The methodology used in creating a Comtrade based dataset of small arms transfers

Memorandum by Nicholas Marsh

Norwegian Initiative on Small Arms Transfers project

at the

International Peace Research Institute, Oslo

28 September 2005

Part 1 – Loading the data

Introduction

Customs data used by Comtrade, and many other national and regional sources, is defined by the Harmonised System (henceforth referred to as HS). This is a set of universal nomenclature which defines, via a series of numeric codes, every good reported as being transported over applicable borders. Each numeric code refers to a defined category of goods, for example 930200 refers to 'pistols and revolvers'.

The existence of such a universal system is of immense benefit to researchers on the arms trade. Other data sources, such as annual reports to parliament, use a variety of methodologies and variables. It is therefore very difficult to use them to develop a picture of global or regional patterns (Haug, Langvandslien, Lumpe and Marsh, 2002).

The HS derived customs data is not without complexities, though. In particular the HS nomenclature has periodically been revised. In recent years, such revisions occurred in 1992, 1996 and 2002. The 1992 revision involved major changes to the categories used concerning the transfer of arms. It is therefore difficult to construct time series data that straddles 1992. The 1996 revision offered no major changes to the arms related nomenclature.

The 2002 revision, however, replaced two of the HS1996 categories 930100 and 930590 with new codes. Thus 930100 was superseded by 930111, 930119, 930120, and 930190. Similarly, 930590 was replaced by 930591 and 930599. Other 1996 codes were not changed in the 2002 revision of the HS nomenclature.

A further complicating factor is that governments have not uniformly used the same nomenclature at the same point. For example, in 2003, countries such as Indonesia or the Philippines continued to report using the HS1996 nomenclature, while other countries had started using the newer HS 2002 nomenclature. Experience indicates that a small number of countries may also still use (as of 2005) the HS 1992 nomenclature.[1]

The next complication is that data can only be downloaded from Comtrade one nomenclature at a time. The final complication is that Comtrade converts the newer nomenclature back into the previous ones (to aid retrospective comparison of the data). Therefore, it is quite normal that data downloaded concerning the same country, but using different nomenclature, will present data in different categories.

Therefore, the conclusion of this introduction is that in order to present a comprehensive picture of the global trade in small arms and light weapons it is necessary to download data from the three nomenclature discussed above (HS 1992, HS 1996 and HS 2002). They must then be thoroughly filtered to ensure that only one record is present in the database, and that this record comes from the most recent nomenclature.

[1] Private communication with Ronald Jansen, Chief Commodity Trade Statistics Section, United Nations Statistics Division, several times, 2004.

Step 1

Delete all pre-existing Comtrade data concerning the years about to be loaded.

Then the following process, described in steps two and three, is carried out, in turn, for all three nomenclature.

Step 2

Convert UN country codes into those used by the NISAT database (based upon the Correlates of War project country codes).

Add descriptive text to the HS nomenclature codes (for example (930200) becomes 'Pistols and revolvers (930200)'.

Amend the records, as necessary, with text notes pertaining to the Comtrade data. For example records concerning South Africa are amended with a not to the effect that they actually concern the South African Customs Union.

Step 3

Filter the records (from the three nomenclature and delete any duplicates, while retaining the record from the most recent nomenclature).

This process uses the following logical steps:

1. IF there is an existing record with the same Data_Source, Year, Country_Code and Weapons_Type and Comment string as the one just added
   THEN delete existing record.
   ELSE do nothing

2. IF year >= 2002
   AND IF the record just added has not been deleted in accordance with Rule 1 above
   THEN the system checks whether the Comtrade weapons code for the record just added is 930111 OR 930119 OR 930120 OR 930190 OR 930100 OR 9301.

   IF the above condition is TRUE
   THEN the system checks if there is an existing record in the table which also contains one of the above weapons codes AND where the Year and Country_Code (partner country) match AND where the Data_Source = 6 (Comtrade) AND where the Comment strings match.

   IF one or more records are returned
   THEN they are ordered by Comtrade weapons type code. The system then locates the record in the set which relates to weapons type 930100 OR 9301, and the record is deleted and an entry made in the log that a record has been deleted under Rule 2.

3. IF year >= 2002
   AND IF the record just added has not been deleted in accordance with Rule 1 or Rule 2 above
   THEN the system checks whether the Comtrade weapons code for the record just added is 930590 OR 930591 OR 930599.

IF the above condition is TRUE
THEN the system checks whether there is an existing record in the table which also contains one of the above weapons codes AND where the Year and Country_Code (partner country) match AND where the Data_Source = 6 (Comtrade) AND where the comment strings match.

IF one or more records are returned
THEN they are ordered by Comtrade weapons type code. The system then locates the record in the set which relates to weapons type 930590, and the record is deleted and an entry made in the log that a record has been deleted under Rule 3.

4. IF the record just added relates to weapons type 930111
   THEN the entry is deleted.

The data has now been entered into the database tables.

Part 2 – Agglomerating the data

Introduction

After the data has been filtered and loaded, it is still unsuitable for the analysis of global and national trends. This is because of the factors which are described below.

First, for every country there will, potentially, be two sources of data. These are the country's reports of its exports or imports (known as 'base data'); and its partners' reports of their exports to, or imports from, it (known as 'mirror data'). For example, Italy may report an export of category 930200 to Canada - this is the base data. Canada may also have reported an import from Italy of the same category. This report would be the mirror data.

In an ideal world, the base and mirror data would correspond exactly. Unfortunately, this only happens very rarely. Discussions with the International Trade Center indicate that this is a well known problem.[2] Explanations range from fraud (customs are used to levy tax), the dates of export and import occurring in different years (export in December 2004 and an import in January 2005), countries used as transit points (exports declared to the Netherlands, for example, may simply transit through that country and ultimately be declared as imports by Germany), and of course human error.

Thus, a comprehensive answer to the question "What did a country export and import during 2002?" would require four different datasets. These are described below.

Exports base	Exports mirror	Imports base	Imports mirror
x	x	x	x

Of course providing four different answers to such a straightforward research question is not a very practicable was of evaluating global trends.

Second, many countries simply do not report any data to Comtrade (or other databases such as Comext for that matter). For example, Austria has not reported any exports of 930200 'Pistols and revolvers'. This is an anomaly, as the Glock Company, located in Austria, is known to be one of the world's largest producers of pistols. This problem, can, of course, be solved by looking at the mirror data – who has reported imports of pistols from Austria.

Third, as noted above, countries have reported using different nomenclature. In such a circumstance, it will be difficult to directly compare base and mirror data.

Fourth, some countries censor the data by reporting that their partners were 'unspecified' countries. These include partners such as 'special categories' or 'Areas, not elsewhere specified'. We therefore have information that an export or import has been made, but no information on which country it concerns.

[2] Interview conducted at the Small Arms Survey, Geneva with Friedrich von Kirchbach, Chief, Market Analysis Section, International Trade Centre, UNCTAD/WTO. 21 December 2004.

These problems are addressed by agglomerating the data. The term ‘agglomerated’ has been deliberately chosen over ‘aggregated’, as aggregation is definitely not what is happening. In a similar manner to the ‘data loading’ procedure outlined above, the agglomeration process filters, and slightly modifies, both base and mirror data, and so derives a single figure for a country’s exports and imports.

Step 1

First, new database tables are created concerning the exports and imports of each country for which we have data.. These tables are known as ‘Aggregated tables’ as they contain all relevant base and mirror data. The following procedure is used:

1. The application opens, for every country, the Import/Export tables in the NISAT database in turn, and returns all of the records which match the year and weapons types criteria.

2. For each matching record, the Agglomerator Program places one copy of the record in the relevant Aggregated table, and a further copy in the relevant mirror Aggregated table, with the IsMirrorData flag set to True (-1). Thus for each record in the NISAT base tables, two records are copied into the aggregated tables.

Step 2

The next step is to remove two general categories that may cause double counting of the data.

Comtrade data includes, for every country, aggregated totals of the summed value of all a country’s trades in a particular category in a given year are. As the Comtrade data also contains the individual trades to each country, the summed totals are superfluous. They are therefore deleted using the following procedure:

```
IF Partner Code = -2 (All Countries) THEN
        Is there at least one other entry using base data in the table for this weapons type?
        IF TRUE THEN
                delete this record
        ELSE
                Retain the All Countries record
                LOG this in the on-screen log and in the log file.
                        END IF
END IF
```

Furthermore, Comtrade data for some countries includes re-exports. Such re-exports are reported in addition to normal exports. Furthermore, the total value of normal exports includes the value of re-exports. Therefore, to prevent double counting, the re-exports data is removed.

Step 3

Step three concerns deciding which data to pick – either from the base or mirror dataset. The Agglomerator application performs a complex set of calculations to determine the reliability of each country's data for each weapons type.

This is achieved by comparing the value of each base transaction with its mirror counterpart. Countries that, on average, have base transactions that are closer in value to their mirror counterparts are viewed as producing more reliable data than countries in which, on average, there are very large discrepancies between their reports and the data supplied by their mirror partners.

Of course, this method is largely 'self referential'. It would be preferable to compare each item of base data with an external data source (other than Comtrade). However, unfortunately, for the vast majority of countries, such an external data source does not exist. Even when we do have additional data sources, they are unsuited to be used to assess the veracity of Comtrade data. This is because alternate data sources, such as reports to parliament on arms exports, often use very different methodologies to Comtrade (such as reporting export licences or using different categories of weapons). Conversely, other sources of customs data (such as Comext) can not be used, as they are based upon the same ultimate source (national customs authorities) as Comtrade. Readers should see Haug, Langvandslien, Lumpe and Marsh (2002) for more information.

The method described below was developed from a publication by the International Trade Centre (2003), discussions at PRIO, and discussions between the Ag.

The reliability statistics are generated as follows.

1. A new table is created in the NISAT database called Reliability, with the following field definitions:

Reporter_Code (Integer)
Reporter_Name (Text)
Year (Text)
Weapons_Code (Long)
IsImports (Boolean)
Reliability (Single)

The Reliability field contains the reliability index for this reporter, for the given year and weapons code. It is a number between 0 and 100 - the lower the number, the more reliable the reporter is deemed to be.

The logic by which the Reliability figure is arrived at is complicated, and is arrived at as follows (this example concerns the procedure for a single export table):

FOR EACH weapons type in the year in question:

dblAggTotalTrade = total trade for this reporter, base and mirror, all partners >0

FOR EACH Partner Country > 0 (i.e. not regions or Unspecified Country)

```
        dblSource = value of base data for this partner(/year/weapons type)
        dblMirror = value of mirror data for this partner(/year/weapons type)
        dblDiscrepancy = 100 * (dblMirror + dblSource) / (dblSource –
                                                    dblMirror)
        dblAbsDiscrepancy = Abs(dblDiscrepancy)
        dblTotalTrade = dblSource + dblMirror
        sngWeight = dblTotalTrade / dblAggTotalTrade
        dblWeightedAbsDiscrepancy = dblAbsDiscrepancy * sngWeight
        dblAggWeightedAbsDiscrepancy = dblAggWeightedAbsDiscrepancy +
                                                        dblWeightedAbsDiscrepancy
         NEXT Partner Country

         Write to Reliability table: Reliability = dblAggWeightedAbsDiscrepancy

    NEXT Weapons Type
```

The above process is repeated for all remaining export tables, and then for all Import tables.

Step 4

The next step is designed to avoid double counting. As noted above, many countries report trades, but do not state the country destination (stating instead destinations such as ‘special categories’). This poses a problem, as such trades can not be filtered via the reliability calculator (described in Step 3). However, they still represent a likely cause of double counting if they are combined with mirror data.

For example, in 2003 the UK just reported exports of category 930190 to ‘special categories’ 5 385 129 USD. It did not report any other trades to individual countries. However, the ‘mirror’ reports of its partner’s imports of 930120 are:

Country	**USD**
Australia	217546
Canada	91876
Ireland	15009
Japan	877681
Korea; South	222345
Maldive Islands	24658
New Zealand	1668
Norway	504046
Switzerland	7300
Turkey	13576
United States of America	310080
Total	2285785

Just adding the UK export to ‘special categories’ to the value of the mirror data would potentially double count UK exports. This is because the mirror trades are likely to have been included in the in the total reported as being exported to ‘special categories’.

The Agglomerator programme therefore searches through the data, locates all the mirror trades that do not have a counterpart in the base data. It then sums the value of these 'unmatched mirrors' and deletes this total from the value of reports concerning unspecified partners, such as 'special categories'. This procedure is described below:

FOR each aggregated table
    FOR each year
        FOR each weapons type
            - Sum the value of "mirror" trades for which no base data is available
            - Subtract this value from the value of the base record for "Unspecified Country"
        NEXT weapons type
    NEXT year
NEXT aggregated table

Step 5

The last step involved the final deletion of duplicated data. As noted in Step 1, the Aggregated tables contain both the base and mirror data. This data has been slightly modified in steps 2 and 4, and the reliability score calculated in step 3.

It is now necessary to finally filter and delete either the base or mirror records, as appropriate, concerning each transaction. (In most cases, each export will have a corresponding mirror import and vice versa).

The first task is to ensure that the most recent nomenclature used, and previous iterations are filtered out. This is achieved using the methods below:

CREATE list of records in the table where Weapons Code = 930100 OR 9301
    FOR EACH record
        Find out if there is a record with the opposite IsMirror flag, with the same
        Partner Code, with Weapons Code 930190 OR 930120 OR 930119
        IF TRUE then
            Delete the record with Weapons Code 930100
            LOG the deletion in the log
        ELSE
            Do nothing
        END IF
    NEXT record

CREATE list of records in the table where Weapons Code = 930590
    FOR EACH record
        Find out if there is a record with the opposite IsMirror flag, with the same
        Partner Code, with Weapons Code 930591 OR 930599
        IF TRUE then
            Delete the record with Weapons Code 930590
            LOG the deletion in the log
        ELSE
            Do nothing
        END IF
    NEXT record

The Agglomerator programme then proceeds with the primary deletion process. This is where the Reliability table is employed to decide which of two corresponding base and mirror records to delete.

```
For each Record:

    IF Partner Code < -2 (a region)
    THEN leave the record in and produce an entry in the log

    IF Partner Code = -1 (Unspecified)
    THEN leave the record in and produce an entry in the log

    IF Partner Code > 0 (distinct country) THEN
        IF there is a neighbouring Mirror record for the same year, weapons
        type and country combination THEN
            Refer to the Reliability table to determine which reporter is
            deemed to be more reliable (which one has the lower
            Reliability factor) and delete the record (base or mirror) from
            the less reliable reporter.
        END IF (mirror record exists?)
    END ID (is this a distinct country?)

NEXT Record

NEXT Aggregate Table
```

The country tables are now complete. To recap, they now contain one trade

Step 6

The final step is the creation of one single dataset out of the circa 400 country tables.

First, the dollar values are converted into ECU/EUR.

Next, the transfers are read from the country tables and inserted, one by one, into the master table. This process will inevitably involve almost all the records being duplicated – an import from one country will have a corresponding and identical record from the exporter.

The final phase in step 6 is to remove the duplicates.

Step 7

The data in the master table is exported to a .csv list.

Users should note that the data in the csv file is arranged as follows:

Item number	Content
1	Year (text, 12 chars)
2	Exporter_Code (integer)
3	Exporter_Name (text)
4	Importer_Code (integer)
5	Importer_Name (text
6	Weapons_Code (long integer)
7	Value_[EUR/USD]

The Exporter and Importer codes are the same as those used by the Correlates of War dataset. However, we have added the following codes:

-2	All countries (this code is not used in the current dataset, but is used in other data in the NISAT database)
-1	Unspecified country (if a government reported a trade but not the identity of the partner)
-100	North America
-200	European Union
-300	Central America
-400	South America
-500	Africa
-600	Asia
-700	Europe
-800	Carribean
-900	Oceania
-1000	Island Territories
-1100	Middle East
-1200	Central Asia
-1300	East Asia
-1400	CIS
-1500	Latin American Integration Association (LAIA)
-1600	Americas
-1700	NATO

-1800	China/Hong Kong/Macao

Users should note that employment of a regional code (-100 to -1800) is a very rare occurrence.

Users should also note that Step 4 (see above) means that many transactions concerning ‘Unspecified countries’ with a code of -1 have indicated that the value of the trade is zero. This is a consequence of the unspecified trade being modified downward to avoid double counting.