

Scanner Data: Initial Data Testing

*Meeting of the Group of Experts on Consumer Price Indices
Geneva, Switzerland, 2 - 4 May 2016*

Heiðrún Erika Guðmundsdóttir, *Statistics Iceland*, erika.gudmundsdottir@stative.is
Lára Guðlaug Jónasdóttir, *Statistics Iceland*, lara.jonasdottir@stative.is

*Statistics Iceland
Borgartúni 21a
IS - Reykjavík 150, Iceland
Tel: +354 528 1200
Fax: +354 528 1299*

Abstract

Initialising the use of scanner data in Consumer Price Index calculations is a large and complex project. The focus of this paper is to address the issue of testing data on arrival. Setting up robust testing procedures for received files, even before data are uploaded to a statistical institute's database, is an important step in building a reliable processing system for scanner data. Good data testing upon reception is important to ensure a steady flow of data, minimise downtime of the system and minimise data loss. It also highlights structural problems that need to be fixed by data providers or worked around in the process following. Sound testing procedures should result in better completeness of the data, increase its overall quality and smooth further processing. This paper describes some of the tests Statistics Iceland is working on implementing to strengthen the reliability of the scanner data the institution receives.

Key words: Scanner Data, Data testing, Store changes, Data changes.

JEL: M11: Production Management, M15: IT Management

Introduction

Statistics Iceland has recently started receiving scanner data from three large grocery store chains in Iceland. These three chains account for at least 80% of the Icelandic grocery market (Icelandic Competition Authority, 2015, p. 27). From the beginning the aim has been to build up a solid system for implementing scanner data for the Consumer Price Index (CPI) and the Icelandic Harmonised Index of Consumer Prices (HICP). As can be expected, several obstacles have had to be overcome, despite efforts to anticipate problems in advance. In order to ensure data quality and reliable data delivery, various testing steps have been undertaken. The plan is to begin to use the delivered scanner data on a small scale with a fixed basket approach for the CPI/HICP as soon as possible and thereby stop manual on site data collection in the participating grocery chains. Further development will be carried out to build a system to utilize the delivered data on a larger scale.

Data structure

Statistics Iceland is obligated by law to compile and disseminate the Consumer Price Index, Act no. 12/1995 and the Harmonised Index of Consumer Prices in accordance with the European Regulation no. 2494/95. The Act on Statistics Iceland and Official Statistics no. 163/2007 gives Statistics Iceland extensive authority to collect data to that purpose. Nevertheless, collecting scanner data is more demanding than ordinary data collection for both the statistical office as well as the data providers. Negotiations with the grocery store chains aimed to build up reliable systems to ensure continuity and timeliness in data delivery. The contracts also describe responsibility for data deliveries and how to respond to system failures as well as requirements for the data structure and content. Furthermore the contracts contain an xml-scheme and a detailed sales description which is a general description for every transaction line that is to be transferred.

Statistics Iceland set up a web service where data providers log in and upload daily data files. The web service is automatic for both parties and requires little or no manual control when everything is in order. This is essential to keep the response burden as low as possible and reduces the risk of data flow interruptions due to absence of employees.

The provided data contains daily sales values for all goods, including quantities and explanatory data such as product numbers, short item descriptions etc. In addition data providers were asked to include their own classification of the goods. The data is broken down to single stores within the chains.

Data reception

All data arrives through the web service in xml data files. The files are built up with a scheme-location and headers that identify the data provider and the stores. Data is received daily but with a few days delay. The delay is for the data providers to reconcile their sales from cash registers all over the country and balance them before the data is delivered. Depending on the data provider it takes about 2-5 days to deliver complete data for all the stores.

Data processing

The first tests are applied to data on arrival. The contracts with the data providers call for daily deliveries and hence delivery is monitored daily. One file per store is delivered every day, i.e. as many files from each data provider as the number of their stores. Every step of the receiving process is logged to ensure traceability. The testing process is now a three step process, where the first step monitors that data is being received and completely transferred. The second step checks the completeness of data, e.g. number of files, whereas the third step deals with the content of the data files.

Step 1: Reception tests

The first step is to monitor whether data is being received at all. This is essential to the early detection of failures in the data transfer processes. Analysing why processes fail can be time consuming. The same goes for the time it takes to restart them and therefore it is helpful to receive data every day in order to minimise down time and shorten periods of missing data.

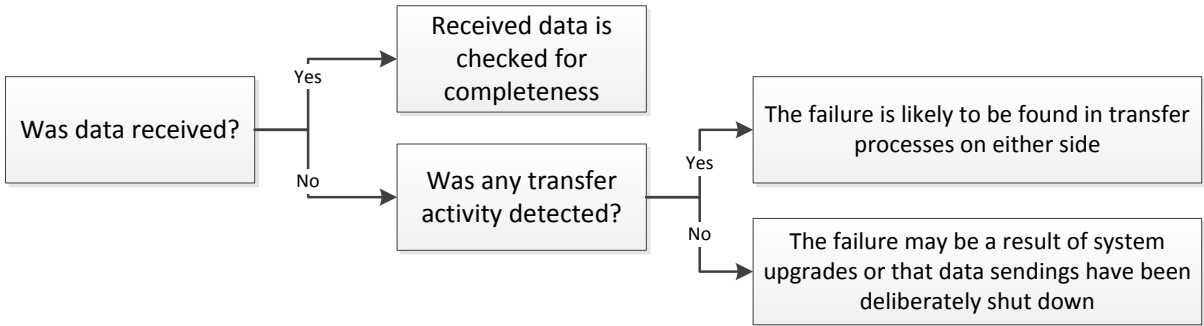


Figure 1 Reception test 1 – Analysing the reception process. Transfer failures may have various reasons. An analysing test helps to suggest failure reasons with the aim to shorten the recovery time.

If data is not received it may be because no attempt has been made to deliver it or because the transfer has failed. Either way; if the reason behind a system failure is logged and analysed, a list of possible solutions could be constructed and used to strengthen the transfer structure, or at least shorten the recovery time.

Other tests in this step include validating the delivery metadata and monitoring packages that move data files into staging areas for further examination.

Step 2: Tests for completeness of data transfer

When data receptions are successful, the next step is to test the actual data files. Here we focus on the store level metadata since one file is delivered per store. Beforehand one would think that the list of stores that each data provider delivers data for would be rather stable. In the short period Statistics Iceland has received scanner data this list has taken several changes where branches have closed down and others opened or supermarkets have been changed to discount stores. Although changes in the number of stores cannot be called common it is an occurrence that needs to be accounted for.

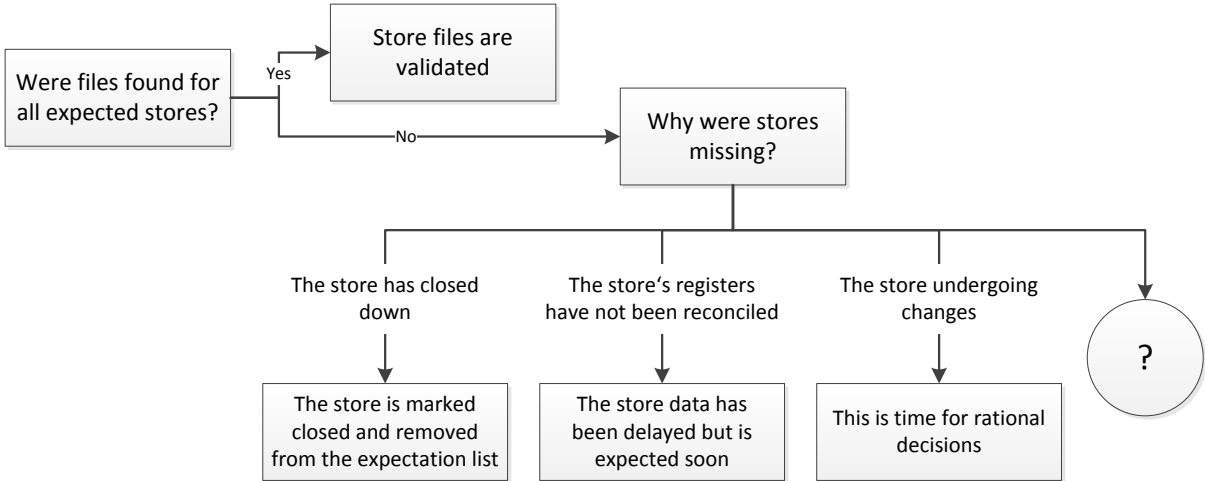


Figure 2 Reception test II-a – Completeness of data delivery. Reasons for stores missing from the data can be diverse but not necessarily a cause for alert. The appropriate reaction depends on the data provider.

There can be various reasons for stores being missing from the data delivery. The reason can be as simple as that one store was not able to finish returning reconciled register data to the data provider in time for the delivery to the statistical office. The expected delay of data from each data provider is a known figure even though it may differ between them. Monitoring delay days for each provider and checking if data delivery for one or more stores is deviating from the expected pattern can be used to raise alerts. Daily reports of the ongoing process are sent to the supervisor of the system. This includes a list of expected stores with statistics about their progress and a list of missing stores with statistics about their absence.

Store / Chain		- 8 days	- 7 days	- 6 days	- 5 days	- 4 days	Today - 3 days	60 day average
Townsquare / Chain 1	Number of data lines	4,363	5,574	5,972	4,514	3,884		4,681
	Turnover (ISK)	7,806,536	13,043,549	13,714,540	8,495,521	6,428,770		9,454,048
	Quantity	25,843	42,715	44,246	28,648	22,099		30,404
Birdstreet / Chain 2	Number of data lines	1,175	1,477	979	690	1,309	1,274	1,185
	Turnover (ISK)	945,400	1,149,798	633,668	399,460	987,293	1,037,490	862,058
	Quantity	3,148	3,786	1,999	1,402	3,275	3,195	2,869
Fishmarket / Chain 3	Number of data lines	885	1,062	897				849
	Turnover (ISK)	769,999	985,504	720,249				714,978
	Quantity	2,180	2,704	1,980				2,086

Figure 3 Example report on the ongoing process.

The same system alerts if unexpected deviations appear. If a store is missing for a few days but other stores from the same provider are being delivered this is not an immediate cause for alert, however if the store's absence is prolonged it is a cause for action. This could be a sign of the store being closed down or undergoing some other changes. In Iceland the grocery chains often use their websites to announce changes in their stores or the changes are reported in the media.

Store	Last data transmission	Number of days since last transmission
Store 1	3/21/2016	9
Store 2	3/21/2016	9

Figure 4 Example of an alert due to missing store data.

With one store or a small proportion of stores missing the price index can still be compiled but the problem should still be addressed and prevented from accumulating. If on the other hand many stores are missing this could be a sign of some malfunction in the delivery system that need to be addressed without delay.

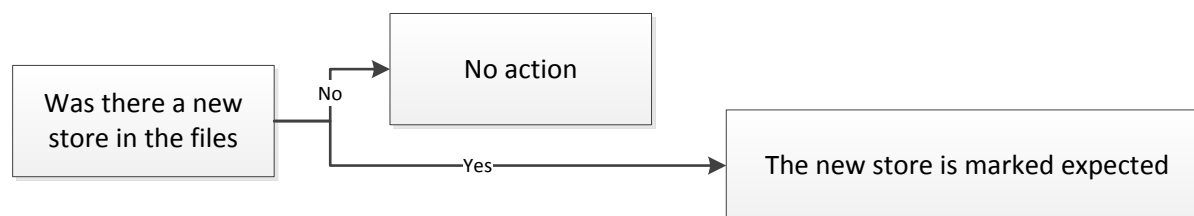


Figure 5 Reception test II-b – New stores.

Receiving a new store is a good sign of the system working the way it should. The system would need improvement if an opening of a new store has been announced but the store's data is not delivered. If a new store appears in the data, the store is added to the list of stores and marked an expected

store for future reference. Checking the chain’s website or contacting it for information about the new store in order to confirm its validity would be advisable.

Step 3: Internal file tests

After the data files pass external checks the next step is to do internal checks. The size of each store file can vary from day to day in accordance with how much data it contains. The amount of data is related to the variety of products being sold every day; however it may be expected that this variety changes in a predictable manner, perhaps a seasonal one.

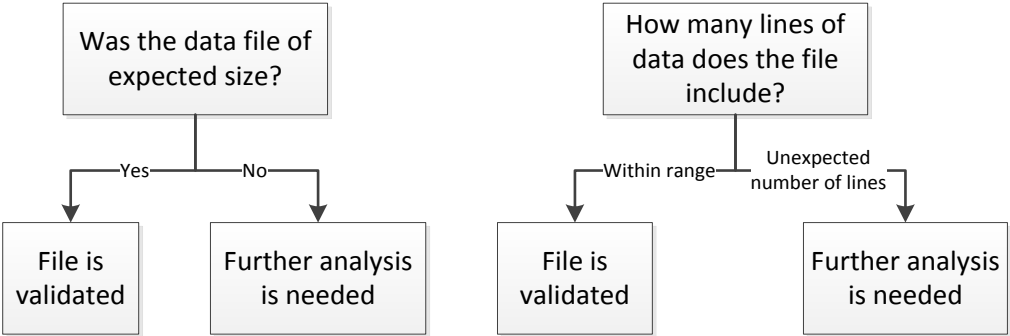


Figure 6 Internal file tests for file sizes and number of lines

The purpose of the file tests is not so much to follow how the product variety changes with season, but rather to detect if a faulty file has been sent. A file that is twice the size as the file from the previous day is likely to contain duplicate sets of the sales lines. On the other hand, if a file contains only a small number of lines compared to the file from the previous day, this might indicate missing data lines or missing supplementary data. In both cases the expected ranges are likely to vary both daily and seasonally. One approach to compensate for normal variations would be to define upper and lower limits for the expected file sizes and data lines as deviations from the measured averages of the same figures accumulated over a period of time. The limits should at minimum account for variations in stores’ opening hours. A variation of these tests is to ask for the aggregated turnover and quantities presented in the file. If the file tests return files that exceed the limits, they are flagged for further analyses. If the data seems in order after inspection it might be an indication that the expected ranges need to be expanded. If the inspection, on the other hand, returns some faults in the data this calls for a system search either by the statistical office or by the data provider to find out what causes the file corruption. Corrupted files are not allowed to enter the databases and faults should be dealt with without delay.

Tests for further data analysis

When data files for the stores have been validated it is time to test the data lines for the sales data.

All lines should include the following variables:

- Date of sales
- Barcode (EAN)¹
- Price lookup code (PLU)
- Store's own product number (OWN)
- Product description
- Unit of sale²
- Category ID³
- VAT rate
- Value
- Quantity⁴

All products are supposed to be identified with at least one of the three identifiers: Barcode/EAN, PLU or store specific product number which is here called OWN for short (Gudmundsdottir and Jonasdottir, 2014, p. 4-5). An EAN code is commonly used and most groceries have one whereas a PLU code is less common but generally used for products sold by weight that often do not have an EAN code, such as fruits and vegetables. Consequently the EAN, PLU and OWN numbers can all be empty in the data file but in each line at least one of them has to have value and this is used as the first test on contents. For example data files would be considered faulty if an unusually high number of lines had all three key variables empty. Where only a few such lines exist in the file, the lines are thrown out but the rest of the file passes the test.

Statistics Iceland is now working on the validation process in this stage and learning about the tests that apply. Other conditions that can be suggested are for example, unusually high number of lines with negative sales values or quantities and no lines with barcodes, to name a few. What is becoming clearer with every step is that few conditions seem to be black or white. Various anomalies can be found in otherwise valid data and therefore margins of error need to be defined.

Comparability of products by description

As mentioned in the previous part, all grocery chains are asked to provide a product number with each product. These numbers should be unique and consistent in time; however, the products can be

¹ The Icelandic data providers do not use GTIN numbers that are the next generation of EAN numbers.

² The unit refers to the unit that the store uses to sell the product, e.g. piece, kilogram, etc.

³ The sales data lines contain a category ID from the store category classification if available. The information on the classification is delivered in the same data file but in a different part of the XML-scheme. The store category is used only to support coding new products with a regular consumption classification.

⁴ The value and the quantity refer to the total amounts that have been sold in one day.

labelled with more than one label. It appears that the stores do all have their own product numbers for all their products and in general they also have barcodes for most of their products.

OWN	EAN	PLU	Around
X	X		98%
X	X	X	1%
X		X	<0.5%
X			<0.5%
Other combinations			0%

Figure 7 Combinations of labelling numbers seen in the first Icelandic scanner data lines

About 1% of products are labelled with all three identifiers and about the same share is labelled either with store’s own product number only or in combination with a PLU code. Despite the requirement that numbers should be unique and consistent there are some deviations that occur. These deviations can probably be attributed to the fact that the stores’ product and pricing registers primary purpose is of course for the stores and not for statistics.

The OWN comes closest of the stores’ labels to being a unique identifier. Yet the OWN is not fully consistent for statistical purposes. It follows that in order to ensure consistency in compilation there is a need to issue an NSI-ID (an ID issued by the statistical office) for the products that is consistent for statistical purposes, namely to compare prices of alike products at different time periods. There are several types of inconsistencies that can arise and their nature determines how they should be treated.

Example 1 – Typing changes

time	OWN	Description	NSI - ID
T1	010014	Grocery bag.	Same product NSI - ID
T2	010014	Grocery bag	can stay the same

An inconsistency does very often appear where a description of a product has been slightly altered, maybe just in order to fix a typing error or just to be trimmed a little. Changes of this kind are inconsequential and can easily be ignored. Procedures can be written to check a newly arrived description string against a current expected string and if the difference is negligible, the product is considered the same as before and keeps its current NSI-ID. This requires that a tolerance level is set for how much a description string may diverge from the old one without the product being considered a new product.

Example 2 – Quantity changes

time	OWN	Description	NSI - ID
T1	23400	Brown bread, 500 gr	Should the ID change?
T2	23400	Brown bread, 400 gr	

An inconsistency where the change in description reflects a resizing of a product is quite common within some chains. The impact of such resizing practices has e.g. been described in an article on Product downsizing and hidden price increases (Imai and Watanabe, 2014). These cases can easily go by without being noticed, since the chains may not bother to change the OWN number. To be fair, a change of this kind could also easily go unseen by a person who is manually collecting prices since changes in package sizes are often so minuscule that the common consumer may not notice them during their shopping. To counteract this it is possible to produce procedures that scan for quantities in a string and compare those parts separately from the rest of string with the current string and quantity measure. The most straightforward way would be to treat this case a new product and assign a new NSI-ID, but as in example 1 it might be worth considering setting up a tolerance test that allows small quantity changes to be overlooked. This would probably depend on the availability of data for the particular consumption class (COICOP).

Example 3 – Ambiguous changes

time	OWN	Description	NSI - ID
T1	33400	White bread	Has the product
T2	33400	White bread, 400 gr	changed?

This third example could be viewed as a special case of the second example. Here there description changes from giving no information on package sizes to specifying the size. Without additional information to compare the two descriptions, this product may have to be treated as a new product and assigned a new NSI-ID.

Example 4 – Stores reuse OWN numbers

time	OWN	Description	NSI - ID
T1	503	Washing powder	A completely different
T2	503	Carrot Cake	Product that needs a new NSI - ID

This last example describes a case where the chain has chosen to reuse the OWN number. Luckily this case is likely to be found by the same procedures that find the typing changes (example 1). The

new string completely fails the tolerance test and therefore the product is assigned a new NSI-ID before being considered valid for index compilation.

These are various examples of product inventory changes as they appear in the data and how they can affect the sample of products that in the end are chosen to represent the price level measurements.

What have we learned so far?

Stores change

It is more common than originally anticipated that the grocery chains change their stores in their attempts to reach the consumers, e.g.:

- a new store opens within a chain or a store closes down
- a store changes types, e.g. a supermarket becomes a discount store
- grocery chains try to offer new store brands with different price levels or product varieties
- grocery chains change business practices and start running chains as separate entities

This calls for our attention to monitor consumer responses to the new setup, as well as possible effect of clearance sales or other unusual circumstances. Often the chains' websites are good resources of information but the first sign of changes can sometimes be found in the scanner data.

Data transfers fail – Recovery is time consuming

No matter how well a system is designed it is always prone to failure. Therefore it is essential to have processes in place to quickly recover the systems. These processes include:

- Quick detection of problems
- Effective communication
- Clear understanding of who is responsible for rectifying the problem and who shall bear the cost
- Clear understanding of the importance of acting quickly to recover data transmission
- Reliable documentation

As the chains' preferred setup of data warehousing is with an external service provider, all correspondence for receiving data and for recovering problems is with a third party. This slows down the recovery process.

Due to the daily frequency of data delivery, Statistics Iceland has to maintain a daily routine to check for anomalies in order to minimise the response time if problems occur.

All contracts clearly state the companies' responsibility to keep the data delivery systems up-and-running and address all problems without unnecessary delays. However, if a problem is detected on a non-office-business day, e.g. on Sunday, Christmas day etc., the allowed delay in the data delivery gives the grocery chains slack to address the problems on the first following working day. The key to handling this appropriately is communication with data providers and Statistics Iceland is more than willing to assist if possible. It is essential for the overall progress that the data providers understand and respect their obligation to respond quickly to all system failures. The data providers also have to communicate this understanding to the IT companies that service their data warehouses. It would be wise to keep track of the recovery time following a failure in the system and use the information for improvements.

It is extremely important to invest in good documentation from the beginning. Even though the systems are supposed to run automatically, changes outside the system may cause failure. Without proper documentation it may be hard and time consuming to fix.

Have a back-up plan in position

Although the grocery chains have an obligation to deliver the data, they may fail to do so for various reasons. Enforcement is likely to be more time consuming than the interval between index disseminations.

The back-up plan is likely to resemble the manual collection method on a smaller scale. Therefore it is important to keep the old processes functioning and practice their use on regular basis.

Deliberation on data collection frequency

Statistics Iceland is collecting data on a daily basis from the three largest grocery chains in Iceland. The amount of data that is received every day is considerable. The upside of this collection frequency method is that it is possible to identify problems in data delivery more quickly than with a lower collection frequency. By collecting daily, effects of special days that have unusual opening hours can be distinguished from the more normal days. In addition, no problems arise due to weeks that are short or overlap with two months. However; it should be taken into account that the total number of stores in the three chains in the country is somewhere between 100 and 200 and in many cases a different collection frequency might be more appropriate.

Next steps

Statistics Iceland next step is to replace the current manual data collection process in the grocery stores by using scanner data with a fixed basket approach. Comparisons between the manually collected prices and the samples from scanner data have been promising. In this step we will only

alter the channels by which we receive the price measurements. The calculation methodology for the index is not changed in this step. However, using scanner data, albeit using only a small number of the products found therein, enables Statistics Iceland to use prices from all stores of the grocery chains where manual prices have been collected. Also it allows for an increase in the basket of products since the basket can be tailored to the most sold products in each store. This step is scheduled for implementation with the annual rebasing of the CPI in April 2016. Sales quantities will not be used at this step except for choosing products for each chains basket.

Statistics Iceland is working on setting up a reliable back-up plan for the data collection as there are many risks involved in relying solely on an uninterrupted reception of scanner data.

Following this Statistics Iceland will set out to develop a processing infrastructure for scanner data that enables the use of more of the scanner data mass than the fixed basket approach does.

References

Guðmundsdóttir, H. E. and Jónasdóttir, L. G. (2014). Scanner Data; Optimal Preservation Policy for Identifiable Datasets

Icelandic Competition Authority (2015) Leiðbeining um samkeppni á dagvörumarkaði – Staða samkeppninnar 2015. *Ritröð Samkeppniseftirlitsins*, Rit nr. 1/2015 – Skýrsla Mars

Imai, S. and Watanabe, T. (2014). Product Downsizing and Hidden Price Increases: Evidence from Japan's Deflationary Period. *Asian Economic Policy Review*, 9: 69–89. doi: 10.1111/aepr.12047