Discussion Paper

Stratification and price index computation

2017 | 20

Leon Willenborg

Content

Summary

Because product populations are dynamic, it is not possible to follow the price development of products at the GTIN level and compute price indices on this basis. In reality each GTIN has a finite lifespan (its time on the market). Products are also renewed by replacing them with others. Sometimes these replacements are simply relaunches of an earlier product, and they are basically as old wine in new bottles. And they come with a new price, typically higher than the product they replace. If all GTINs are treated as separate products price jumps due to relaunches will be missed. In order to avoid this one could try to link relaunches to their predecessors. But this is not so easy as it may seem. Another option is to classify the items on the basis of some common, key features, and consider subgroups of COICOP groups as strata and follow their price development. In order to do this, however, these features should be extracted from the product descriptions available in scanner data or internet data. The present paper discusses which properties stratifications of product populations should have.

Keywords

Product descriptions, attributes, characteristics, feature extraction, stratification, product population, dynamics of product populations, price index computation.

1. Introduction

This report investigates the problem how to stratify populations of products in such a way that they are suited for computing price indices. The first question is: Which properties make a stratification suitable for index computations? It should be stated at the outset that we will not provide the definitive answer to this question here. Our aim in the present paper is more modest. It aims to make some explorations of the problem, and to point out some issues. To make further progress computations on real data are necessary. These are beyond the present paper, and are reserved for future work.

The main problem that we are dealing with when trying to answer the stratification question is the dynamics of product populations. If such a population would be static, the answer is easy: choose the most detailed level, i.e. the GTIN[1] level. But this answer is not the obvious choice in case of a dynamic product population, which one typically encounters in practice.

Why is the GTIN level in a dynamic population of products not a good choice? Items in such a population have a finite lifespan. Items are regularly removed from a market or introduced to it. Comparing the prices of the same product can, of course, only be done during its lifespan, which is limited.

But apart from this limitation, this method is also unable to pick up price changes that are due to relaunches, i.e. similar products that are introduced to the market in order to replace an older product, usually at a higher price. If the product would have the same price during its entire life cycle, producing price indices in the way just indicated, would suggest that no price changes have occurred over some period of time. But the price increases due to the phenomenon of relaunches would have been entirely missed in this approach. So this method is inadequate for producing reliable price indices in the case of a dynamic population.

So the idea may arise to save the method by introducing matches of relaunches to previous products. This would yield chains of prices for what is considered essentially the same product, and the price jumps associated with relaunches would be observed. This is true, but producing the matches is not so easy. Which item or items a relaunch replaces is not reported by a retailer or producer. The statistical office has to produce them on the basis of their knowledge of the situation, which is limited. Given certain characteristics of the items it would be possible to look for similar items. But there is never certainty that in this way the right items are matched. And also, the method is rather laborious. And it yields results that are probably

[1] GTIN = Global Trade Item Number. It can be used by a company to uniquely identify all of its trade items. In Dutch: EAN.

comparable to the next method that we discuss, which is based on stratifying a group of items (such as a COICOP[2]) into subgroups, also on the basis of product features.

These subgroups should consist of similar products and they should show a kind of continuity that is lacking at the GTIN level for a dynamic population. The subgroups will change in composition in due course, but at the aggregate level they should exist all the time, that is, each month. As long as no major changes occur that change the character of a subgroup in a fundamental way[3] this should be acceptable. So the subgroups are composite entities: they exist, although their constituent elements constantly change. So comparing 'like with like' is lifted to a higher level of aggregation (subgroups within a COICOP group), as this makes no sense at the GTIN level, in case of a dynamic population.

Composite entities are encountered frequently in life, and also in statistics. Think of municipalities, schools, factories, etc. Their composition constantly changes over time. At different points in time, different persons live or work there, and the ones that stayed have a different age, or possibly a different home or job title. Schools at different points in time may have different pupils, differently composed classes, different teachers, etc. For factories similar observations can be made. These kind of changes, slowly and with most of the environment remaining essentially intact, are the ones that one would accept as 'natural development' and make direct comparisons of these composites over time, possible and sensible.

But there may also be big changes, that create new entities: municipalities can merge, or 'acquire' or 'loose' major parts of their territory and their inhabitants, etc. A school may merge with another school. Factories may split or may merge with other factories, or they may be bought by bigger companies. In such cases these changes are substantial, and the ‘higher level entities’ involved have been changed considerably. Direct comparison of such entities before and after the changes took place does not make sense, pretending that nothing dramatic has happened and it is 'business as usual'.

So at a detailed level changes are allowed but at the subgroup level there should be stability. But one should be on guard for major changes at the subgroup level. In the present paper we shall assume that at the aggregate level no dramatic changes occur.[4] The question is then which subgroups (strata) to choose. This is the central question of the present paper. Its aim is not to settle it, but, more modestly, to develop some thoughts that go some way in this direction. The present paper only gives some deliberations about how COICOP groups should be stratified in a way that

[2] COICOP = Classification of Individual Consumption by Purpose. A reference classification published by the United Nations Statistics Division. For more information see: https://en.wikipedia.org/wiki/Classification_of_Individual_Consumption_by_Purpose .

[3] For instance when a new and quite different product is introduced. Think of the first smartphones that would suddenly appear in a subgroup ‘telephones’ within ‘consumer electronics’. In such a case a new subgroup ‘smartphone’ should be created and the subgroup ‘telephones’ should exclude ‘smartphones’ from then on.

[4] Nondramatic changes are those that do not cause subgroups to be redefined. This is in fact a change of the underlying classification used.

would be suitable for price index computations. We assume that the strata will be defined by combinations of certain features of the items involved.[5]

So part of the stratification problem is the selection of suitable features (attributes and characteristics of products). For many product populations (e.g. as represented in scanner data) the available features are rather scarce. If there are no sources to enrich them, selection of suitable features from the few available is not a real problem. In some cases (like consumer electronics) there is an abundance of features available from the internet and the selection of a suitable subset is more of a challenge. The aim is to find a limited number of attributes which are powerful predictors of the average stratum prices (for instance). But selection of suitable auxiliary variables to stratify a COICOP group may not be enough. A second step may be required in which suitable categorizations of these variables have to be found. We can also view this as producing aggregates of groups of items within the product populations considered, or as the formation of strata.

However, the present paper is not devoted to these questions, which are best answered by considering concrete COICOP groups, each with their own specific conditions and problems. This is in fact the plan for the near future.

The remainder of the paper is organized as follows. In Section 2 we consider product descriptions and the features (attributes and their characteristics) they may contain in more detail. We discuss the five groups of products from a Dutch retailer. To one of them – pastries – a separate discussion paper (in Dutch) is dedicated, namely Willenborg (2017c). In Section 3 we provide some examples of product descriptions, as they are used by Dutch retail chains. They are intended to illustrate the variety of descriptions encountered in practice, both in form and level of detail. In Section 4 we discuss GTINs in a COICOP group and how to stratify them in such a way that the strata (subgroups) can be used for price index computations. The GTINs are the most detailed level at which products are specified and observed in scanner data. It is argued why the GTIN level is too detailed to be a suitable level for index calculations. GTIN matching is laborious and cumbersome. It is therefore necessary to consider suitable strata / subgroups for the items in a COICOP group. Section 5 continues the discussion in Section 4 and pursues the question how suitable strata should be defined: what aspects should be taken into account? We discuss what distinguishes desirable strata. Section 6 contains a discussion of the main results and insights and has some suggestions for possible future work in the line of the present discussion paper. It also contains some suggestions for future activities, both theoretical and more practical. The text is concluded with a list of references. [6]

[5] Another option is that each subgroup is characterized by a set of product descriptions. This would be the case if supervised learning would be used. The characterizing set of descriptions would form a training set. This should be build using a classification to link the descriptions (from real products) to. This is an interesting approach, but it is not considered in the present report.

[6] The author would like to thank Sander Scholtus for reviewing this document and for suggesting some improvements.

2. Descriptions and features

2.1 Descriptions

The descriptions of the GTINs that are the base input that we want to use are typically in a free format[7], so that they cannot be used 'as they are'. They have to be transformed into a form more suitable for statistical use, in particular to partition the items into groups or strata. More concretely we wish to transform the descriptions into a matrix, in which the rows correspond to the GTINs and the columns to the variables (implicitly) found to be in the data (attributes) through their values (characteristics). So the matrix elements are the characteristics in the descriptions to which attributes have been assigned. The attributes are variables and the characteristics their values. An attribute-characteristic pair will be sometimes be referred to as a feature in the present paper.

Typically no attributes are mentioned in the descriptions, only characteristics. A user has to identify the characteristics in descriptions that belong to the same attribute, which he should name. So a description like 'Blue cotton HappyPeople jacket', there is a colour attribute (with characteristic 'blue'), a fabric attribute (with characteristic 'cotton'), a brand attribute (with characteristic 'HappyPeople') and a clothing type attribute (with characteristic 'jacket') involved. The attributes have to be identified by a consumption analyst. This process of 'attribute identification' is impossible[8] to fully automate, as it requires quite some knowledge 'of the real world', in this case of clothing.[9] The characteristics are part of the description, but the assignment of the characteristics to attributes to be specified is the task of such a consumption analyst. It is very well possible that such a person decides not to use all characteristics in a description, only those that are correlated to the prices of products. For instance 'size' of clothing is independent of the price, at least in a certain range. Baby clothing, clothing for pregnant women or for big or tall people, tends to be more expensive.

Of course, being data, descriptions are not without errors, flaws or irregularities:

- Typos occur, more or less randomly (at first sight),
- Spelling variations occur,
- Words have been separated wrongfully,
- Words have been concatenated wrongfully,
- Words have been omitted,
- Synonyms or abbreviations have been used,[10]

[7] But not always. See Section 3.5 on consumer electronics.
[8] At least at the current state of affairs at CBS.
[9] But such a user would be tremendously helped by an appropriate specialized tool for feature extraction from descriptions. We expect that this tool operates interactively, where a consumption analyst has to make decisions, but the tools collects the relevant information so the analyst can make these decisions conveniently. The decisions themselves are not difficult to make, basically what characteristics to use and how to name the corresponding attributes.
[10] Their meaning may sometimes be obscure.

– Order of the characteristics in descriptions (in the same file) may vary.

All of these can be expected to turn up in descriptions (among others). In Section 3 some examples are presented. Besides, different retailers are likely to use different descriptions of the items they sell. These descriptions also may differ in the kind and number of characteristics they contain. This variation of descriptions over the various retailers whose data are used for price index computations, may imply that attribute extraction may be (slightly) different for different retailers.

Descriptions are usually very short pieces of prose, in telegram style. They typically consists of lists of characteristics, with rudimentary sentences, free text (to some degree), with a certain freedom in which order these characteristics are listed. Also there may be variation in the terms or abbreviations that are used to indicate characteristics: 'jacket', 'jack.', 'jckt' could be used to indicate 'jacket'. There is no indication which attribute belongs to a characteristic. So a colour such as 'white', 'blue', 'red', 'yellow', etc. may be mentioned (or ad hoc abbreviations, such as 'wht', 'yllw', etc.), but we have to infer that we are dealing instances of an attribute 'colour'.[11] Sometimes the meaning of certain abbreviations is ambiguous and has to be guessed (or inferred) from the context by a consumption analyst.

2.2 Features

Suppose that we have products in some product group (say office supplies) that are described by a limited set of attributes. These attributes could have been specified explicitly by the retailer selling the item, or they could have been derived from the descriptions used by a particular retailer.

The descriptions typically contain characteristics (like 'dark blue', 'cotton', 'stitched', etc.) but lack the corresponding attributes ('colour', 'material' / 'fabric', 'manner of making', etc.). They have to be identified by a human being and also which characteristics belong to which attributes, say an expert in the type of products sold in a particular retailer ('clothing', 'groceries', 'drugstores', etc.). The attributes act as variables that can be used to describe the products. As indicated before, we call an attribute-characteristic pair a feature, like ('colour' and 'blue').

Suppose that we have m items in a particular product group $I_1, \ldots, I_m$, each identified by a unique GTIN. And furthermore that n attributes $A_1, \ldots, A_n$ are used to characterize these items. For item I_j and attribute A_k the characteristic is denoted by χ_{jk}. We can arrange this information in the form of an $m \times n$ table as in Table 2.2.1.

[11] This is only true if we start collecting information about certain attributes and characteristics. But if such information has been collected in the past a computer program could be used to 'guess' to which attributes certain characteristics belong.

2.2.1 Items and their attributes (columns) and characteristics (cell values)

	A_1	...	A_n
I_1	χ_{11}	...	χ_{1n}
...	...	...	...
I_m	χ_{m1}	...	χ_{mn}

2.2.2 Indicators of characteristics of the attributes for the items.

		A_1		...		A_n	
	A_{11}	...	A_{1k_1}	...	A_{n1}	...	A_{nk_n}
I_1	$\Delta_{1,11}$	...	$\Delta_{1,1k_k}$	...	$\Delta_{1,n1}$	...	Δ_{1,nk_n}
...	...	...	...	...	...	...	...
I_m	$\Delta_{m,11}$	...	$\Delta_{m,1k_1}$	...	$\Delta_{m,n1}$	...	Δ_{m,nk_n}

In its simplest form, Table 2.2.1 is complete, in the sense that none of the χ's is missing. But in practice, some of the χ's may be missing. That a characteristic for an item is missing may be due to the fact that the information is simply lacking, or that the attribute is not applicable to a particular item or group of items.[12] Think of an attribute as 'Paper quality' in the product group 'Office supplies'. This attribute does not apply to 'perforators' or 'scotch tape', to mention but a few items in this product group. For this particular product group, the item indicators ('scotch tape', 'ring binder', 'perforator', 'stapler', etc.) can also be viewed as characteristics of an attribute like 'Office supply item'.

In Table 2.2.2 the same information is presented as in Table 2.2.1, but the various characteristics have been made to correspond to columns in the matrix. Characteristics corresponding to the same attribute have been put next to each other and the corresponding attributes are shown in a row above these. As in Table 2.2.1 the rows of Table 2.2.2 correspond to GTINs. The Δ's are indicators taking the values 0 (= 'does not apply'), 1 (= 'does apply') or 'missing'. Not that the Δ's in the various cells are characterized by three indices, in the following order: item indicator (GTIN), attribute indicator, characteristics indicator.

The advantage of representation in that in Table 2.2.2 over that in Table 2.2.1 is that it is convenient to use in computations. For instance, it is easy to aggregate characteristics: this is simply a matter of adding columns using binary addition.[13]

Remark

The information in Tables 2.2.1 and 2.2.2 is semantic information, that is, it is language independent, apart from the labels used to denote the items (GTIN-identifiers). But it could also be a description in some language, a picture or a set of pictures per item, or a short film showing the product) and the labels to describe the attributes or characteristics. ■

[12] In the latter case the meaning of NA is 'not applicable' rather than 'not available'.
[13] With the following rules: $0+0=0, 0+1=1+0=1, 1+1=1$.

Remark

Suppose that the labels of these attributes and characteristics are specified in a language that one does not understand (say Japanese, as in the author's case). Then by comparing the items with the same characteristics one could perhaps infer what attribute is in fact involved and which characteristic, simply by looking at what different items have in common.

This also implies that once Tables 2.2.1 and 2.2.2 have been compiled using labels in some language, only the labels need to be translated in order to make them suitable for use for all NSI's in the European Union, say. The relationship that the tables provide is the same, for all languages, or, put differently, is language independent. ■

2.3 Quantity indicators

One should be aware that some characteristics may in fact be quantity indicators: the number of socks in a set, the volume of glue in a bottle, the size of a fitted sheet, etc. In the case of bed clothing (see Section 3.1) in many descriptions sizes can be found of bed linen, such as: 'Molton waterdicht hoes 90x200', 'AA MOLTON.HS 140x200', 'Hslk 160x200', etc. (see Table 3.1.1). Such descriptions contain sizes: 90 cm x 200 cm, 140 cm x 200 cm and 160 cm x 200 cm, respectively. We have added the units of measurement: centimetres, that is lacking from the description but that is implied.

These size indications can be used to define unit value prices, that is prices per square meter, per (single) item, per litre, etc.. So for the bedding examples given we would have to divide the prices by 1.8, 2.8 and 3.2 respectively to obtain the corresponding prices per square meter. The description minus the size indications can then be used to classify the items. In case of the examples we gave we would then have: 'Molton waterdicht hoes', 'AA MOLTON.HS', 'Hslk'. This approach would lead to less strata, and probably to more stable ones, and, generally, to strata with a better 'filling' as well.

Furthermore, it should be remarked that there is no need to use all characteristics. T-shirts of different sizes, but otherwise the same, have the same price. So in this case the attribute 'size' does not have to be extracted from the descriptions.[14]

3. Examples of descriptions

To illustrate the variety of descriptions that are used in practice we consider the following groups of items from Dutch retailers: bed clothing, office supplies, pastries, men's T-shirts and consumer electronics. The information on the first four groups of

[14] Although it is only true for a certain range of sizes. Very big or very small sizes (for babies or toddlers) can have deviating prices.

products is from scanner data. Those on consumer electronics can be found on the web. It can in principle be linked to scanner data about such items. But this has yet to be done.

3.1 Bed clothing

In Table 3.1.1 a sample of descriptions used for bed clothing items is presented.

3.1.1 Some descriptions for bed clothing

AA ST.MOLTON HS 1 PER
AA ST.MOLTON 2 PERS
AA WATERD MOLTN MAT
AA jers.hslk 90x200 wt
Molton waterdicht hoes 90x200
AA MOLTON. HS 140X200
AA HSLKN 90X200 WIT
AA slopen 60X70 wit
AA Jers.hslk 90x200 ec
AA Molton sl 60x70 cm
Jers.hslk 180x200 wit
Molton stretch 90x220
Jers.hslk 180x200 ecru
HSLN 90X200 ECRU
Hslk 180X200 wit
HSLKN 140X200 WIT
HH Hslkn 90x200 bright white
Hslk 160X200 wit
B.hslk 2p+ 180x200 wit

Closer inspection of Table 3.3.1 reveals a number of peculiarities:

- In most descriptions it is explicitly indicated what the item involved is (like 'hoeslaken' = 'fitted sheet', 'sloop' (cf. 'sl')= 'pillow case', 'hoes' = 'cover'), but not always: 'AA ST.MOLTON 2 PERS', 'Molton stretch 90x220' are examples of descriptions where the item intended is not explicitly mentioned, but can be inferred (certainly by a consumer analyst).
- The various abbreviations used for the same thing can be treated as synonyms. For instance, for 'hoeslaken' the following abbreviations are used: 'hs, 'hslk', 'hoes', 'HS', 'HSLKN', 'HSLN', 'HSKLN' and possibly others as well.
- As remarked in Section 2.3 in some descriptions size information is used, namely the sizes of fitted sheets. This information can either be used as a characteristic of the items. Or it can be used to calculate a unit value price, say a price per square meter. In the latter case, only the remaining characteristics would be used to classify the items, and hence results in an aggregation.
- Characteristics on the type of item, size (exact, material ('molton', 'Jersey', and colour / design (e.g. 'wit' = 'white', 'ecru' = 'light fawn', 'MAT' = 'matt', 'bright white') can typically be found in the descriptions. But not always. Sometimes at least one of these indications is missing. In a few cases additional information is given (e.g. 'waterdicht' = 'watertight'). The size information is sometimes exact

('90x200' the units of measurement are not always given, but we may assume that 90 cm x 200 cm is meant) or sometimes less exact ('1 PER'= '1 person', '2PERS'= '2 person').
- The order of the characteristics is not fixed but tends to follow the pattern: fabric, item, size, colour. When classifying items automatically, this information could also be used.
- Spaces are sometimes absent. For instance ST.MOLTON (which should be 'ST. MOLTON' = 'stretch molton'), 'Jers.hslk' (which should be 'Jers. hslk' = 'Jersey hoeslaken'). An automatic classification system should be able to cope with such errors.

It is interesting to investigate if a stratification based on the item (without the adjectives and specializations) is good enough for index calculations. And also if the use of unit values is attractive for this purpose.

3.2 Office supplies

The product group 'office supplies' is the most diverse of the five product groups we consider here, as a glance at Table 3.2.1 already shows, although it only contains a small fraction of the items in this lot.

3.2.1 Some descriptions for offices supplies

AA VP 2 Brievenbak transp
AA Burolijm 100 ml.
AA LP Tijdschriftcass.transparant
Tijdschriftcassette transparant lux
AA Documap wit
AA Documap zwart
Documap
Documap
Documap
Documap
Documap aqua
Documap dessin roze
Documap groen bloem
Documap roze stip
Documap dessin
Documap geblokt
Documap groen
Documentenmap roze
AA Elastiekjes

'Office supplies' contains a variety of items that are used in, or are associated with, an office, but actually used in households, such as:

- Glue
- Containers (for letters, magazines or documents)
- Rubber bands
- …

Because there is quite a variety of items available, the characteristics used are very diverse. But the feeling is that the characteristic key word in each description (like 'documap' = 'container for letters', 'elastiekjes' = 'rubber bands', etc.) should be used as the basis for a partitioning of 'office supplies'. It is less detailed than GTIN but still a refinement of 'office supplies' into more homogeneous strata.

So if we look at Table 3.2.1, we talk about 'brievenbak' = 'containers for letters', 'burolijm' ='office glue', 'tijdschriftencassette' = 'magazine holder', 'documap' ='document holder', 'elastiekjes' = 'rubber bands'. In some cases the items in such a group differ in size (the number of pages in a package of printer paper, the contents of a glue bottle, etc.). This information should not be ignored. But attributes like 'colour' or 'design' probably can, as they are less likely to be important determinants for the prices of the products.

On the basis of these key characteristics we should try to find suitable hyperonyms, like 'filing systems' (e.g. 'magazine holders', 'document holders'), 'binding materials' (e.g. glue, 'scotch tape',' sticky tape',' sellotape', 'staples', 'paper clips', 'rubber bands'), 'office equipment' (e.g. 'perforators', 'staplers', 'calculators').

3.3 Pastries

Table 3.3.1 contains a small sample of descriptions for pastries that are being used.

3.3.1 Some descriptions for pastries

AA Appeltaart
AA Appelkruimelvlaai
AA Tompouce
AA Slagroomtaart
AA Vanille-slagroomvlaai
Taart op Maat gekleurd
AA Slagroomtaart 25 cm
Mokkaslagroomtaart
Aardbeien-roomvlaai
Maxi Taart op Maat kleur
Aardbeien slagroomvlaai
Mini tompouce op maat
Taart op Maat amandel
Slagroomschnitt
Drievruchtenvlaai
Chocoladetaart
Gebak
Slagroomrijstevlaai
Moorkop

The product group 'pastries' involves the most complicated descriptions of the four product groups considered in the first four subsections of Section 3. A lot of compound words are used, with various characteristics involved. However, the descriptions tend to be short in terms of the numbers of words used. Quite often a description consists of only one (compound) word.

The product group 'pastries' is studied in more depth in Willenborg (2017c).Because characteristics and attributes for pastries are described in Dutch, that document is also in Dutch. Please refer to Willenborg (2017c) if you are interested in the kind of information that can (or cannot) be extracted from the descriptions used for pastries (and are able to read Dutch).

3.4 T-shirts

Men's T-shirts form a rather homogeneous group of items. Table 3.4.1 contains a small sample of available descriptions. This population can be subdivided on the basis of characteristics such as neck shape ('O-shape', 'V-shape'), fabric ('organic'/'basic'), sleeve length ('KM' = 'korte mouw' = 'short sleeve', 'LM' = 'lange mouw' = 'long sleeve'), pack size /set size (2,3,… items in a pack / set), colour ('wi'= 'wit' = white, 'zw' = 'zwart' = 'black', 'mgrm' = <an unknown colour>), form ('stretch' / 'normal'), size ('L' = 'large', 'M'= 'Middle', 'XL' = 'eXtra-Large', etc.). For sizes that are not extreme, the prize is usually the same. But extra-large sizes might cost more. But if these extreme sizes cannot be bought from the retailers considered it is unnecessary to select 'size' as a possible stratification variable. Otherwise 'size' should be included among the potential stratification variables.

It is clear these descriptions are rather limited in content. But it should be noted that we are dealing with a specialized group of clothing items already. It is a priori not clear to which extent this subgroup of products should be further stratified. But as Table 3.4.1 shows there are some additional characteristics available in the descriptions that could be used to refine this product group.

3.4.1 Some descriptions for T-shirts

Organic stretch t-shirt KM O-neck, wi, L
Organic stretch t-shirt KM O-neck, wi, M
Organic stretch t-shirt KM O-ne, mgrm, M
Organic stretch t-shirt KM O-nec, wi, XL
VA Organic stretch t-shirt KM O, wit M
Organic stretch t-shirt KM V-neck, wi, L
Organic stretch t-shirt KM V-neck, zw, M
Organic stretch t-shirt KM O-ne, mgrm, L
Organic stretch t-shirt KM V-neck, wi, M
T-shirt O-neck 3-pack, wi, XL
T-shirt O-neck 3-pack, wi, L
Organic stretch t-shirt KM V-ne, wi, XXL
Organic stretch t-shirt KM V-neck, zw, L
Organic stretch t-shirt KM V-nec, wi, XL
Basic t-shirt KM O-neck 2-pack, mgrm, XL
AA T-shirt O 2-pack, wi, L
Basic t-shirt KM O-neck 2-pack, wi, L
Basic t-shirt KM O-neck 2-pack, mgrm, L
O-neck T-shirt LM stretch, wi, M

Looking at the descriptions in Table 3.4.1, it is clear that alternative spelling is used to denote 'neck'(like 'ne', 'nec') and sometime this indication is discarded altogether

and simply ‘O’ is mentioned. Sometimes a space is left out by accident and ‘KMV-neck’ or ‘KMO-neck’ has been written instead of ‘KM V-neck’ and ‘KM O-neck’.[15] Looking at the order of the characteristics mentioned, it is clear that this is not fixed. Typical is an order like in the first record in Table 3.5.1: ‘Organic stretch t-shirt KM O-neck, wi L’ (i.e. fabric, form, ‘t-shirt’, neck shape, colour, size). But also deviations like ‘T-shirt O-neck 3-pack, wi, XL’ (i.e. ‘t-shirt’, neck shape, pack size, colour, size) occur or ‘AA T-shirt O 2-pack, wi, L’ (i.e. ‘AA’, ‘t-shirt’, neck shape, pack size, colour, size) occur. Probably the number of such variations is rather limited, but this has not been investigated.

3.5 Consumer electronics

To contrast with the product descriptions of the previous subsections, we now consider those used for items within the consumer electronics group. This information can be found on the web. The information of the four products considered above is found in scanner data.

At CBS, consumer electronics is defined in terms of the ECOICOPS that belong to the set of products. Roughly, it contains devices such as white goods, cleaning devices, air conditioners, phones, mobile phones, televisions, cameras, PCs. In Table 3.5.1 it is specified (in Dutch) which ECOICOPs are considered by CBS to be part of consumer electronics.

3.5.1 ECOICOPs that define 'consumer electronics' at CBS

ECOICOP	Omschrijving
53110	Koel- en vrieskasten
53120	(Af)wasmachines en wasdrogers
53130	Fornuizen, ovens, magnetrons e.d.
53140	Verwarming, airconditioners
53150	Schoonmaakapparaten
53210	Keukenmachines
53220	Koffiezetapparaten, waterkokers en dergelijke
53230	Strijkijzers
82010	Vaste telefoontoestellen en toebehoren
82020	Mobiele telefoons
91110	Audio-opname- en weergaveapparatuur
91120	Televisietoestellen en videoapparatuur
91130	Draagbare beeld- en geluidsapparatuur
91190	Overige apparatuur door de opname en weergave van audio en video
91210	Camera’s
91220	Accessoires voor foto- en filmapparatuur
91310	Personal computers
91320	Accessoires voor gegevensverwerkende apparatuur
121210	Elektrische toestellen voor lichaamsverzorging

As Table 3.5.1 shows, consumer electronics is a rather diverse lot. This implies that the product descriptions one can expect are also different, as different features are needed for different classes of consumer electronics items. To illustrate this, we have included information on two different items, namely a washing machine and a tablet.

[15] It is not clear to the author what the abbreviations ‘VA’ and 'AA' mean. But a consumer analyst is probably familiar with them.

3.5.2 Features of a washing machine

Attribute	Characteristic
Prestatie	-
Wascapaciteit	6 kg
Centrifugesnelheid (max)	1000 RPM
Geluidsniveau (wassen)	59 dB
Geluidsniveau (centrifugeren)	76 dB
Uitgestelde start timer	Ja
Startvertraging	9 uur
AquaStop-functie	Nee
Design	-
Type lader	Voorlader
Kastontwerp	Vrijstaand
Trommelinhoud	38 l
Ingebouwd display	Nee
Kleur van het product	Wit
Ergonomie	-
Resterende tijd indicatie	Nee
Kinderslot	Ja
Energie	-
Energie-efficiëntieklasse	A++
Jaarlijks energieverbruik wassen	173 kWu
Jaarlijks waterverbruik wassen	9240 l
Gewicht en omvang	-
Hoogte	85 cm
Breedte	59.5 cm
Diepte	47 cm

Table 3.5.3 contains information on a tablet. This information was scraped by an internet bot from a web site with extensive information on consumer electronics items. The information shown in Table 3.5.3 is a 'prettified' version to increase readability. [16] But the features are in the data as presented in the table. No extraction was needed to obtain this information. In fact, more information is available, such as GTIN number, brand and type. These are important matching variables that can be used to enrich scanner data records with features collected from the internet.

[16] This was easily achieved using some standard functions in Excel.

3.5.3 Features of a tablet

Attribute	Characterisc	Attribute	Characteristic
Design		Wi-Fi standaard(en)	802.11a/b/g/n/ac
Soort	Tablet	Ethernet LAN	Nee
Formaat	Slate	AirPlay	Nee
Kindertablet	Nee	Bluetooth	Ja
Kleur van het product	Zilver	Bluetooth-versie	4.0
Software		4G	Nee
Besturingssysteem	Windows	3G	Nee
Besturingssysteem versie	10	Poorten & interfaces	
Processor		Hoofdtelefoon uit	Ja
Processorfamilie	Intel Core i5	Hoofdtelefoon/microfoon combo poort	Ja
Processormodel	i5-6300U	Dockingconnector soort	Surface Connect
Processor frequentie	2.4 GHz	Aansluiting voor netstroomadapter	Ja
Aantal processorkernen	2	Aantal Mini-HDMI poorten	0
Geheugen		Aantal Micro-HDMI poorten	0
Intern geheugen	4 GB	Aantal USB 3.0 poorten	1
Opslagmedia		Aantal USB 2.0 poorten	0
Interne opslagcapaciteit	128 GB	Aantal Micro-USB 3.0 poorten	0
Opslagmedia	SSD	Aantal Micro-USB 2.0 poorten	0
Geïntegreerde geheugenkaartlezer	Ja	Prestatie	
Compatibele geheugenkaarten	Micro-SD	GLONASS	Ja
Totale opslagcapaciteit	128 GB	Gyroscoop	Ja
Beeldscherm		Versnellingsmeter	Ja
Beeldschermdiagonaal	12.3 inch	Orientatie sensor	Ja
Resolutie	2736 x 1824 Pixels	Omgevingslichtsensor	Ja
Beeldverhouding	3:2	Barometer	Ja
LED backlight	Ja	Beveiliging	
Touchscreen	Ja	Trusted Platform Module (TPM)	Ja
Touch screen type	Capacitief	Accu/Batterij	
Touch technologie	Multi-touch	Accu/batterij gebruiksduur	10 uur
Pixeldichtheid	267 ppi	Continue video-afspeeltijd	9 uur
Grafisch		Gewicht en omvang	
Grafische adapter-familie	Intel HD	Hoogte	20.14 cm
Grafische adapter	Intel HD graphics 520	Diepte	8.45 mm
Camera		Breedte	29.21 cm
Camera voorzijde	Ja	Gewicht	768 g
Camera achterzijde	Ja	Inhoud van de verpakking	
Resolutie camera voorzijde (numeriek)	5 MP	Inclusief toetsenbord	Nee
Resolutie camera achterzijde (numeriek)	8 MP	AC-adapter meegeleverd	Ja
Video recording	Ja	Geheugenkaart meegeleverd	Nee
Maximale videoresolutie	1920 x 1080 Pixels	Styluspen	Ja
Automatisch scherpstellen	Ja	Koptelefoon	Nee
Audio		Opbergetui	Nee
Audiosysteem	DolbyÂ®-audio	Snelstartgids	Ja
Ingebouwde luidsprekers	Ja	Gebruikershandleiding	Ja
Aantal ingebouwde luidsprekers	2	Garantiekaart	Ja
Ingebouwde microfoon	Ja	Overige specificaties	
Netwerk		Basisstationaansluiting	Ja
Wi-Fi	Ja		

3.6 Comments

The examples in the previous subsections show that the product descriptions can vary significantly. In four cases – all concerning scanner data - the characteristics and attributes have to be extracted from the descriptions. And these descriptions tend to be short and contain only a few features. In one case - consumer electronics - the

features are explicitly given.[17] And there are much more attributes (and characteristics available). In this case the data are from the web, and they can potentially be used to enrich scanner data, provided both types of data can be linked. There seem to be sufficient keys available to make the matching possible.

If the route is chosen to work with attributes and characteristics to define subgroups, an interactive[18] software tool would be very handy in case we are dealing with product descriptions where attributes and characteristics are implicitly available.[19] The examples given above show that there is variety among the descriptions used.

1. The product groups bed clothing, office supplies, pastries and T-shirts differ in heterogeneity. The most heterogeneous of these is 'office supplies' and the most homogeneous (in terms of the items concerned) is 'men's T-shirts'.
2. The descriptions of 'pastries' often consist of compound words, with information on several attributes (main ingredients, size, plural / singular, intended occasion / festivity for the pastry, etc.).
3. The descriptions, except for consumer electronics, are to be considered free text. This implies that there are errors (e.g. spaces that have been deleted between consecutive phrases), there are spelling variations (using different letters, in capital or in normal fonts). Sometimes information is missing (e.g. an indication of the item, dimensions used (length in cm's, for instance). But probably a consumer analyst looking at these descriptions knows immediately what is intended and can readily supply the missing information. This is a strong argument in favour of an *interactive* extraction tool.
4. The information on consumer electronics is (far) more detailed than that of the four remaining product qualities. The information seems also to be of higher quality.
5. The characteristics are to be interpreted as categories of attributes that have to be named by an expert. For instance, the descriptions contain phrases like 'wit', 'bright white', etc. (among many other phrases) and the expert has to understand that they are all categories of a variable 'Colour'.
6. If some characteristics occur only in a limited number of descriptions, this is no reason to dismiss them as uninteresting. It is possible that they are categories of a variable with many categories. Colour could be an example of this. Or it is possible that the product group splits into several different subgroups (as in case of office supplies) and some characteristics only apply to one or a few such subgroups. Another possibility is that characteristics have alternative spellings and should be treated as synonyms. So the relevance of characteristics should not be dismissed as irrelevant purely on the basis of simple frequency counts.

[17] Apart from one field which contains a string, with brand and type of the item.

[18] Such a tool is feasible. But maybe a more ambitious one using machine learning techniques would also be feasible. But we stick to the more modest tool.

[19] But it may well be that it turns out to be more attractive to use another approach not based on attributes and characteristics but on a machine learning technique applied to entire descriptions. In this case one would have to use training sets of descriptions that characterize certain product subgroups. However, for the moment we stick to the more traditional approach, using attributes and characteristics.

7. In principle, the descriptions are sources of characteristics of the items. But in some cases also quantity information is available (sizes of fitted sheets, numbers of t-shirts in a set, the amount of glue in a bottle, etc.). This allows the calculation of unit value prices for these items (e.g. the price of a fitted one person sheet per square meter, the price of one T-shirt from a set of three, the price of glue per 100ml, etc.). Items can then be stratified on the basis of the remaining attributes.

From the fact that the focus is so strongly on all the attributes and characteristics available in the descriptions, it should not be concluded that they should necessarily all be used to stratify a product population. The problem of selecting the attributes for stratification, which would be the next step after attribute and characteristics selection is not considered here. In an extreme case, it may even be decided not to stratify a product group at all, for instance because its contribution to the CPI is only small, and stratifying it is next to meaningless.

4. GTINs, subgroups and groups

4.1 Introduction

GTINs are the base ingredients for price index calculations. To use them directly, however, has some complications, as is explained below. The lowest level at which price index results are published are (E)COICOPs.[20] But these groups are often too heterogeneous to use them as strata directly. In this case it may be preferable to use an intermediate level of aggregation, a subgroup level. There may in fact be many subgroups levels between the GTIN level and the group level, and the challenge is to find a good one. And also to know, which is actually a good subgroup level and which is not. Without this knowledge it is impossible to judge a stratification into sub-groups.

If populations of products would be stable, the GTINs would be the natural level to base the price index computations on. One would simply compute price indices per item and aggregate them to a price index at the group level, and beyond.

But static product populations do not exist in practice; they are theoretical abstractions. In practice product populations are dynamic. But for dynamic populations it is not straightforward how to compute these indices starting at this level. If one only considers items separately, one may miss price increases, due to the fact that some

[20] ECOICOP = European Classification of Individual Consumption to Purpose. For more information see: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrLanguageCode=EN&StrNom=COICOP_5&StrLayoutCode=HIERARCHIC .

item are relaunched, by packaging them differently, and changing some superficial characteristics such as lettering and images on the package.

So if the GTIN level is chosen one has to cope with relaunches and hence one faces a matching problem. But matching relaunches to items that were previously on the market is not trivial. No such information is available from a producer or a retailer. So the statistical office has to find a method to produce these matches. This process is not error-free. And it also uses available attributes and characteristics in the data. These are the same as one would use in case of stratifying the product population.

But it is questionable whether this is the right approach. It certainly is not the simplest one. A simpler approach is to consider stable subgroups of GTINs. Comparing 'like with like' at the subgroup level is easy. We will consider both approaches in more depth in the remainder of the present section.

4.2 GTIN matching

GTINs are identifiers for product items. They are used for logistic purposes. It is convenient use them for production, ordering and transportation. It may be tempting to view GTINs as keys in the sense of data bases. But in fact, they are not. They can be defined by companies and retailers for logistics and sales purposes. This tends to produce rather heterogeneous definitions.

For the same product a retailer can use two GTINs, for instance one for the items in the sale and the other one for the same items in the regular collection. For similar products, however, a company may decide to use a new GTIN. The product in question may have had a facelift (new packaging, new print, bigger bottle or package, etc.) but nothing essential has changed. That is, except for the price, that is likely to be increased.[21] Such a product is a relaunch.

So items and their relaunches should be viewed as essentially the same thing, and if item and relaunch are not matched, a vital price jump may be missed. But this raises the question: how does one recognize relaunches? Typically such information is not provided by retailers or producers. So a statistical office should solve this problem. And it should be solved in such a way that it can be done largely by computer. There are simply too many scanner data to assume that consumer analysts can do this, certainly not 'by hand' but also not with the help of interactive matching software.

At first sight it may seem obvious , this matching process. But on closer inspection it is not. It raises a number of questions. The first question is: Which items should be matched? Obviously, the relaunches. But how does one recognize relaunches? To this one could answer: the GTINs that do not have predecessors. But how far should one

[21] Taking into account that the relaunched product may be packaged in a package with a different volume, weight or number of items inside. If a relaunch of a package of cigarettes has less cigarettes per package while the selling price is the same as that of its predecessor, the price per cigarette is thus increased by the relaunch.

look back for this? One month? Two months? Three months? Half a year? The problem is that an item may have been unavailable for several months. And, suppose these questions have been answered, the next question is: How should such (potential) relaunches be matched? Obviously GTIN code cannot be used. So perhaps one can match items using common characteristics? It should be borne in mind that these problems have to be solved by a computer program, at least for the bulk of the data. This is complicated enough, and it is likely to lead to quite arbitrary decisions as to what items are relaunches and which are not.

But which item to choose if there are several items with the same characteristics (but different prices)? Can we randomly choose one such item? How far back in time do we have to look to find a predecessor of a relaunch? It is clear that this is a fairly complicated process and one that is likely to produce false matches. Due to the volume of data, this process has to be carried out automatically, by some dedicated software tool.

So it is clear, this is a quite complicated matching process, when the matching should be based on characteristics. And also it tends to be error-prone, as one does not know when an item is a relaunch and how to match relaunches with predecessors. It is not sure that the right matches are found in this way, as well. So why stick to it if there is so much uncertainty associated with it, and it is so complicated to implement?

In the next subsection we consider a simpler approach, which we also consider superior to the GTIN matching method described in the present subsection in several ways.

4.3 Subgroup matching

Would it not be easier to classify all items on the basis of (some of) their characteristics? Then if an item and its relaunch have the same characteristics, they are classified (automatically) to the same group. This is not only far more attractive, as one does not have to bother about relaunches and handling them in a specific way. It also seems the superior approach if one does not consider the individual items as leading but the groups into which they are classified. We should focus on them and their properties, and consider the items at GTIN level.

The situation is comparable to, say municipalities. If we want to produce statistics about them, we do this not by linking individual citizens of a municipality at different points in time. We do this by taking the citizens that are present at a particular moment. Some may be present at both moments, others only at one of them. It is inherent for municipalities that they are entities that are constantly changing.

If one thinks about it, such changes do not hold only for municipalities, but for essentially all entities build from more fundamental parts, that is, composite entities. For human bodies, factories, schools they have in common that they all change

constantly at the micro level, but they remain unchanged at a more abstract level.[22] The same is true for the items that we see constantly being put on the market and taken out of it after some time. So, to be more concrete, specific types of jeans come and go, but 'jeans' as a group of products exists for a much longer time. The message is that we should look for the wood rather than for the trees.

The purpose of the subgroups is to act as a layer between the GTIN and the group level.[23] They should provide the continuity the GTINs cannot provide in a dynamic product population. On the other hand, the subgroups should be sufficiently homogeneous and stable. These are properties that tend to be conflicting. So the challenge in practice is to find the right balance.

Once a product population has been stratified into strata/subgroups, the idea is to apply the principle to compare 'like with like' to subgroups rather than to GTINs. The composition of the subgroups in terms of GTIN may differ from month to month. This is not an issue anymore. The GTINs that compose a subgroup in a given month are only used to compute a total turnover and a total quantity for that subgroup for that particular month, simply by adding the respective values for the GTINs involved.[24]

$$v_{ij} = \sum_{k \in A_{ij}} v_{ij,k}, \tag{4.1}$$

$$q_{ij} = \sum_{k \in A_{ij}} q_{ij,k}, \tag{4.2}$$

Here A_{ij} denote the items in subgroup i in month j. From (4.1) and (4.2) a price for the subgroup i in month j can be computed:

$$p_{ij}^{av} = \frac{v_{ij}}{q_{ij}}. \tag{4.3}$$

The subgroup prices (4.3) are then in turn used for price index computation: per subgroup, these prices are compared across different months:

$$\Pi_i^{j_1 j_2} = \frac{p_{ij_2}^{av}}{p_{ij_1}^{av}} = \frac{v_{ij_2}}{v_{ij_1}} \frac{q_{ij_1}}{q_{ij_2}}. \tag{4.4}$$

Of course, this only makes sense if the subgroups do not change fundamentally. But this cannot be assumed automatically and should be monitored. This is considered in the next subsection.

[22] All of this reminds one of an observation by the pre-Socratic philosopher Heraclitus that no man ever steps in the same river twice. Although the contents of a river is constantly changing, what remains the same is the entity that is 'the river', which transcends its constituent parts at any given time. Despite its different contents, we have rivers like the Rhine and the Meuse. The fact that these rivers have been given names is indicative of a kind of persistence that is lacking at the micro level. At that level existence is a much more fleeting affair.

[23] In case the GTINs in some group are fairly stable, these subgroups could be made to agree with the GTINs, or most of them.

[24] This is possible because these variables are additive, in contrast to the price, being a ratio of turnover and quantity.

4.4 Monitoring subgroup composition

It was stated before that changes of the composition of subgroups should not worry us per definition. As long as the nature of the subgroup remains the same there is no problem. But certain changes may lead to essentially new entities. It may be necessary to reconsider the subdivision of a product group. Perhaps a subgroup should be split into two (or more) new ones. Or certain subgroups are not really different anymore and should be united.

The goal is to carry out price index computations automatically as much as possible, due to the sheer volume of the data to be processed each month. This carries the danger that new data do not fit very well with the chosen stratification of COICOPs into subgroups. So it should be monitored that subgroups do not change drastically. But how should this be done?

A first idea is to monitor the price per subgroup, which is an average price of all the items in this subgroup. If this price (or its variance) changes dramatically in a particular subgroup in a short period of time, this could be an indication of a structural change in this subgroup. It should cause an inspection of this subgroup. This approach is an example of macro-editing, where aggregates are being monitored, and only if they show a deviant behaviour there is a reason to drill down to the most basic level to see what is going on.

But macro-editing only reacts to prices. It could also be that a subgroup is changing qualitatively, in the sense that new products creep in that do not necessarily have a very different price. But they simply do not fit the mould and should be placed in a separate subgroup. Such a change can only be noticed looking at the items inside the subgroups, in particular their characteristics that are not used to classify them into a subgroup. If this is to be done automatically, these extra characteristics should also have been extracted from the descriptions. Every month, the frequencies of their respective occurrences should be monitored. Following these over time seems adequate to monitor the stability of the composition in terms of the multiplicities of the various extra characteristics.

Of course, both control measures do not help to detect items that have new features in their descriptions but that are sold for prices within the range of the original items in the 'affected' subgroup. Such changes can only be detected by regular inspection of the composition of all subgroups, or a sizeable sample thereof. But as long as the prices of such items are within the range of the regular products in the subgroups, there is not much reason for concern for the index computations.

5. Desirable stratifications

5.1 Approach

We may assume that the groups of items that we are considering (COICOPS) are persistent, in the sense that they will have sufficient 'filling' each month, so that price indices at that level are possible to compute for each pair of months. But as they may be rather heterogeneous, the idea is to improve them by splitting (some of) them up into subgroups. These subgroups can be expected to exist above the GTIN level, as GTINs are often too fleeting, as we have seen before. The question is now: which properties should these subgroups have in order to make them suitable for index computation?

A first requirement is that they should be stable in time: it should be possible to allocate GTINs in our group of items (say 'fruit') to any of these subgroups. This means that the stratification should in fact be a partition of the population. There should be no need to create new subgroups all the time (as would be the case for GTINs).[25] Also it should not be the case that subgroups regularly 'peter out', in the sense that they do not contain any items, which limits the possibility of making comparisons over time. This requirement is weaker than persistency, as it should be possible that in some months there are no sales to report in a subgroup. But this should not happen too frequently. We refer to this property of subgroups as 'stability'.

At the group level, it was assumed possible to compute a price index. In order to be able to produce price indices at the subgroup level for every pair s, t of months there should be at least one subgroup σ for which a price index $P_\sigma^{s,t}$ can be computed. When the subgroups are stable this condition is not guaranteed to be met but is likely to exist. We settle for this property as it is easier to check than the property that the property that for each pair of months s, t there should be at least one subgroup σ for which $P_\sigma^{s,t}$ can be computed.

But stability of subgroups is not enough. We want to improve on the supposed heterogeneity of the product group, so we should consider homogeneity of subgroups as well (see Subsection 4.5 for a discussion of this subject). The general idea is to make all the subgroups within a group as homogeneous as possible, not losing sight of the stability requirement. If we succeed comparison of 'like with like' is achieved at the subgroup level. Trying to produce stable subgroups with maximum homogeneity is what we should look for.[26]

[25] This can always achieved by adding a category 'other'. But this category should contain only exceptional cases. And if this is hard, it should be closely watched as it develops. It should not become too big. If so, perhaps another stratification is called for.

[26] In practice optimality is an ideal to pursue, but we should not take it too literal. In practice we should look for solutions that work and are 'good enough' rather than 'optimal'.

In Figure 5.1.1 an example is presented to illustrate persistence and stability. A group G of products consists of 5 stable subgroups that together form a partition of G. So every GTIN belonging to group G should fall into exactly one of the subgroups G.1,…, G.5. Persistence of G is suggested by the presence of a black square for each of the months in the 9 month period[27].

5.1.1 Presence (black) or absence (white) of items at group (G) and subgroup (G.i) level

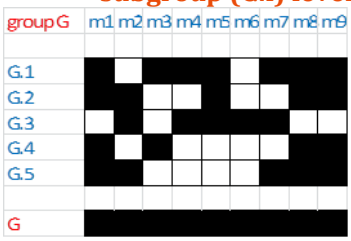
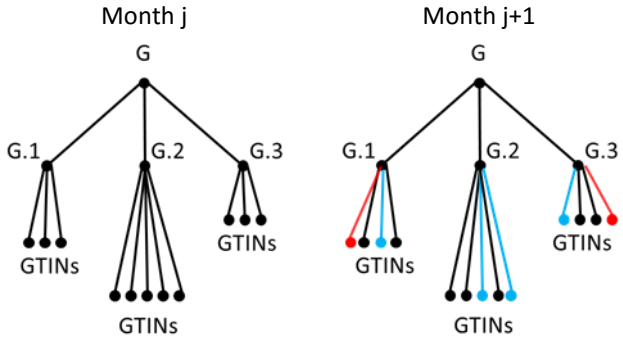


Figure 5.1.2 is to convey the idea that the groups and subgroups are stable, that the group is persistent, but that there are changes at the GTIN level. The red points and edges symbolize new GTINs and the subgroups they belong to, respectively. The blue points and edges symbolize GTINs that have been deleted and the subgroups where this was the case, respectively. Of course, homogeneity of the subgroups cannot be shown by this picture, but is tacitly assumed that they are homogenous.

5.1.2 Mutations within subgroups in a group



If after some time it appears that a product population is developing in a different direction than anticipated, it should be part of a maintenance procedure to revise the stratification of a group in stable homogeneous substrata, and even to redefine the partitioning into groups of products. But if done well such revisions should not occur too frequently.

It should be noted that the approach suggested here to obtain a desirable stratification is independent of the choice of a specific index method. The properties that we

[27] This is only for the sake of presenting an example to illustrate some ideas. In reality the period should be much longer to judge persistence of a group and stability of its subgroups.

have mentioned above (persistence, stability and homogeneity) are purely defined for the product populations that we are interested in. This is attractive, as it is possible to adopt or change an index method without bothering to adapt the stratification. Or the other way round: to change the stratification while keeping the same index method.

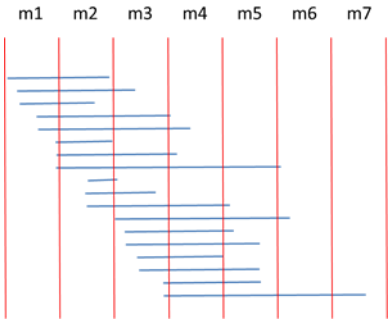
5.2 Persistence of groups

Persistence[28] is a property of the group of items that we are considering for stratification. Intuitively, it means that every month sufficient data should be available so that price indices can be produced every month. This is the case with scanner data if every month items (GTINs) of this group are sold or can expected to be sold (for the future). If groups coincide with COICOPS (or several COICOPs combined), this is the case. In fact, in the present paper we assume that the groups we consider consist of one or more COICOPs,

In Figure 5.2.1 a situation is depicted in which GTINs in a certain product group are sorted on the date of entry to the market. Each line segment symbolizes the lifetime of a GTIN, that is , the period that the GTIN is on the market. It may be unavailable temporarily, or have not been sold temporarily, but has not been withdrawn from the market.[29] Figure 5.2.1 shows persistence (in the 7 months depicted), because in every month at least one GTIN was sold (or present in the supposed scanner data). But persistence is also about expected continuity in the future, say several years ahead. This can be judged by a consumer analyst specialized in the products involved.

Notice that new GTINs appear and existing ones disappear. The set of GTINs in the group changes constantly, so comparisons of prices for the same items can only be made during their lifetime. But in that case price jumps due to the introduction of relaunches would be missed, and a wrong picture of the price development within this group would be created.

5.2.1 GTINs sorted on date of introduction to the market



[28] Also to be referred to as 'continuity', or 'flow'.

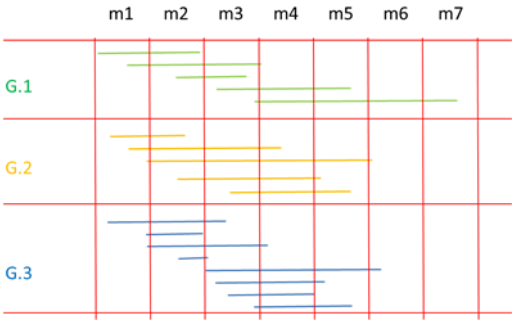
[29] We are aware that in practice there may be a problem to decide whether a GTIN is temporarily not available or has been withdrawn from the market. But in retrospect, this problem may be easier to settle.

As we have argued before, describing the price development at the GTIN level of a dynamic population is doomed to fail. Trying to repair matters by identifying re-launches and matching them to predecessors is both complicated (especially to do this on a computer) and error-prone (one has to guess which predecessor(s) match to each relaunch), and not guaranteed to be successful in a (very) dynamic population. It is then easier and preferable to consider subgroups of items and follow their price development, as was also argued before. The problem then is only to classify each GTINs in the correct subgroup. One does not have to worry about matching GTINs otherwise.

5.3 Stability of subgroups

Continuing with the example of the previous subsection, we can say that what we want is a classification of the GTINs into subgroups that are stable in time. Any new GTIN appearing on the market should be assignable to one of these subgroups. This is shown in Figure 5.3.1. The same GTINs as in Figure 5.2.1 are shown, but this time each of them is assigned (each month) to one of three subgroups G.1, G.2 and G.3. These subgroups form a partitioning of the entire group. The GTINs are sorted on the date of entry to the market within each subgroup.

5.3.1 The same GTINs as in Figure 5.2.1, grouped into three subgroups, sorted on date of introduction in each subgroup.



The subgroups are - or the stratification they collectively represent is - supposed to be stable. This means that for an extended period items (GTINs) can be allocated to (exactly) one of these subgroups. Of course, one can always define a partition of a group of items by introducing a remainder category ‘Other’. But then this category should remain one that contains only a few, odd items, over an extended period of time.

It is not required, though, that each subgroup is persistent. The persistence of the group of which the subgroups are part implies that at any month at least one sub-group is nonempty. This implies that for the group prices can be computed, as well as price indices using average subgroup prices.

In practice there may be GTINs with a quite long lifespan. Such GTINs may be stable enough to form a subgroup by themselves. But to do this is a bit risky concerning persistence. If it disappears from the market the subgroup seizes to exist.

5.4 Homogeneity of subgroups

In order to reach stable subgroups there is a tendency to define them broadly. But this would jeopardize their homogeneity. So we have to look at stability of the subgroups but keeping homogeneity in mind. Homogeneity of subgroups acts as a counter-balance to subgroup stability.

Homogeneity can refer to two things:

a. Homogeneity of the items in terms of (the distribution of) their characteristics.
b. Homogeneity in terms of the variance of the prices, or rather an estimate of this parameter (EV, or estimated variance), as the population mean is not known. So a stratification with a smaller EV would be preferable than one with a larger EV, given persistence in both cases, as it would mean a more homogeneous stratification.

As to the first point, it can be remarked that the items in a group have common characteristics, that is, as far as they were used to define the group. The items therefore can only differ on the basis of characteristics that we did not take into account. But are these additional characteristics known? If so, why have they not been used for stratification? Besides, if they were known, how should they be used? They could be used to define a further stratification and produce more homogeneous strata, but this time in the second sense of the word. And in (the most likely) case that they are not known we have to refer to this second interpretation of homogeneity. So ultimately this second interpretation is the one we would use in practice. This is also a quantified property.

Let μ_i^t and σ_i^t be the average and the standard deviation of subgroup i in month t. With these quantities we can estimate the variation coefficient (vc) per subgroup which is a measure for the precision:

$$vc_t = \frac{1}{m_t}\sum_{i\in N_t}\frac{\sigma_i^t}{\mu_i^t}, \tag{5.1}$$

where the sum is taken over the nonempty subgroups N_t in month t, and $m_t = |N_t|$ the number of non-empty subgroups in month t. For the entire period we have

$$vc_T = \frac{\sum_{t\in T} vc_t}{|T|}. \tag{5.2}$$

We have tacitly assumed that the group is persistent.

5.5 Detail

We quantify the detail of a stratification by the number of strata used. It seems preferable to use this measure rather than the number of variables used to stratify the group, as this is not well defined if the number of variables in the domain can vary.

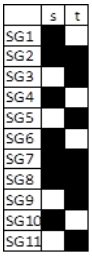
5.6 Overlap

This property directly relates to the possibility of price index calculations at the sub-group level. This means we have to look at pairs of months in a given time window. For each such pair we need to consider how many subgroups there are for which data are available in both months. To illustrate what is intended, consider the next example.

Example

Figure 5.6.1 indicates for a group with 11 subgroups, considered at two different months, s and t, when data (prices) are available (black square) or not (white square).

5.6.1 Fillings of a group with 11 subgroups at month s and t.



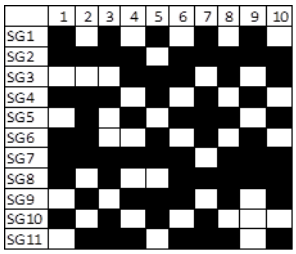
If we want to know for how many subgroups we can calculate price indices (or price ratios) for months s and t, we have to count the number of pairs in which price information is available in both months. In this case there are 3 such pairs (for SG2, SG7 and SG8). For the remaining 8 combinations no price indices (price ratios) can be computed, as for either month s or month t a price is lacking. ■

In the example only two months are considered. If we are considering a time window T with $|T|$ months there are $\binom{|T|}{2}$ pairs of months to consider to judge the overlap.

Example

In Figure 5.6.2 a group with 11 subgroups and a time window of 10 months. For each month in this window it is indicated for which subgroups price information is available: black (or 1) if there is, white (or 0) if there is not. ■

5.6.2 Fillings for a group with 11 subgroups in a period of 10 consecutive months, presented in two ways.



	1	2	3	4	5	6	7	8	9	10
SG1	1	0	1	0	1	0	1	0	1	0
SG2	1	1	1	1	0	1	1	1	1	1
SG3	0	0	0	1	1	1	0	1	0	1
SG4	1	1	1	0	1	0	1	0	1	0
SG5	0	1	0	1	0	1	1	1	0	1
SG6	1	1	0	0	1	0	1	0	1	0
SG7	1	1	1	1	1	1	0	1	1	1
SG8	1	0	1	0	0	1	1	1	1	1
SG9	0	1	0	1	1	1	0	1	0	1
SG10	1	0	1	0	1	0	1	0	0	0
SG11	0	1	1	1	0	1	1	1	0	1

In order to compute the overlap we introduce an indicator function δ_i^t, for subgroup i and month t. Let $\delta_i^t = 1$ if there are data (prices) for subgroup i in month t, and $\delta_i^t = 0$ otherwise. Similarly for month s we have δ_i^s. Then $\sum_{i=1}^m \delta_i^t \delta_i^s$ counts the number of subgroups that are nonempty in both month t and month s, and for which therefore price indices (price ratios) can be calculated. So the total number of indices that can be computed for different month pairs in period T is: $\sum_{t<s\in T} \sum_{i=1}^m \delta_i^t \delta_i^s$. The maximum number of pairs for which an index can be computed is $m\binom{|T|}{2} = \frac{1}{2}m|T|(|T|-1)$. So the number of pairs for which no price index can be calculated is

$$\tfrac{1}{2}m|T|(|T|-1) - \sum_{t<s\in T} \sum_{i=1}^m \delta_i^t \delta_i^s. \tag{5.3}$$

Relative to the maximum number of pairs this number equals:

$$1 - \frac{\sum_{t<s\in T} \sum_{i=1}^m \delta_i^t \delta_i^s}{\frac{1}{2}m|T|(|T|-1)}. \tag{5.4}$$

As a penalty term, we want to minimize this expression, which amounts to maximizing the overlap.

It should be noted that (5.4) gives equal weight to all pairs of months. The question is whether that is reasonable. Certain price comparisons involve months that are not far apart and others those that are widely separated. It seems reasonable to attach more weight to those pairs that are not far separated than to those that are (see Subsection 5.7).

In view of this remark, we could replace (5.4) by one that only measures overlap on a MoM-basis, that is for adjacent months. They yield, in a sense, the most important pairs, because they are the month pairs where one can expect the largest overlaps (in terms of GTINs) for the various subgroups. We then obtain a measure that is closely related to the flow concept (see Willenborg, 2017a):

$$1 - \frac{\sum_{t=1}^{|T|-1} \sum_{i=1}^m \delta_i^t \delta_i^{t+1}}{m(|T|-1)}. \tag{5.5}$$

In the next section we elaborate on this idea of overlap in a subgroup for different pairs of months, making overlap a matter of degree instead of a 'yes-no affair'.

5.7 Degree of overlap

It is possible to refine the notion of overlap defined in the previous subsection. The idea is to actually look beyond the surface of the subgroups to see how they are composed of GTINs. If we have two months, s and t , and a subgroup i, we can look at the number of GTINs in this subgroup present in both months. So if $\mathcal{S}_i^t, \mathcal{S}_i^s$ are the sets of GTINs in subgroup i in months s and t, $\mathcal{S}_i^t \cap \mathcal{S}_i^s$ is the set of our interest. In particular we are interested in its size, i.e. $|\mathcal{S}_i^t \cap \mathcal{S}_i^s|$. Or rather, we are interested in its relative size

$$\mathcal{R}_i^{s,t} = \frac{|\mathcal{S}_i^t \cap \mathcal{S}_i^s|}{|\mathcal{S}_i^t \cup \mathcal{S}_i^s|}. \tag{5.6}$$

Of course, to be able to apply (5.6) it is assumed that $|\mathcal{S}_i^t \cup \mathcal{S}_i^s| > 0.$, which means that subgroup i should be nonempty for months s or t. We use (5.6) to quantify the concept of degree of overlap in a subgroup. Obviously, it holds, for any subgroup i, that $\mathcal{R}_i^{s,t}$ is a nonnegative rational number, and $0 \leq \mathcal{R}_i^{s,t} \leq 1$, for all s and t, $\mathcal{R}_i^{s,s} = 1$, for all s, $\mathcal{R}_i^{s,t} \downarrow 0$ if $|s - t| \to \infty$. In practice it will be the case that $\mathcal{R}_i^{s,t} = 0$ if s and t are far enough apart, because the lifetimes of the GTINs are finite.

$\mathcal{R}_i^{s,t}$ is useful for determining for subgroup i , for which months s and t direct price comparisons should be computed: if this number is close to one, there is enough overlap in terms of GTINs to warrant the computation of a price index $P_i^{s,t}$. This idea has been suggested and used before by the author in connection with the cycle method (see e.g. Willenborg, 2017b); see also Willenborg (2017d). In this context it was used to derive a weight matrix which controls the adjustment of a nontransitive price index to a transitive one.

For applications in penalty functions (see Subsection 5.8), we should rather look at the complement of $\mathcal{R}_i^{s,t}$, that is

$$1 - \mathcal{R}_i^{s,t} = 1 - \frac{|\mathcal{S}_i^t \cap \mathcal{S}_i^s|}{|\mathcal{S}_i^t \cup \mathcal{S}_i^s|}. \tag{5.7}$$

Remark

$\mathcal{R}_i^{s,t}$ is a simple measure to quantify overlap, within a subgroup at different periods in time. But it is feasible to apply, as such, although considerably more effort is required to compute this quantity than (5.4) or (5.5).

A more sophisticated version of $\mathcal{R}_i^{s,t}$ would take relaunches into account as well. But this would complicate matters considerably (taking the discussion in Subsection 4.2 into account), in fact beyond what is practically feasible and manageable. ■

5.8 Penalty functions

We want to use penalty functions to quantify the various aspects that play a role by selecting a desirable stratification of a product group. In fact, the class of penalty

functions that we consider is based on homogeneity, detail and overlap, as discussed in previous subsections of the present section. The idea is that more desirable stratifications have smaller scores for a penalty function.

There is no need to include other aspects such as persistence at the group level and stability at the subgroup level. These aspects can be judged before, by inspection (for known data) and introspection (for future developments). If any of these aspects would be absent, there is no sense in continuing with such a group or such a stratification. First, there should be some adjustments at the group or stratum level. So we tacitly assume that persistence at the group level and stability at the subgroup level is secured.[30]

To find a suitable class of penalty functions we add all the terms for each of the three aspects together, using tuning weights for two of the three terms, that allow us to find the right balance between them.[31]

Proceeding in this manner we find the following class of penalty functions for a period T of $|T|$ consecutive months for a group of products with m subgroups:

$$\mathcal{P}_T = \frac{\sum_{t\in T} vc_t}{|T|} + \alpha m + \beta\left(1 - \frac{\sum_{t<s\in T}\sum_{i=1}^{m}\delta_i^t\delta_i^s}{\frac{1}{2}m|T|(|T|-1)}\right), \tag{5.8}$$

where m is the number of subgroups in the group, and $\alpha, \beta > 0$ are tuning parameters, to balance the various terms relative to each other; they have to be found empirically and depend on judgment.

If we take (5.5) instead of (5.4) as the overlap component, we would obtain the following class of penalty functions:

$$\mathcal{P}_{T,MoM} = \frac{\sum_{t\in T} vc_t}{|T|} + \alpha m + \beta\left(1 - \frac{\sum_{t=1}^{|T|-1}\sum_{i=1}^{m}\delta_i^t\delta_i^{t+1}}{m(|T|-1)}\right). \tag{5.9}$$

Another class of penalty functions would be obtained if we would replace the overlap terms in (5.4) or (5.5) by (5.7), expressing a degree of overlap:

$$\mathcal{P}_{T,dov} = \frac{\sum_{t\in T} vc_t}{|T|} + \alpha m + \beta\left(1 - \frac{\sum_{t<s\in T}\sum_{i=1}^{m}\frac{|\delta_i^t\cap\delta_i^s|}{|\delta_i^t\cup\delta_i^s|}}{\frac{1}{2}m|T|(|T|-1)}\right). \tag{5.10}$$

Another approach[32] is possible, namely one in which one of the terms in (5.8), (5.9) of (5.10) are minimized, under the condition that the two remaining ones are bounded from above by certain thresholds. For example, choose a boundary value

[30] The first is easy to ascertain. The second is not in the way discussed before. But we assume that non-stable stratifications will also score badly on the penalty functions that we discuss below (that is, high values).

[31] This is the reason why we talk of a class of penalty functions. There are many such functions, parameterized by the parameters $\alpha, \beta \geq 0$

[32] Suggested by Sander Scholtus.

for the sum of detail and overlap, and then minimize the average variation coefficients (first in term in (5.10)) for stratifications that satisfy this restriction. This yields a problem that can be solved by OR-techniques, as is the case for the other optimization problems mentioned in the present section.

Whatever class of penalty functions is chosen, experiments have to be carried out on real data to see if reasonable results can be obtained when using them. And also to find out the computational burden when they are applied. Then it also becomes clear how to tune the weights α and β, and how easy / difficult it actually is to find 'optimal' / useful values.

6. Discussion

In the present report we discussed the problem how to stratify groups of products (typically COICOPs or unions of COICOPs) in such a way that they are suitable for index computations. It was argued that groups of items should be used instead of GTINs, in case the populations are dynamic, which in practice they always are, although there are differences in dynamics between the various product groups.

At the GTIN level one is limited in making price comparisons, as the lifespan of each (or most) is limited. Considering only separate GTINs is incorrect as it misses price development due to the mechanism of relaunches. Trying to repair this deficiency by matching perceived relaunches to their perceived predecessors is complicated, error-prone and, generally, not very attractive.

A better option is to stratify a product population into stable strata (or subgroups). Instead of looking at the development of prices at the GTIN level, the alternative is to consider the price development at the subgroup level. The subgroups are now considered as composite products that may change in composition, while retaining their character. Such changes should make them comparable over time as similar products.

The question to be solved remains: what properties should the stratification have to be acceptable for price index computations? In the present paper this question is answered by considering relevant aspects that should be taken into account. It is argued that the group (say a COICOP) should be persistent (or continuous) , whereas the subgroups should be stable and sufficiently homogeneous, to start with. These aspects should be judged by inspection and introspection. Furthermore the following aspects should be taken into account: the homogeneity of the subgroups, the level of detail achieved by the stratification, and overlap which measures how many price comparisons can be made. The method proposed is independent of a particular price index method. The advantage of this is that we can change the index method but maintain the stratification. This allows us also to compare different index methods using the same stratification.

Also classes of suitable penalty functions are derived, based on homogeneity, detail and overlap, as are relevant for a stratification. Such penalty functions can be used as objective measures to judge their suitability of stratifications for price index computation. They have a similar function as the AIC (= Akaike Information Criterion)[33] or the BIC (Bayesian Information Criterion)[34] for the selection of suitable statistical models. They could also be used in a dedicated software tool to find good stratifications. The assumption is that features found in product descriptions are used to define strata. If there are many such features the question is which ones to pick. Various examples of descriptions used in practice (in scanner data or web data, which are a rich source of supplementary information) are presented to give the user an idea of the variety of what is available.

The present paper is one of ideas, but not one that puts these ideas to the test using real data. That should be the next step: apply the ideas presented to real data, and modify them, where necessary.

References

Willenborg, L. (2017a). Quantifying the dynamics of populations of articles. Discussion paper, Statistics Netherlands, The Hague, The Netherlands.

Willenborg, L. (2017b). Transitivizing elementary price indices for internet data using the cycle method. Discussion paper, Statistics Netherlands, The Hague, The Netherlands.

Willenborg, L. (2017c). Characteristics of pastries in product descriptions (in Dutch). Discussion paper, Statistics Netherlands, The Hague, The Netherlands.

Willenborg, L. (2017d). Price indices and transitivity. Discussion paper, Statistics Netherlands, The Hague, The Netherlands.

[33] See e.g. https://en.wikipedia.org/wiki/Akaike_information_criterion
[34] See e.g. https://en.wikipedia.org/wiki/Bayesian_information_criterion

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information