Expert Group meeting on CPI 26-28 May 2014	**One size fits all? The need to cope with different levels of Scanner Data quality for CPI computation.**
Workshop 4. Scanner data	
Presentation Topic 3	
Abstract	Considerable resources need to be applied for classifying and coding scanner data. The task to make scanner data ready for CPI compilation might be covered by retailers, NSIs or outsourced to market research companies. Statisticians interested in using scanner data for CPI compilation should be aware that the data cleaning and data processing requirements before actual index calculations are considerable high. This paper highlights the range of scanner data characteristics from different sources and points at the effects of the scanner data quality on the methods that can/must be chosen to treat the data. There is a danger of using resources inefficiently when building up a (or worse: several) system(s) to process scanner data for CPI index compilation.
Contact	Ingolf Boettcher STATISTICS AUSTRIA Directorate Macro-economic Statistics Prices and Purchasing Power Parities Guglgasse 13 1110 Vienna phone: +43 (0) 1711 28-7917 mail: ingolf.boettcher@statistik.gv.at

Characteristics of scanner data

There is no ISO standard which requires retailers to structure their point of sales data and internal product classification system in a harmonized way. Instead, most retailers have built up their own systems and classifications which often support the organizational structure and strategy of the retailer. Classification systems might be organized according to marketing and organizational principles, respectively, rather than according to classification principles of product purpose (e.g. in our scanner data we found ice cream bars assigned to internal product groups called “cash point zone products”, “frozen food” and “ice cream”). Depending on the retailer, the structure and characteristics of the scanner data of retailers might be still based on 1980s database conventions with very few characters and digits per variable due to restricted storage place.

Scanner Data used by Statistics Austria

Currently, Statistics Austria deals with two kinds of scanner data sets. On the one hand, we test neatly formatted market research data from AC Nielsen, on the other hand, we are building up processes for "raw" scanner data of a medium sized Austrian retailer[1].
Table 1 compares the scanner data characteristics of the two Austrian scanner data providers. The example values describe for both providers the same article (example): a SPRITE non-returnable 1 Litre PET bottle. Obviously, the scanner data information about this article is structured very differently in both datasets.

Table 1 *- Characteristics of the Scanner data available to Statistics Austria*

	AC NIELSEN			Austrian Retailer			
Data Structure	**Variable**	**Format**	**Example value**	**Variable**	**Format**	**Example value**	**Available in both data sets + same format?**
Characteristics	City	Text	Vienna	Store ID	Numeric		No
	Manufacturer	text	Coca-Cola	-			No
	Brands	text	Sprite	-			No
	With CO2/ without CO2	text	'With CO2'	-			No
	Single use / returnable	text	Single use	-			No
	Flavour	text	Citron	-			No
	Type of packaging	text	PET	-			No
	Volume per unit	Alpha-umeric	1000ML	Volume per unit	Numeric	1000	No!
	Volume	Alpha-umeric	2000ML				
	Multipack	Alpha-umeric	1 er	-			No
	Description	Text	SPRITE TM NA ZITRONE LL 1000 ML PET1 ER NA	Description	Text	Sprite , 2x1,0 L Pet.	Yes
		-		Unit	Text	ML	No
				Unit of Sale	Text	ST (Piece)	No
	Class of goods	Text	CO2 LIMONADE	Class of goods	Numeric	23413	No
	Segment	Text	NON-COLA	-	-	-	
	EAN	-	-	EAN	Text	3046920028370	No
Facts	Sales in Euro	Numeric	28	Sales in Euro		28	Yes
	Sales in units	Numeric	34	Sales in units		34	Yes
	Sales in Volume	Numeric	34	Sales inVolume		-	No
Period	Calendar week	text	201202	-			No
	-	-		Start Date (YYYYMMDD)	Numeric	20131229	No
	-	-		End Date (YYYYMMDD)	Numeric	20140104	No

[1] For a detailed description and analysis of the Austrian scanner data project, see the report in the 'relevant papers' section of the workshop webpage:
http://www1.unece.org/stat/platform/display/CPIworkshop/Workshop+4.+Scanner+data

Thus, a CPI compilation system has to cope with diverging data structures of the retailers and scanner data providers. In the two scanner data sets available to Statistics Austria, only three variables exist in both data sets and have the same format (product description, Sales in Euro, Sales in Units).

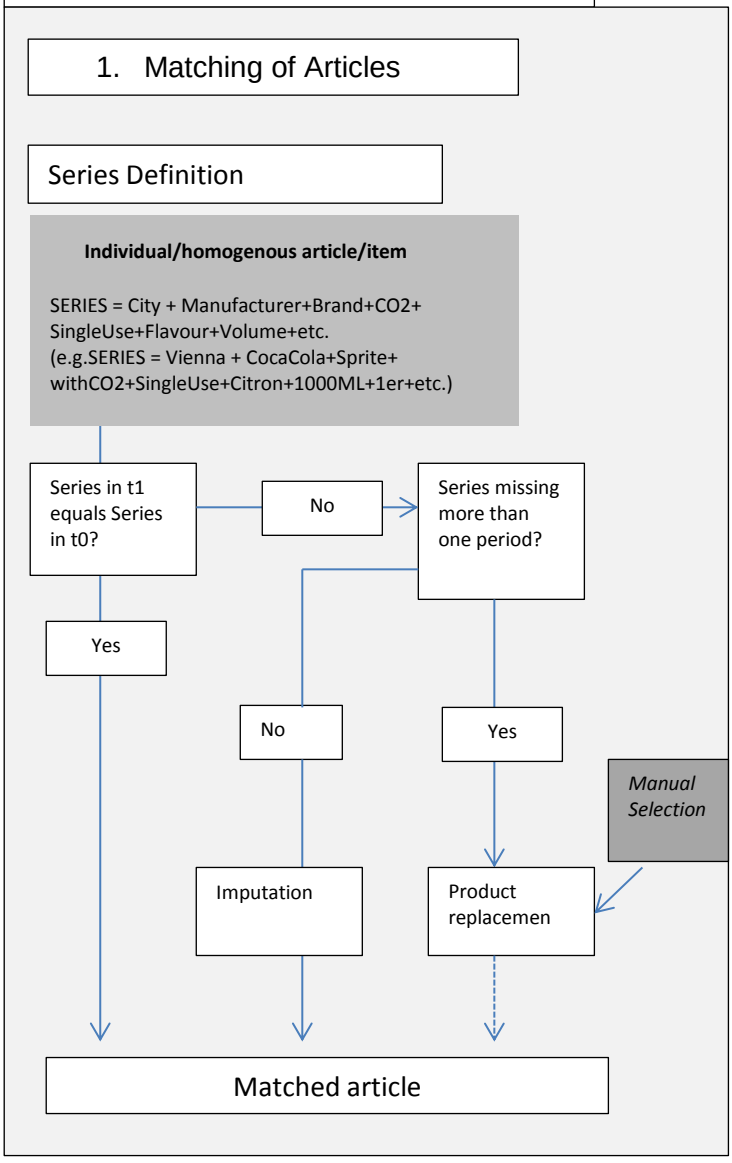
There are two main tasks to achieve when processing scanner data after receiving it from retailers and before CPI compilation: matching individual articles between time periods and assigning/mapping individual articles (EANs) to a CPI elementary aggregate (EA) and COICOP (sub-)class.

Correct matching of individual articles between two different time periods;

At first glance, the EAN-Code seems to be the obvious choice for the correct identification of individual articles in different time periods. However, one of the major problems concerning scanner data is that EAN codes of articles change over time, e.g. when a product re-launch occurs[2], the place of production changes, internal changes of the manufacturer take place, etc. In the data sets available to Statistics Austria, EAN attrition rates are about 35% over a 5-month period. Other studies find attrition rates of up to 45% in a one year period.[3] Thus, product characteristics might be taken into account when identifying identical or at least homogenous articles over time.

Charts 1 and 2 depict the two methods used in the Austrian scanner data project to match articles over time for price index compilation. When using scanner data from the *market research company ACNielsen*,

Chart 1: ACNielsen – Article Matching



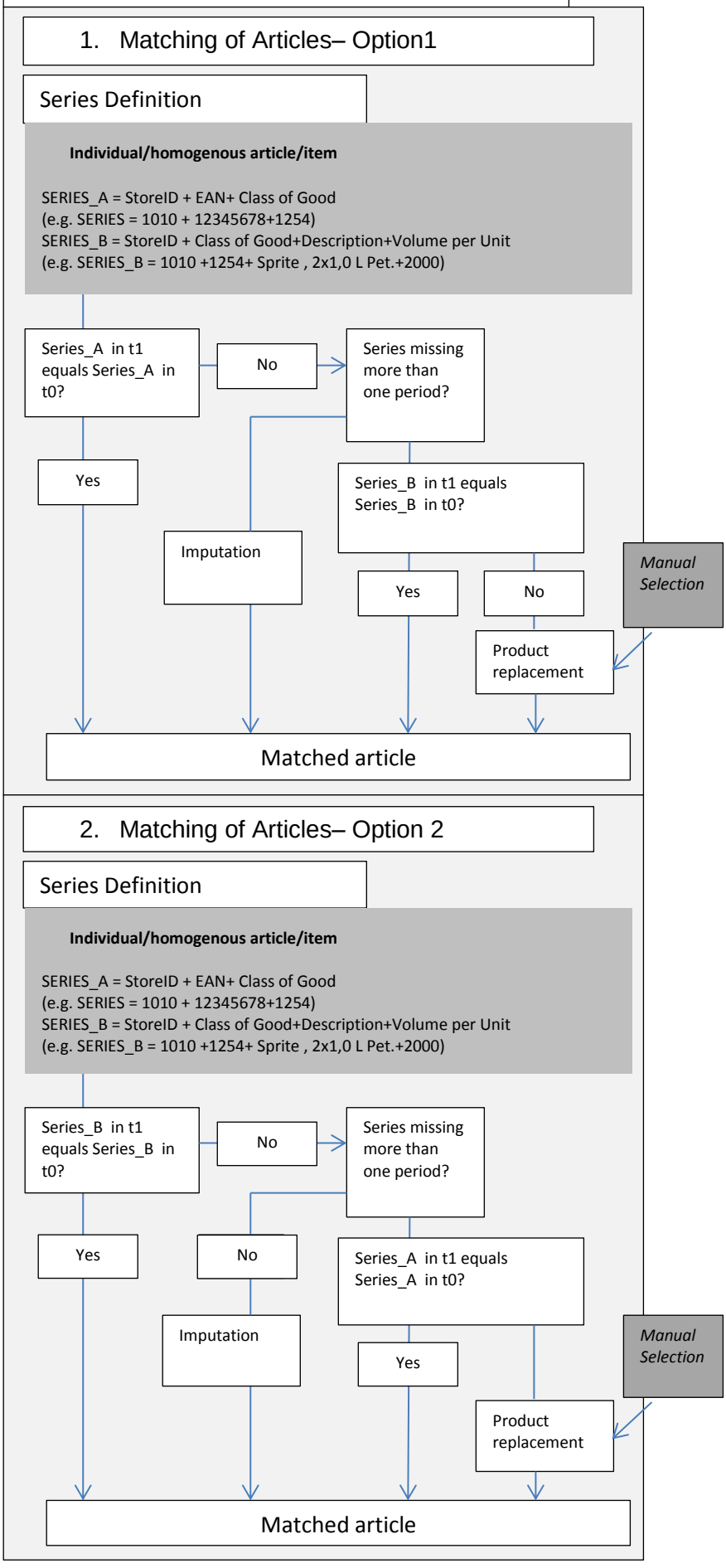
[2] See Antonio G. Chessas study „Comparing scanner data and survey data for measuring price change of drugstore arcticles“:
http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_sem_lista&tipo=r&detalhe=165101941

[3] Faivre, Sébastien(2012) INSEE, 'Would scanner data improve the French CPI?' P.4
http://www.scb.se/Statistik/PR/PR0101/_dokument/Would%20scanner%20data%20improve%20the%20French%20CPI.pdf

the definition of precise product characteristics into a product key is very easy. In that case, using the EAN-code is not necessary as eventual EAN attritions would influence the selection of identical articles.

The data set provided by an *Austrian retailer* provides less information for the task of article matching. The EAN code, Store ID, and the internal number of 'Group of Goods' allow for the computation of a series that identifies a unique article. In case of EAN attrition, an alternative Series_B is checked. It includes the product description and the volume per unit to identify an individual article but exludes the EAN code. In Option 1 the Series_A is matched between time periods before checking whether there are matches for Series_B. In Option 2 the opposite sequence is applied with Series_B being matched before Series_A. Both options have advantages and disadvantages. In Option 1, the replacement of an articles identified by EAN might take place too late (e.g. if a re-launched article with new EAN appears). In Option 2, the matching process does not work in case of insignificant format changes and is therefore prone to errors.

Chart 2: Austrian Retailer – Article Matching

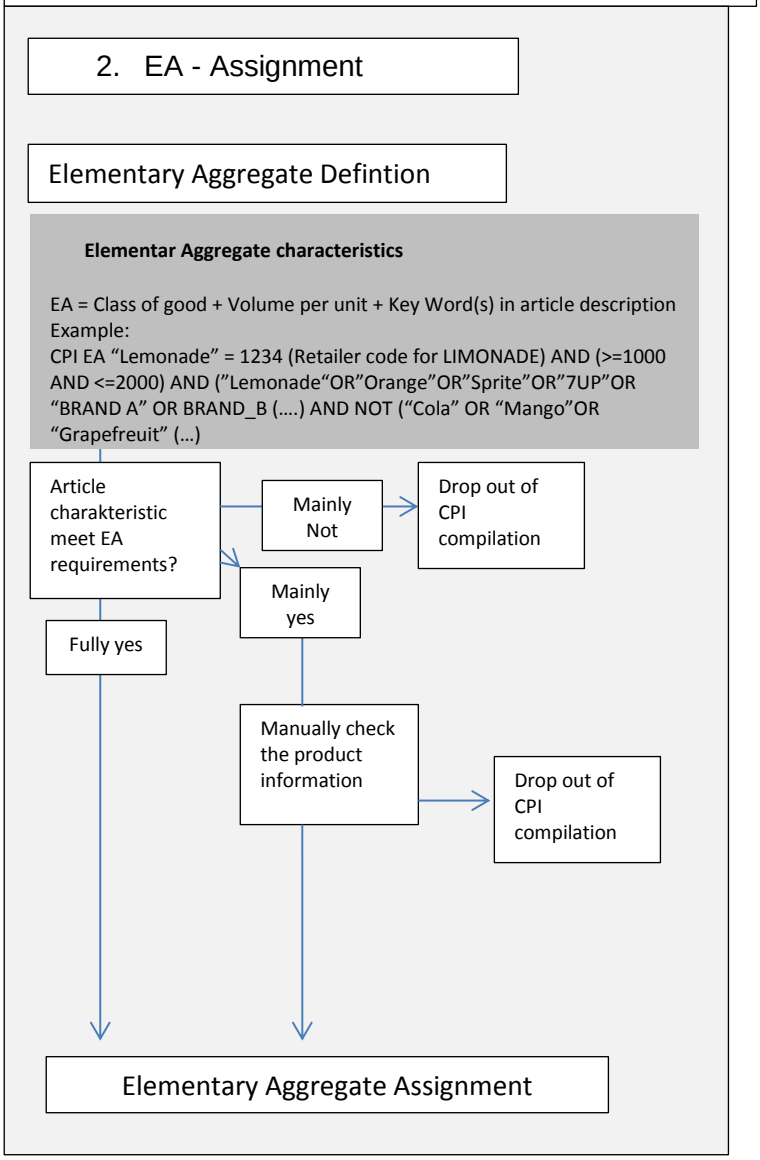


Correctly filter and assign individual articles to their respective Elementary Aggregate (EA) and COICOP class (EAN-Mapping).

The data set by ACNielsen provides many detailed product characteristics to allow for a relatively easy attribution of the articles to the target EAs and target COICOP classes. It is the main advantage of market research scanner data that it includes a full set of variables with article characteristics (See Chart 1). In the Austrian test-project, the market research scanner data from AC Nielsen includes sufficient characteristics to precisely filter and assign articles which complied with the internal CPI product descriptions for elementary aggregates.

The scanner data available to Statistics Austria from an Austrian retailer requires much more sophisticated data processing. Chart 3 depicts the method currently worked on in the Austrian scanner data project. Every article is being processed and filtered according to pre-defined elementary aggregate characteristics. In order to assign a scanner data set to an EA the internal 'Class of Good' must be correct, the 'Volume per unit' must range according to the specific Austrian CPI product definition and certain key words can be /cannot be part of the product description variable. Naturally, the resources needed to build up such an elementary aggregate assignment process are considerable large. Especially, the key word list has to be filled and maintained at considerable costs. Advanced statistical programming is required to perform the data management efficiently (e.g. spelling errors should be taken into account using 'like'-Functions).

Chart 3: Austrian Retailer – Article-Mapping to COICOP/EA



There are other options to solve the problem of insufficient article characteristics when mapping scanner data to elementary aggregates and COICOP (sub-)classe. EAN-Dictionaries may be obained from market research companies, hereby upgrading the rudimentary product information in the original scanner data from the retailers can with additional product characteristics[4]. Also, retailers might be willing to build up and maintain a harmonized classification system based on COICOP. However, retailers usually want to avoid any additional data management tasks.

Table 2 lists the different options currently used by NSIs to assign article scanner data sets to an Elementary Aggregate and COICOP (sub)-class, respectively. (Please note that the list is neither complete nor absolutely precise as many NSIs use a mixture of methods when working with scanner data.)

Table 2 Assigning Scanner Data sets to Elementary Aggragete / COICOP (sub)-class

Article Mapping Method	**Description – Advantages – Disadvantages**
Using EAN Dictionary from Market Research Company	Description Merging the scanner data information from retailers with the detailed product characteristics from the EAN Dictionary by EAN code. Advantage -low work-load to assign articles to CPI Elementary Aggregates Disadvantage -costs to obtain EAN Dictionary -not all articles use universal EAN Codes (e.g. private label products that are exclusively sold by retailers) NSIs applying the method INSEE France
Harmonizing the classification with retailer	Description The retailer develops / changes / streamlines in cooperation with the NSI a new or existing classification in order to comply with the products groups of the NSI and with COICOP. Advantage -low work-load to assign articles to CPI Elementary Aggregates Disadvantage -willingness of retailer to harmonize own classification usually low NSIs applying the method CBS Netherlands

[4] INSEE in France purchases EAN documentation (EAN dictionary) from a market research institute.
Statistic Sweden promotes the development of an extensive international EAN database in collaboration with GS1, the producer of EAN Codes.

Reducing the Sample Size and manually assigning articles to EAs	Description Drawing a representative sample of articles with significant turn over . and manually assigning articles to Elementary Aggregates Advantage -good overview and understanding of the used data - Disadvantage - high manual work load - not making full usage of Scanner data potential NSIs applying the method BFS Switzerland; Statistics Denmark; Statistics Sweden
Automatically assigning products using statistical software data management	Description Data management that assigns articles to EAs according to existing information (internal group of products, item description etc.). Pre-requirement: Development of key-word list and data management programming. Advantage -no need to draw sample from the scanner data census -no need to obtain additional article information from market researchcompanies Disadvantage -long implementation time (building up key word lists + re-programming) -high work load needed for quality control and key word list maintenance NSIs applying the method Statistics Portugal and Statistics Austria (both test projects)
Acquisition of Scanner Data from market research companies	Description Scanner data from market research companies include detailed article classifications. Advantage -low work load Disadvantage -high costs -Discounters (Aldi, Lidl) not included -no real supervision of data management procedure possible NSIs applying the method Statistics Austria (test-project

Conclusion

Users of scanner data will be faced with the challenge to make certain choices when it comes to classifying and coding the data for CPI index compilation. This paper highlighted the necessity to keep in mind the potentially large resources needed to integrate scanner data into the CPI compilation process. In the end, the optimal choice will have to take into account different dimensions: availability of financial and human resources, ability and/or willingness of retailers to cooperate on the classification of articles, access to EAN dictionaries with detailed product descriptions. There is a danger of using resources inefficiently when building up a (or worse: several) system(s) to process scanner data for CPI index compilation.