Identifying duplicate Divorce records in the data for 1988 to 2009

This document provides background information about the revision of the historical series of Divorce figures in March 2012.

How ‘duplicate’ records were created

Courts used to provide information on divorces to National Records of Scotland (NRS)’s, formerly the General Register Office for Scotland (GROS), on paper. The procedures that GROS used had some weaknesses. One of these was not checking the newly-received data against what was in the records already stored in our statistical database. Staff would just allocate a new ‘GROS unique identifier’ to each form that they processed. As a result, if a court sent a second form and extract for the same divorce, GROS would process them as if they were data for a completely new divorce, with the result that the statistical database would contain two records for the same divorce.

The information that was held in GROS's statistical database for the Divorce records which were created from paper forms

The details that GROS keyed into the records for its statistical database included:

- a code for the court;
- the court reference for the case;
- the marriage reference details (year, registration district and entry number - in cases where the couple had been married in Scotland);
- the date of the marriage;
- the dates of birth and/or the ages at marriage of the husband and the wife; and
- when the decree was granted (the month and the year - but not the exact date).

These data items were used (as described in the next section) to identify apparent ‘duplicate’ records. GROS also keyed into the statistical database several other data items (e.g. the type of action and the grounds for divorce) whose values were not used to identify apparent ‘duplicate’ records.

There were occasions when some of these details were not provided on the form, and therefore could not be keyed into the statistical database. For example, the court reference might not have been entered on the form, or a Scottish marriage's reference details might be missing because they were not available to the court staff, or the boxes for the dates of birth of the husband and/or the wife might have been left blank. There could also be occasions on which the same information was recorded in different ways on different forms (e.g. the court reference for a particular divorce might have been entered as ‘M812/96’ on one form, and as ‘M812-96’ on another form; or might have been put with two digits for the year on one form, and with four digits for the year on another form).

The GROS statistical database did not contain the names of the husband and the wife (or even their initials), so we could not see immediately (in the vast majority of cases) whether or not two records were likely to be ‘duplicates’ (in the sense of both records relating to the divorce of the same marriage).

Identifying apparent 'duplicate' records in the historical statistical database

While developing its new statistical database, NRS found some pairs of divorce records which had the same marriage reference details, and therefore appeared to relate to the same marriage. Such records were not 'exact' duplicates of each other, because they might have different values for some of the other data items, or one record might have the values missing for some data items that were available for the other record. They all had different GROS unique identifiers, so it was clearly not a case of duplicate records having been generated by a flaw within the computer system.

Because the historical statistical database did not contain any names, and because some data items might be missing or completed inconsistently, there was no simple way to identify all its records which appeared to be for divorces of the same marriage, and so determine which records were 'duplicates' of each other. To do that, NRS had to use combinations of the information that was held in several fields. NRS used three approaches to identify sets of possible duplicate records, by matching together sets of records for which the same values had been provided for:

- **Court code** and **court reference** - NRS then determined whether they were records for the same divorce by considering the values of the following 'checking' fields: marriage reference details (year, registration district and entry number); date of marriage; dates of birth of husband and wife; ages at marriage of husband and wife; and when the decree was granted.

- **Court code**, **when the decree was granted** and **date of marriage** - NRS then determined whether they were records for the same divorce by considering the values of the following 'checking' fields: court reference; marriage reference details (year, registration district and entry number); dates of birth of husband and wife; and ages at marriage of husband and wife.

- **Marriage reference details (year, registration district and entry number)** - NRS then determined whether they were records for the same divorce by considering the values of the following 'checking' fields: court code; court reference; date of marriage; dates of birth of husband and wife; ages at marriage of husband and wife; and when the decree was granted.

For each approach, NRS identified as:

- **'probable duplicates'** cases where there were no differences in any of the 'checking' fields;

- **'possible duplicates'** cases where there were, say, only 1-2 differences in the 'checking' fields, and 3 or more of the 'checking' fields had the same value; and

- **'unlikely duplicates'** cases where there were, say, 3 or more differences in the 'checking' fields.

Note: NRS did not count as a 'difference' a case where a particular field had a value in one record but was blank/missing in another record, presumably because no information had been provided on the form that was keyed to produce the second record. Nor did we count as 'the same' a case where a particular field's value was blank/missing in both records.

Deciding which of the apparent ‘duplicates’ related to the same divorce

NRS produced separate lists of the records that were classified as ‘probable’, ‘possible’ and ‘unlikely’ duplicates. Each such list was in order of the number of ‘checking’ field values that were the same and, within that, the number of ‘checking’ field values that differed. The lists showed the value of each ‘checking’ field for each record in each pair, side-by-side. NRS used that information to decide where, within each list, was the ‘cut-off’ between (a) records that appeared to relate to the same divorce and (b) records that, apparently coincidentally, had the same values for the fields used by the matching process but whose ‘checking’ fields' values were so different that it was most unlikely that they related to the same divorce.

NRS looked at the extracts of the decrees of divorce for (in total) 54 of the pairs of records ‘around’ the ‘cut-off’ points, in order to check that the overwhelming majority of those that appeared to be ‘duplicates’ were indeed pairs of records for divorces of the same couple's marriage, and that the overwhelming majority of those that appeared to be ‘coincidences’ were indeed records for the divorces of different couples' marriages. So, scrutiny of the extracts provided evidence to support the choice of the ‘cut-off’ points.

Why ‘duplicate’ forms and extracts had been sent to GROS by the courts

The results of scrutinising of the extracts also suggested that ‘duplicate’ paper forms and extracts of decrees for the divorce of the same marriage had been sent to GROS by the courts for a number of different reasons, including:

- the submission of revised or corrected information about the divorce and/or a new version of the extract of the decree of the divorce. In some cases, the earlier extract was marked ‘cancelled …’ because the staff responsible for the Register of Divorces had been able to associate the two versions of the extract, and mark the earlier one as ‘cancelled …’. Unfortunately, GROS had no procedure for cancelling the original statistical record for a divorce;
- two extracts of the same decree of divorce being produced by the same court, sometimes several years apart. Perhaps court staff thought that, in such cases, they should also resubmit the data for the divorce; and
- separate proceedings in different courts, perhaps several years apart. For example, one record might state that the husband was the Pursuer/Applicant and the wife was the Defender/Respondent; the other record might show the wife as the Pursuer/Applicant and the husband as the Defender/Respondent - so two separate decrees of divorce were granted in respect of the same marriage.

Choosing which ‘duplicate’ records to cancel

NRS decided simply to keep the most recently-processed record for the divorce of any given marriage, on the grounds that it might have been submitted by the courts to replace the data which had been received earlier (e.g. to correct errors in what had been supplied previously, or because additional information had become available). While almost all the ‘duplicates’ occurred in pairs, there were a few instances of there being three (or more) statistical records which related to the divorce of the same marriage. In such cases, NRS kept only the one which had been processed most recently.

In total, NRS cancelled 810 ‘duplicate’ records. 278 of these had been identified as such by all three of the approaches described earlier, 321 had been identified by two

of the approaches, and 211 had been identified by only one of the approaches. It should be noted that the lack of suitable data meant that it was not possible for some approaches to identify some ‘duplicates’. For example:

(a) the ‘court code and court reference’ approach could not identify duplicates if the court reference was not available (which was the case for almost every pre-1998 record); and
(b) it was not possible to use the ‘marriage reference details’ approach in cases where that information had not been supplied (e.g. because the couple had been married outwith Scotland).

The ‘GROS unique identifier’

As mentioned above, GROS staff would allocate a new so-called ‘GROS unique identifier’ to each paper form that they processed. Strictly speaking, the ‘unique identifier’ for a data record created from a paper form was actually the combination of the values of two fields, both of which were allocated by the staff:

- the so-called ‘proxy year’ - which was normally the year in which the decree was granted, but which could sometimes be a later year (generally, if GROS had received the data more than a couple of months after the end of the year in which the decree was granted); and
- a serial number - the value of which was unique within the ‘proxy year’.