: http://www.minedu.govt.nz/

Tertiary and secondary information

This data covers all students enrolled or who have completed formal/non-formal tertiary qualifications at government funded tertiary education organisations.

Secondary school transitions; tertiary provider enrolments and qualifications; industry training.

Data includes:

- personal identifiers and details; ie NSN student number, TSEC student number, name, date of birth, sex, ethnicity
- secondary school data; ie school standards and qualifications and transitions
- tertiary workplace data; ie industry training, modern apprenticeships and targeted training
- tertiary provider data; ie enrolments in courses and qualifications, completions, provider, student ID numbers, provider type, field of study.

Student Loans Account Manager

This dataset is consolidated unit record data containing key information about student loans and allowances from the Information Analysis Platform. The Student Loans Account Manager is no longer in use.

Data includes:

- personal identifiers and details; ie IRD number, name, date of birth, sex, ethnicity, address postcode
- enrolment details (for where a loan was received); ie provider, student ID number, attendance type (ie full time/part time)
- student loan details; ie student loan payments (including the amounts borrowed for fees, course-related costs, and living costs), interest and repayments, and the loan start and end dates.

Ministry of Business, Innovation and Employment

The Ministry of Business, Innovation and Employment (MBIE) contributes to the Government's goals of building a more competitive and productive economy, delivering better public services, rebuilding Christchurch, and creating more affordable housing.

Website: http://mbie.govt.nz/

Migration and movements

The data contains all individuals who have had a visa application decided by Immigration New Zealand within the reference period and all individuals who have a record of movement across New Zealand's border within the reference period.

Data includes:

- personal identifiers and details; ie client code, passport number, name, date of birth, sex, nationality
- international movement details; ie arrival and departure date, ship number, visa type and status

- visa application details; ie visa type, policy criteria, occupation, education institution.

Migrant survey

The Migrant Survey covers migrants aged 18+ who were admitted under the Skilled Business (principal and secondary applicants), Family (Parent and Partner), and Work Visa schemes. Excluded from the survey were migrants admitted under the Students scheme and those admitted under the Pacific Quotas schemes.

Migrant Survey data:

- personal identifiers and details; name, date of birth, country of birth, sex, passport number, citizenship, marital status, occupation
- departure and arrival details; dates, New Zealand and overseas port codes, overseas zip codes, travel month, aircraft details, permits, reasons for travel.

New Zealand Customs Service

The New Zealand Customs Service is the government agency with the job of ensuring the security of New Zealand borders.

They protect the economy from illegal imports and exports. They promote New Zealand's international trade. They collect revenues, investigate illegal activity, and prosecute where necessary.

They also make sure that lawful travellers and goods can move across our borders as smoothly and efficiently as possible.

Website: http://www.customs.govt.nz/about/Pages/default.aspx

Departure and arrival cards

This data contains overseas visitors, New Zealand resident travellers, and permanent and long-term migrants entering or leaving New Zealand. Statistics NZ receives a sample of departure and arrival cards from the New Zealand Customs Service, which is then supplied to the IDI.

Departure and arrival cards data:

- personal identifiers and details; name, date of birth, country of birth, sex, passport number, citizenship, marital status, occupation
- departure and arrival details; dates, New Zealand and overseas port codes, overseas zip codes, travel month, aircraft details, permits, reasons for travel.

Ministry of Justice

The Ministry of Justice delivers court and tribunal services including collection of fines and reparation, provides policy advice and negotiates Treaty of Waitangi claims on behalf of the Government. The Ministry of Justice is the lead justice sector agency and also supports the judiciary.

Website: http://www.justice.govt.nz/

Charges data

This data covers all persons who have had a charge laid against them in court and where the case has been disposed.

Ministry of Justice charges data includes:

- personal identifiers and details; name, date of birth, sex, ethnicity, address,
- charges details; offence code, offence details.

Department of Corrections

The Department of Corrections works to make New Zealand a better, safer place by protecting the public from those who can cause harm, and reducing re-offending.

Website: http://www.corrections.govt.nz/

Sentencing data

The sentencing data contains period information about the sentences served by offenders.

Corrections sentencing data includes:

- personal identifiers and details; offender id, name, date of birth, sex, birthplace
- offending details; sentencing start and end dates, sentencing details, first and last hearing dates.

Accident Compensation Corporation

The Accident Compensation Corporation (ACC) provides comprehensive, no-fault personal injury cover for all New Zealand residents and visitors to New Zealand.

Website: http://www.acc.co.nz/

Injury data

This data contains information on workplace injury claims.

Injury data:

- personal identifiers and details; name, date of birth, sex, ethnicity, employee and employer IRD numbers, address, occupation, fund account
- Injury details; dates, activity, medical fees paid, injury causes, claim number, injury diagnosis code.

Statistics NZ

Statistics NZ is a government department and New Zealand's national statistical office. It is New Zealand's major source of official statistics and leader of the Official Statistics System.

Website: http://www.stats.govt.nz/

Household Labour Force Survey

The Household Labour Force Survey (HLFS) provides a regular, timely, and comprehensive portrayal of New Zealand's labour force. Each quarter, the HLFS produces a range of statistics relating to employment, unemployment, and people not in the labour force.

Data includes:

- personal identifiers and details; reference number, name, date of birth, sex, ethnicity
- labour force and work variables; ie part-time / full-time status, labour force status, work for pay or profit, absent from work, more than one paid job, actual hours worked, usual hours worked, prefer to work more hours, occupation, industry.

New Zealand Income Survey

The New Zealand Income Survey (NZIS) is a supplement to the HLFS that is run on a yearly basis to provide information about average weekly income. The results of the NZIS are used by the Government when making decisions about the minimum wage.

Data includes:

- personal identifiers and details; reference number, name, data of birth, sex, ethnicity
- labour force status
- income source
- average hourly, weekly and overtime earnings.

Longitudinal Business Database

The Longitudinal Business Databases integrates survey and administrative data about New Zealand businesses.

Longitudinal Immigration Survey: New Zealand

The Longitudinal Immigration Survey: New Zealand (LisNZ) collected information on how well migrants settle, both socially and economically, during their first three years as permanent residents of New Zealand.

Longitudinal Survey of Immigration New Zealand data includes:

- personal identifiers and details; address, income within New Zealand and overseas, education, ethnicity
- migration details; language competency, family details, spouse occupation and income, health, occupation, travel experience, reasons for migration, satisfaction within New Zealand.

Survey of Family, Income, and Employment

The survey of Family, Income and Employment (SoFIE) collects information on respondents' work, family, household circumstances, income and net worth.

Survey of Family, Income and Employment data includes:

- personal identifiers and details; name, sex, date of birth, address, marital status, country of birth, school details, income, assets, family relationships dependencies,
- health details; physical injury or illness/ limited mobility, emotive details, addictive substances, fruit and vegetable intake, doctor/dentist visits
- labour market details; employment information, other income details, occupation hours worked, average earnings, benefits, KiwiSaver, education information
- household details; rates, mortgage, rent, number of bedrooms, appliances.