

TTP DEMOGRAPHIC DATA QUALITY REPORT

DATA SOURCE

DATA SUPPLIER:	Department for Employment and Learning(DEL)
DESCRIPTION:	Annual downloads from 2000 onwards by academic year. The academic year for student data begins on August 1st and finishes on the July 31st of the following calendar year.
COVERAGE:	Students, at all ages, who have been enrolled in a government funded Higher Education Institute (HEI) within NI or students who are enrolled in a HEI within the rest of the United Kingdom and have a non term-time address in Northern Ireland from the academic years 2000/2001to 2013/2014.
NUMBER OF RECORDS:	1,130,963 records covering academic years from 2000/2001 to 2013/2014. (Approximately 80,783 per year) https://www.delni.gov.uk/publications/enrolments-uk-higher-education-institutions-northern-ireland-analysis-2013-2014

DATA BACKGROUND

The Enrolment data, which is collected by the Higher Education Statistics Authority (HESA) and passed on to DEL, contains details of all students, at all ages, who are enrolled in a government funded Higher Education Institute (HEI) within NI or students who are enrolled in a HEI within the rest of the United Kingdom and have a non term-time address in Northern Ireland from the academic years 2000/2001 to 2013/2014. Therefore, the information supplied does not cover those students that have enrolled in a third level education course that is taught by a body not recognized as a HEI, e.g. within a Further Education college. The information also omits postdoctoral students and those awarded higher documents. The data does include details of part-time students, foreign students including visiting and exchange students. These are students based at an institution abroad who study at a UK institution for at least 8 weeks.

Enrolment data is collected based on returns by the HEI's through student record. This return covers all students who attend each institution, therefore issues of sampling error and estimates are not relevant to this data. Accuracy of the information returned and the scale of missing data are more important. These are kept to a minimum by having in place stringent data quality checks and validation procedures.

The HE institution at which each student is registered is responsible for submitting the data to HESA about that student. The institutions data must go through over 700 validation checks in order for a return to be accepted. These checks ensure that the data are accurate in terms of format and logic. There are specific validation checks for NI HEIs which cover variables collected from NI HEIs only, for example, religion, dependents and marital status. Year on year changes are examined closely to see if they fall outside of an expected range and counts of students are also compared annually with returns made to funding bodies in respect of state funding allocation. Any issues arising from any of the above stages of quality assurance are returned to the institution for verification.

In terms of missing data items, the majority of data items are collected for all students. Some data items may include categories for 'unknown' or 'information refused' and not all data items are mandatory for a student to provide. The level of unknown entries within data items are routinely monitored during the data collection process. Any HEI recording abnormally high levels of unknown values in key data items are strongly encouraged to reduce this level over time .

DEMOGRAPHIC INFORMATION

Information available to the TTP for linkage purposes -

- Student Name which includes Forename(s), Surname and Previous Surname
- Home Address information
- Term-time Address information
- Gender
- Date of birth

DATA QUALITY & PRE-PROCESSING

Record linkage is highly dependent on the quality of the data being linked. Many key variables can be presented quite differently between datasets as the information has been collected and recorded using different methods. This can greatly complicate record linkage unless understood ahead of time. Pre-processing can be used to standardise these variables into a consistent format prior to linkage.

NAME INFORMATION- Almost all records have forename and Surname information. The small number of cases with missing name information relate to the earliest downloads.

- In some instances multiple names will be recorded as the Forename. This can be edited to leave the first name that appears as the Forename and any additional names as the Midname;
- Punctuation and characters which do not appear in the English alphabet (e.g. Á, %, \$...) can also appear in the name information. A function can be used to replace these with usable characters;

ADDRESS INFORMATION- 12% of Home postcode and 57% of term-time postcode information are null, not known or incomplete ('BT'). It should be noted that for individuals who are usually resident outside of Northern Ireland (e.g. exchange students), it be unlikely to match these to another data set as the usual address will be outside Northern Ireland.

- For purposes of matching Home postcode is initially used, followed by Term- time postcode to maximise the number of matches
- Postcode information is standardised prior to linkage.

GENDER- Less than 0.01% of all records do not contain the gender of the student.

- Gender information should be standardised prior to linkage.

DATE OF BIRTH- All records contain the Date of Birth information.

- Date of birth is provided in two date-time formats (dd/mm/yyyy and yyyy/mm/dd). It can be parsed into its constituent parts (day, month and year) prior to the linkage process to allow data to be linked.

MATCHING METHODS

Using the demographic information provided the following matching can be undertaken-

- Person matching
- Address matching of the students home address
- Address matching of the students term-time address

PREVIOUS LINKAGE MATCH RATES

The HESA information provided by DEL had previously been matched to the 2011 Census data with an adjusted match rate of 97.9% (90.6% raw match rate).



EVALUATION

The demographic data from the higher education enrolments provided by DEL is of very good quality and can be used for matching without extensive pre-processing. It should be noted that when matching the HESA information to another dataset there may be a small number of records that we are unable to match. These will include students from Northern Ireland who are studying elsewhere (term-time address outside NI) and those students from elsewhere who study in Northern Ireland (home address outside NI).