SDMXUSE

MODULE TO IMPORT DATA FROM STATISTICAL AGENCIES USING THE SDMX STANDARD

Sébastien Fontenay
sebastien.fontenay@uclouvain.be


**Introduction**

- **sdmxuse** is a user-written command available from the SSC archive since Sept. 2016
  - [https://ideas.repec.org/c/boc/bocode/s458231.html](https://ideas.repec.org/c/boc/bocode/s458231.html)

- The package allows users to
  - download and import statistical data from international organizations using the SDMX standard
    - The complex format of the datasets will be reviewed to show how users can send specific queries and import only the required time series
  - format the dataset into a panel or time series

- **Motivation**
  - It might prove useful for researchers who need frequently updated time series and wish to automate the downloading and formatting process
    - One can think of modern methods for forecasting economic series that exploit many predictors, often hundreds time series, which could be used as soon as they are released
# September 2016

<table>
<thead>
<tr>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>30 ESTAT – B&amp;C surveys</td>
<td>31 ESTAT – Unemployment</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5 ESTAT – Serv. turnover</td>
<td>6 ESTAT – GDP</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>20</td>
<td>21 ESTAT – Flash consumer conf.</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>26 ECB – Monet. aggregates</td>
<td>27</td>
<td>28 ESTAT – B&amp;C surveys</td>
<td>29</td>
<td>30 ESTAT – B&amp;C surveys</td>
<td>31 ESTAT – Unemployment</td>
</tr>
</tbody>
</table>

**MOTIVATION**
SDMX stands for Statistical Data and Metadata Exchange

- Initiative started in 2001 by 7 international organisations:
  - Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat (ESTAT), International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD), United Nations Statistics Division (UNSD) and the World Bank (WB)
  - More info at: [https://sdmx.org/](https://sdmx.org/)

- Their objective was to develop more efficient processes for sharing of statistical data and metadata:
  - Metadata = data that provides information about other data
  - e.g. the data point 9.9 is not useful without the information that it is a measure of the total unemployment rate (according to ILO definition) for France, after seasonal adjustment but no calendar adjustment, in June 2016
The initiative evolved around three axes:

- **setting technical standards**
  - for compiling statistical data
    - the SDMX format (built around XML syntax) was created for this purpose

- **developing statistical guidelines**
  - i.e. a common metadata vocabulary to make international comparisons meaningful (e.g. seasonal or price adjustments)

- **promoting tools to deploy web services**
  - that facilitate the access to data and metadata (RESTful web services)

The primary goal was to foster data sharing between participating organisations using a “pull” rather than a “push” reporting format:

- i.e. instead of sending formatted databases to each others, statistical agencies could directly pull data from another provider website
  - Dissemination of data to final users was somehow secondary even though the web services are accessible to the public
Concretely, users can download a dataset (when they know its identifier) by sending a request to the URL of the service:

- The result is a structured (SDMX-ML) file
  - [http://stats.oecd.org/restsdmx/sdmx.ashx//GetData/RPOP/BEL+FRA+CAN+USA.2024.2./all?](http://stats.oecd.org/restsdmx/sdmx.ashx//GetData/RPOP/BEL+FRA+CAN+USA.2024.2./all?)

- The output is really just a string of characters with text elements (data and metadata) and structural markers (called tags)
  - The tags are encapsulated between lower-than and greater-than symbols to distinguish them from the content
In order to process the file in Stata, it is important to distinguish two types of tags:

- `<SeriesKey>`, which contains the identification key of a given series
- `<Obs>`, which contains a set of observations with a time element `<ObsDimension>` and a value element `<ObsValue>`
How do we convert the file into a human-readable format

Before importing the file into Stata, we add a carriage return to the <SeriesKey> and <Obs> tags (using the command `filefilter`)

```
. filefilter sdmxfile.txt sdmxfile2.txt, from("<Obs>") to ("\r\n<Obs>") replace
```

Then, we separate the data and metadata from the structural markers

- This is facilitated by the use of the package `moss` created by Nicholas J. Cox and Robert Picard that allows for finding substrings matching complex patterns of text using regular expressions
  - This package must be installed for `sdmxuse` to work properly

```
. moss v1, match(`"value="([a-zA-Z0-9-_]+)"`) regex
```
But datasets are often very large and users may be seeking to download only a few series.

This is the reason why the statistical agencies have decided to offer a genuine database service that is capable of processing specific queries.

The organisation of this database relies on a **cube structure** commonly used for data warehousing.

The dataset is organised along dimensions and a particular series (stored in a cell) takes distinct values for each dimension (the combination of these values is called a key and it uniquely identifies this cell).
"Slicing" a data cube by processing a specific query

- To obtain only the total female population aged between 20 and 24 years in four OECD countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Female population between 20 and 24 years in 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEL</td>
<td>343190</td>
</tr>
<tr>
<td>CAN</td>
<td>1159659</td>
</tr>
<tr>
<td>FRA</td>
<td>1986450</td>
</tr>
<tr>
<td>USA</td>
<td>10900133</td>
</tr>
</tbody>
</table>
The total number of cells of the cube in the example above is 6498 corresponding to all possible crossings of the dimensions

- age groups (38) * countries (57) * sex (3)
  - But new dimensions could be added - In fact, even though it is called a cube, it is actually multi-dimensional (i.e. it allows more than three dimensions)

The user should therefore identify the dimensions to be able to make a specific query

- This is the reason why the SDMX standard provides structural metadata describing the organization of a dataset in the form of a Data Structure Definition (DSD) file
  - giving information about the number of dimensions of the dataset, the order of the dimensions, as well as the values for each dimension
The DSD gives the user enough detail to write a query for data, but it does not make any guarantees about the presence of data.

- It is quite possible that the dataset is a sparse cube (i.e. there may not be data for every possible key permutation).

```plaintext
  sdmxuse data IMF, dataset(PGI) dimensions(A1.AIPMA...)

The query did not match any time series - check again the dimensions' values or download the full dataset.
```
The program `sdmxuse` allows for retrieving three types of resources:

- **Data flows**
  - complete list of publicly available datasets with their identifiers and a description
- **Data Structure Definition**
  - metadata describing the structure of a dataset, the order of dimensions for the query and the distinct values for each dimension
- **Time series data**

The syntax varies accordingly:

- `sdmxuse dataflow provider`
- `sdmxuse datastructure provider, dataset(identifier)`
- `sdmxuse data provider, dataset(identifier)`

6 providers are currently available:

- European Central Bank (ECB), Eurostat (ESTAT), International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD), United Nations Statistics Division (UNSD) and World Bank (WB)
  - Their acronym should be written in capital letters
The following example uses `sdmxuse` to import and format population data in OECD countries.

**Step 1:** find all publicly available datasets from OECD and search for those whose description contains the word “population”

<table>
<thead>
<tr>
<th>dataflow_id</th>
<th>dataflow_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALFS_POP_VITAL</td>
<td>Population and Vital Statistics</td>
</tr>
<tr>
<td>ALFS_POP_LABOUR</td>
<td>Population and Labour Force</td>
</tr>
<tr>
<td>SNA_TABLE3</td>
<td>3. Population and employment by main activity</td>
</tr>
<tr>
<td>POP_FIVE_HIST</td>
<td>Population</td>
</tr>
<tr>
<td>SAH_URBA_CITY_LIST_7</td>
<td>Africapolis List and Population of West African urban agglomerations 1950-2010</td>
</tr>
<tr>
<td>RPOP</td>
<td>Total population by sex and age</td>
</tr>
<tr>
<td>SNA_TABLE3_SNA93</td>
<td>3. Population and employment by main activity, SNA93</td>
</tr>
<tr>
<td>POP_PROJ</td>
<td>Historical population data and projections (1950-2050)</td>
</tr>
<tr>
<td>WATER_TREAT</td>
<td>Wastewater treatment (% population connected)</td>
</tr>
<tr>
<td>AEO2012_CH6_FIG5</td>
<td>Figure 5: Employment Rate to working age population in Africa and comparators</td>
</tr>
<tr>
<td>AEO2012_CH6_BOX6_EDUDEM</td>
<td>Box 6: Rural vs. Agricultural population in Nigeria (1980-2010, in thousands)</td>
</tr>
</tbody>
</table>

- `sdmxuse dataflow OECD, clear`
- `list if regexm(lower(dataflow_description), "population"), noobs`
Step 2: find the Data Structure Definition of the RPOP dataset

- The command also returns a message to indicate the names and order of the dimensions:

  Order of dimensions: (COUNTRY.DAGEGR.DSEX.DSTATUS)
Step 3: Customized request to obtain total population aged between 20 and 24 years

- We leave the first and last dimensions empty, meaning that we want all values for those dimensions

```
. sdmxuse data OECD, clear dataset(RPOP) dimensions(.2024.90.)
```
We can reshape the dataset to get all time series in separate variables or build a panel dataset

Here, we ask separated series for men and women by specifying the values “1+2” in the dimension DSEX

- [, timeseries]
  - reshapes the dataset so that each series is stored in a single variable - variables' names are made of the values of the series for each dimension

```
.sdmxuse data OECD, clear dataset(RPOP) dimensions(.2024.1+2.) timeseries
```

- [, panel(panelvar)]
  - reshapes the dataset into a panel - `panelvar` must be specified, it will often be the geographical dimension

```
.sdmxuse data OECD, clear dataset(RPOP) dimensions(.2024.1+2.) panel(COUNTRY)
```
More options are available

- Filtering the time dimension
  - [, start()] or [, end()]
    - defines the start/end period by specifying the exact value (e.g. 2010-01) or just the year (e.g. 2010)

- Attributes
  - [, attributes]
    - downloads attributes that give additional information about the series or the observations, but do not affect the dataset structure itself (e.g. observations' flags)

- Merge dataset and Data Structure Definition
  - [, mergedsd]
    - adds new variables with labels for dimensions' values – useful if the meaning of a specific value is not transparent (e.g. ZAF for South Africa)
Many thanks to Robert Picard & Nicholas J. Cox for their program moss

I believe that SDMX is an initiative that is worth investing in because it is sponsored by leading international statistical agencies
   • Joined by more and more national organizations
     - INSEE: https://www.insee.fr/en/information/2868055

Some initiatives have already been implemented to facilitate the use of SDMX data for external users but they rely on the Java programming language
   • Formatting the data directly within Stata has proved to be quicker for large datasets

sdmxuse could become an alternative to private data providers
   • e.g. Thomson Reuters Datastream, Macrobond

Stata 15 integrates a module to import data from the Federal Reserve Economic Database (FRED)
   • built after the package freduse by David Drukker