nsm2019
Nordic Statistical Meeting
HELSINKI 26.–28.8.2019

Error Detection for the Statistics of External Trade in Goods

Garðar Páll Gíslason, gardar.gislason@hagstofa.is, Statistics Iceland

Violeta Calian, violeta.calian@hagstofa.is, Statistics Iceland

Auður Ólína Svavarsdóttir, audur.svavarsdottir@hagstofa.is, Statistics Iceland

Bryndís Bjarnadóttir, bryndis.bjarnadottir@hagstofa.is, Statistics Iceland

Kolbrún Ýr Jóhannsdóttir, kolbrun.johannsdottir@hagstofa.is, Statistics Iceland

Abstract

Statistics Iceland is modernizing the production process of the statistics on external trade in goods. This is a monthly publication, primarily based on customs declarations for imports and exports supplemented by data from the Icelandic Transport Authority and Statistics Iceland's survey regarding trade in ships, aircraft and fuels purchased abroad. Statistics Iceland also receives data from the Icelandic Post. The published statistics describe the quantities and values of imported and exported goods by customs tariff numbers /types of goods and by country of origin for imports and country of destination for exports.

The goal of our project is twofold:

1. to automatize the detection and correction of data errors to a larger extent

to provide up-to-date and flexible tools to the analysts and experts in the subject matter

The production process consists of the following stages, iterated in some cases:

S1. Data input and validation of the expected IT structural requirements

S2. Error detection and localization, by verifying content and statistical consistency of data

S3. Error correction, which includes: data imputation and editing, both automatic and interactive

S4. Advanced validation procedures

S5. Testing and assessment of validation rules and imputation methods

In this paper we focus on the stage S2, mainly on the implementation of error detection rules in a flexible way. We also explore new solutions for the development of stage S3 and S4, based on rigorous data analysis, imputation and machine learning methods. The validation rules collection, results of data analysis and the newly developed open source software will be shared within ESS and on public repositories.

***Keywords**: Error detection, error correction, external trade in goods, validation*

Introduction

In this paper we describe the modernization of the production process of the statistics on external trade in goods at Statistics Iceland (SI), based on using open source R-packages and an improved repository of standard validation rules.

The main reason why this work is necessary is that the current production process and in particular the error detection and correction systems are based on old software with little flexibility, relying on an out of date validation rule repository and on too high a proportion of manual corrections. A survey of several NSIs and of the relevant literature showed that a modern and dedicated tool is not available for such a task although open source R-tools of high quality for general data validation purposes do exist.

Therefore, we decided to build a bridge between the multiple-source and multiple language error detection rule collections that are available at SI, Eurostat and elsewhere and these R-tools. In first stages of the project, our original contribution consists of creating an R-package for translating the validation rules from the available sql-sources into R on the one hand, and on the other hand providing, maintaining and updating a unified (share-able) repository of validation rules for external trade data.

Our aim is to build a highly reliable system, unique to our knowledge at the time of writing, based entirely on open source code and able to perform the following tasks specific to production of external trade statistics:

(i) to perform a systematic and *reproducible*, error detection and error correction in a highly automatic way

(ii) to provide a statistical analysis of errors and validation *rules*, not only of data

(iii) to manage the validation rules in a flexible way

(iii) to create efficient reporting tools

(iv) to build an interface easy to use by subject matter experts

(v) to share our products as open rule repository and open R-code.

The external trade in goods statistics is a very important indicator of the situation and the development of an economy of a nation, also being a data source for National Accounts and Balance of Payments statistics. Therefore it is of vital importance to try to secure as high quality of the data as possible although it is a challenging task given its characteristics of large and detailed datasets.

The trade in goods statistics is published monthly. The largest data source for the Icelandic trade in goods are customs declarations from the Customs Authorities supplemented by data from the Icelandic Transport Authority and Statistics Iceland's survey regarding trade is ships, aircraft and fuels purchased abroad. Statistics Iceland also receives data from the Icelandic Post. The published statistics describe the quantities and values of imported and exported goods by customs tariff number (types of goods) and by country of origin for imports and country of destination for exports.

The project is focused on data from the Customs Authorities. The data is retrieved monthly from the Customs database and inserted into a processing system which test the data for errors and flags both definitive and potential errors. However, as we mentioned earlier, the current system is both old and with limited flexibility, leaving space for manual corrections to a significant extent.
The production process consists of the following stages, iterated in some cases:

S1. Data input and validation of the expected IT specific structural requirements

S2. Error detection and localization [Van der Loo, M., De Jonge, E. (2018)], [De Waal, T., Pannekoek, J. and Scholtus, S. (2011)], by verifying content and statistical properties of data

S3. Error correction, which includes: data imputation and editing of an optimum set of errors [Fellegi, I. P. and Holt, D. (1976)] as well as their impact [Statistics Sweden (2006)], both automatic and interactive

S4. Advanced validation procedures

S5. Testing and assessment of validation rules and imputation methods

In this paper we focus on the stage S2, mainly on the modernization of the error detection stage. We currently explore new solutions for the development of stage S3

and S4, as well, based on rigorous data analysis, imputation and machine learning methods.

Methods, data and tools

Data validation, as defined by the Eurostat manual[1], is "an activity verifying whether or not a combination of values is a member of a set of acceptable combinations". We understand it here as ensuring that micro-data and aggregated data are clean, correct and useful. In addition, advanced data validation tasks should be considered at later stages. These involve data analysis and mining for exploration and content checking. Our new system for error detection as a first stage of the validation process on external trade data is based on a rule repository and a suite of open source R-packages.

The focus of our present project is micro-data. By contrast, the validation of aggregated data at SI uses already a systematic error detection system based on sql-version of standard Eurostat rules.

As shown in the metadata[2], Statistics Iceland follows the guidelines given in the "United Nations: International Trade Statistics, Concepts and Definitions" as regards what to include in external trade of goods, how and when. The statistics extend to merchandise trade, and by a general definition any imports or exports of goods which add to or subtract from the stock of material resources of a country should be included in external trade statistics. A distinction is made between two systems of international trade in goods, the general trade system and the special trade system. The main difference between these systems involves the method of registering goods imported to customs bonded warehouses and free zones. According to the general trade system an item of goods is registered as import on entry into a bonded warehouse or free zone, whereas according to the special trade system such an item would be registered on entry into a country from a bonded warehouse or free zone. In Iceland the general trade system was replaced by the special trade system in 1998.

External trade of goods shows detailed information on import of goods to Iceland and

[1] https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_methodology_for_data_validation_v1.1_-_rev2018_0.pdf
[2] https://www.statice.is/publications/metadata?fileId=19622

export of goods from Iceland by countries (countries of origin for imports and countries of final destination (consumption) for exports), divided by various classifications.

The basic classification and the most detailed one is the Icelandic Customs Tariff, which is an eight digits classification that complies with the six digits of the HS classification, with the addition of two digits that are used in some instances for more detailed breakdown according to Icelandic requirements. The current tariff entered into force on 1 January 1988, including approximately 8,000 customs tariff numbers. Information on weight and fob-value for exports and on weight, FOB-value and CIF-value of imports are published. For chosen products supplementary unit is published (square metres, units, pairs and litres).

Figures on the value of imports and exports of goods are reached by converting the foreign currency value of the commodity to the Icelandic króna (ISK) based on the daily midrate of the currency concerned. The reference rate of exchange is a so-called customs exchange rate, which is the official exchange rate as registered by the Central Bank of Iceland on the last working day before the date of customs clearance.

Statistics Iceland also publishes information on external trade of goods by other classifications, mainly SITC (Standard International Trade Classification, UN), Classification of Commodities (Icelandic classification) and BEC (Broad Economic Classification, UN), as well as information on external trade by market areas. The data are not seasonally adjusted.

For the purpose of detecting errors in such micro-data, the validation rule repository is built by using in-house expert theoretical and practical experience in addition to Eurostat rules or their generalized versions. Technically, the starting point of the newly built repository consists of:

(i) sql scripts which implement local expert rules

(ii) sql scripts based on text sources like Eurostat manuals and guidelines

(iii) external expert rules

Every rule that has been collected is stored in a database, even though some of them are not relevant anymore. All rules are accompanied by attributes like use status and origin.

The main tools and resources we employ are:

- the R-packages *validate,* validatetools, errorlocate, simputations, rspa, lumberjack[3], built by Statistics Netherlands, well described in [Van der Loo, M., De Jonge, E. (2018)]

- a shareable error detection *rule repository* compiled at Statistics Iceland. This repository is already available in a beta (Icelandic) version as a database and will be shared publicly on Github in its English version[4]

- an open source R-package built at Statistics Iceland[5] for *translating* the sql - validation rules into R-syntax, necessary for the application of the R-validation packages.

The repository of error detection rules

We have created a database containing all validation rules available at this point, for the whole set of 67 variables which are currently defining the data.

There are about 220 validation rules, 170 of which are inherited from the old production system and still valid, as well as almost 100 rules derived from the general Eurostat guidelines. Some of these are overlapping but we are deliberately building a redundant set which is then streamlined by using the R-package *validatetools*[6].

Several **examples** of errors defined by the validation rules inherited from the old system are:

- fields which have NULL values or have values which do not belong to pre-defined admissible sets

- outliers in unit prices when compared to their past 12 months values

- outliers in total prices by comparing them with their past 12 months values (19% of warnings in the old system)

[3] https://cran.r-project.org/web/packages/validate/, and links therein.
[4] https://github.com/hagstofan/TranStoR/tree/master/examples

[5] https://github.com/hagstofan/TranStoR

[6] https://cran.r-project.org/web/packages/validatetools/

- rare events and extreme values, like country codes which have not been used for the last 12 months (6% of warnings in the old system) or import / export prices higher than given thresholds (26% of past warnings)

- wrong codes due to typos or confusion of very similar codes

- impossible combinations of codes for nature of transaction and statistical procedure variables

- custom tariff numbers which have been error prone according to past experience.

The list is long, evolving with time and the analysis of the whole set is a topics on its own. All rules are accompanied, in the dedicated database, by tags which define their type (like warnings, strong errors), their status (such as in use, disabled, testing stage) and few other attributes.

We emphasize again that statistics, consistency and redundancy of rules can be systematically managed by using the R-packages cited above.

R-code for the translation of rules from sql into R

The set of sql – queries which define the validation rules need to be translated into R, in order to be able to apply the suite of R - validation packages listed above. An added bonus would be the possibility of translating the rules to/from VTL, since there are examples of VTL-into-R translation code in an advanced stage of development[7].

Differences between basic SQL and R syntax are relatively simple to map. See for example the R “flow == X“ and the sql “FLOW=X” conditions, or operators like the sql “IN” and the R “%in%”. Same simplicity is found when mapping statements involving “NULL” in sql, which are re-phrased into “is.null==TRUE” (or FALSE) in R.

The most difficult translation concerns *composite* sql queries, where two or more queries are nested. For example, “select … where country not in (select country from lookup_table)”. Since the validate package allows for auxiliary data frames, we have exploited this feature. Therefore, we work in a step-wise manner: first the *included* (in the composite) queries are translated and associated with lookup data frames, then the *including* queries of the corresponding composite are translated and are referring to the newly created lookup data frames. We make full use of dplyr and tidyverse R

[7] https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_15_Technical_aspects_of_VTL_to_SQL_translation.pdf

ecosystem for an efficient mapping of the sql syntax into R and for keeping the code easy to read and maintain.

5. Examples of validation tests

We show here an example of the output from the R-package *validate*, which was applied to a test set of rules and a real data set.

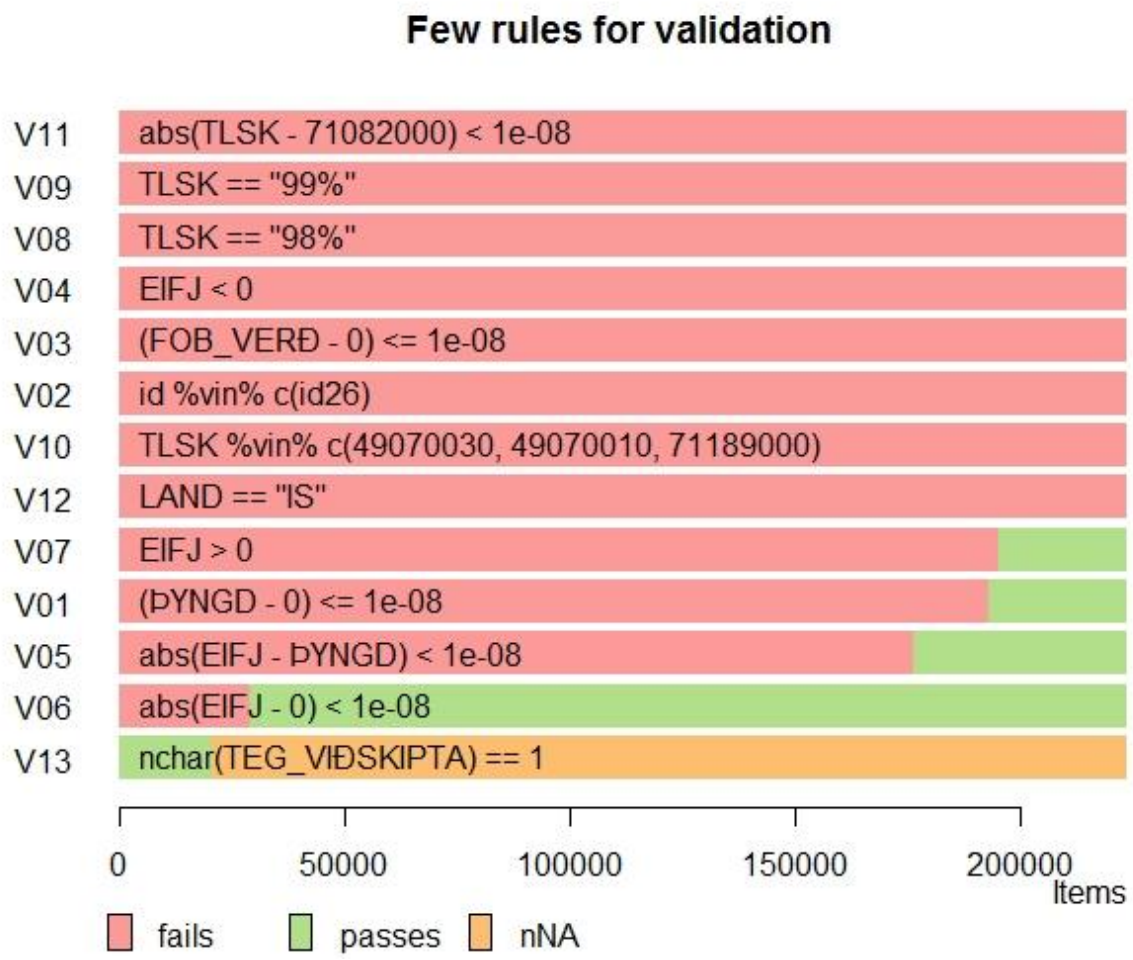

Figure 1 Example of data validation test

Table 1 Detailed results of data validation in Figure 1

name	items	passes	fails	nNA	error	warning	expression
V01	224054	30939	193115	0	FALSE	FALSE	(ÞYNGD - 0) <= 1e-08
V02	224054	0	224054	0	FALSE	FALSE	id %vin% c(id26)
V03	224054	0	224054	0	FALSE	FALSE	(FOB_VERÐ - 0) <= 1e-08
V04	22405	0	22405	0	FALSE	FALSE	EIFJ < 0

	4		4				
V05	224054	47461	176593	0	FALSE	FALSE	abs(EIFJ - ÞYNGD) < 1e-08
V06	224054	195320	28734	0	FALSE	FALSE	abs(EIFJ - 0) < 1e-08
V07	224054	28734	195320	0	FALSE	FALSE	EIFJ > 0
V08	224054	0	224054	0	FALSE	FALSE	TLSK == "98%"
V09	224054	0	224054	0	FALSE	FALSE	TLSK == "99%"
V10	224054	1	224053	0	FALSE	FALSE	TLSK %vin% c(49070030, 49070010, 71189000)
V11	224054	0	224054	0	FALSE	FALSE	abs(TLSK - 71082000) < 1e-08
V12	224054	193	223858	3	FALSE	FALSE	LAND == "IS"
V13	224054	20315	0	203739	FALSE	FALSE	nchar(TEG_VIÐSKIPTA) == 1

Figure 1 represents the number of records passing or failing a test set of rules such as: supplementary units and standard units of measurement coincide, country codes satisfy certain conditions, custom tariff numbers satisfy some constraints, nature of transaction should belong to the values in a given lookup table.

Table 1 shows that for these rules there are no errors or warnings. That means that the rules involve fields which are found in the data and therefore they can be used for validation. The total number of records of the test data is 224,054. The tolerance in coding equality of quantities is 1e-8.

For instance, from the total number of records, 30,939 do pass the rule V01, which is testing the fact that the quantity of those goods is zero. Another rule, V03, refers to the FOB value of goods which is tested against zero. The result of such test is that all records fail, i.e. they do not have a zero FOB value, which is in fact the desirable result.

The example illustrates how efficient it is to perform validation tasks by using the newly designed error detection process. The only conditions are to create a file of rules and apply them to the data set. Compiling and improving the rule collection is a non-trivial task, but with the recent advances in sharing code, information and

translating between SQL, R and even VTL (see for example the last two papers in the list of references), such repositories are easier to improve.

The companions of the *validate* R-package make it also very easy to verify the consistency of a given set of rules, to locate the errors, impute and analyze the results.

Conclusions

In this paper we described the process of reviewing and re-designing the production system of foreign trade statistics, with a focus on the error detection stage. This, as well as the imputation and advanced validation are made automatic to a much higher degree than the older system employed at Statistics Iceland. The new editing process is based on a flexible collection of validation rules and on a suite of open source packages, some of which are newly built for the purpose. This guarantees continuous improvement via sharing and peer review of such products, as well as reproducible and analysable results. Our new production system becomes much more automatic, reliable, flexible and share-able within ESS:

6. References

Compilers Guide on European statistics on international trade in goods by enterprise characteristics (TEC). Available at:

https://ec.europa.eu/eurostat/documents/3859598/9546348/KS-GQ-19-001-EN-N.pdf/6f64a2c4-f4f7-4005-94f1-82ca2d4f531b

Van der Loo, M., De Jonge, E. (2018) Statistical data cleaning with applications in R, J.Wiley, N.Y.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011) The editing of statistical data: methods and techniques for the efficient detection and correction of errors and missing values, Statistics Netherlands, Discussion paper (201132).

Fellegi, I. P. and Holt, D. (1976) A systematic approach to automatic edit and imputation. Journal of the American Statistical Association, 71, 17–35.

Statistics Sweden (2006), Background Facts, Economic Statistics 2006:3, A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics.

Statistical Office Poland (2017) Technical aspects of VTL to SQL translation, Conference of european statisticians, The Hague. Available at:

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg2/Paper_15_Technical_aspects_of_VTL_to_SQL_translation.pdf