Discussion Paper

Quantifying the dynamics of populations of articles

2017 | 10

Leon Willenborg

Content

Summary

Most populations of articles change in due course, due to innovation or changing demands of consumers. Existing articles disappear from the market after some time, and new ones are introduced, continuously. Each article has a finite life span. In classical index number theory populations were assumed to be static, that is without change, which in fact means that articles 'live' forever. This ideal situation, however, is quite unrealistic. Especially in our times article populations tend to change constantly, some even dramatically so. The choice of a suitable index formula depends on how dynamic such a population is. Not all index methods are suited for highly dynamic populations. This paper does not go into the choice of a suitable index method for a certain dynamic population. It only deals with the problem of characterizing the dynamics of populations of articles. Several characteristics are proposed, and are applied to a few populations of articles, with different dynamic behaviour. It is shown that certain statistical problems arise when we have to deal with samples of articles, instead of the entire population.

Keywords

Dynamical populations, characteristics, change, flow, inflow, outflow, births, deaths, cohorts, censoring, decay.

1. Introduction

This document is written in the context of an on-going research effort at CBS to calculate several price indices, apply them to the same data sets, and compare the results (cf. De Haan et al, (2016) and Chessa et al. (2016)). The ultimate aim is to get insight into how these results differ when applied to different article populations, and what explains these differences. The idea is that the dynamicity of the population being studied holds the key to the answer, or at least one of the keys. But this begs for another question to be answered: how does one adequately describe the dynamics of a population, in particular of articles in scanner data (transactions) or of articles offered by web shops? This is a first step to answer the more complex question of how the dynamics of article populations can be applied to understanding the behaviour of price indices. And also how it can guide the choice of a price index method in case of a highly dynamic population of articles? The present report only focuses on the first step. To answer the last question would require more research. Maybe a full answer to the guiding question requires not only a characterization of dynamicity of article populations, but also an understanding of the nature of the demand for the items that, collectively, define the various article populations: how is this demand guided by product characteristics and prices, in particular items temporarily on offer for discount prices. This topic, however, is beyond the scope of the present report.

In the present paper some characteristics of dynamic populations of articles are proposed. They are studied using real data from a retail shop in the Netherlands. These are the same data used in Chessa et al. (2016). It concerns 4 types of articles: office supplies, bed clothing, pastries and men's T-shirts. These characteristics are applied to the four articles mentioned above, to get a feeling of what they mean, and whether they seem to be useful or not. To be useful they should be able to distinguish between different dynamic behaviour of article populations. For instance to see if a population has a large churn or not, or whether the introduction of new items is on a regular basis or has irregular bursts of introductions of new items, etc. Whether these characteristics ultimately will be useful to make a reasoned choice for a suitable price index remains to be seen. A separate use for such characteristics would be in a dashboard application, so that analysts at the CPI[1] department can get a quick overview of the dynamic behaviour of populations of articles in terms of a select set of salient features. This can help in identifying populations that have to be watched carefully and populations that do not need so much attention and supervision.

The remainder of the paper is organized as follows. In Section 2 some thoughts are devoted to dynamic populations of articles and the problems they pose for price index methods. Section 3 introduces measures for dynamicity that are considered in the present report, as well as related issues and preparatory material. These measures are applied to scanner data from a Dutch retailer. The results are collected

[1] CPI= Consumer Price Index

in the nine appendixes at the end of the paper. They contain the results from various analyses applied to the data, using Excel and the R package, in harmonious co-operation. Comments on the results can be found in Section 4. The results clearly show different dynamical behaviour between the four article populations studied. Section 5 concludes the core of the report with a discussion of the main findings. It also contains some suggestions for future activities.[2]

2. Dynamic populations of articles

Standard price index theory tacitly assumes that population of articles described is static, that is, does not change in composition in due course. In practice this condition is never met. Populations of articles do change in time, and sometimes at a (very) rapid rate. Existing articles at the end of their life cycle are replaced by newer versions. Sometimes entirely new products are introduced to the market.

The existence of dynamic article populations implies that classical index theory cannot be applied straightaway. This applies in particular to the well-known price indices named after Laspeyres, Paasche, Fisher and Törnqvist. One can find adaptations to apply them to dynamic populations. But these are fundamentally different from the original ones for a static population.

This points to the fundamental problem with dynamic populations of articles. The exact same articles may not exist at different points in time. So in order to make a comparison one is compelled to make some compromises: only a subset of articles is taken into account for a price comparison for a given pair of months. Or one is forced to compare similar, but different articles. In the latter case the question arises when articles can be considered similar, or similar enough, in an operational way? New price index formulas may have to be invented, or existing ones may have to be modified. Or articles need to be grouped in such a way that the groups formed are both sufficiently homogeneous and stable. Stability means that an article population exists for a longer period of time, much longer than the average life span of the items in that population.

The choice of a suitable price index formula or method for a particular population of articles is somehow related to the dynamics of the population at hand. For it is intuitively clear that the dynamics may be different for different groups of articles: for agricultural products there may be a seasonal influence; for smartphones and other ‘hot’ gadgets new types come to market regularly and disappear quickly; clothing is highly sensitive to change in fashion, but there are also seasonal changes.

[2] The present document was reviewed by Sander Scholtus.

We can get a feeling for a dynamic population in different ways. For instance we can look at local behaviour, whether items are being sold in consecutive months or not, or whether there are gaps in the months in which they have been sold (in scanner data) or are available (in web data). This leads to the idea of 'flows', that is elaborated in the present paper. Another way to understand a dynamic population is to study the introduction of new products to the market and their withdrawal. How often are new items introduced to the market, or withdrawn? Are these processes steady, or are there bursts of introductions, or withdrawals? And for how long are products on the market? These questions can be answered for all items in an article population given equal weights, or by weighting with their economic importance (measured by turnover). It is insightful to view each article population of being a set of cohorts, say of items that were introduced to the market in the same month, or that were withdrawn from the market in the same month. An appropriate description of this structure gives good insight in the kind of population one is dealing with. In particular one can study a specific cohort, for instance the one consisting of all items present in a certain month, and study how they, or some of their characteristics, behave, more particularly, die out or fade away. In this report we shall consider the items present at the first month of the period of study.

Because we have to study our dynamic populations by means of a time window – a period of consecutive months for which we have data – we are faced with the effects of censoring, which is a result of the fact that we cannot look beyond that time window. It introduces biases in the data, for instance of life spans (which are cut-off at the length of the time window).

In the next section we shall delve a bit deeper into these matters.

3. Measuring and quantifying dynamicity

3.1 Characteristics

In order to characterize the dynamics of populations of items we consider a few options. These are:

- Development of sales status
- Development of value / turnover
- Development of items sold
- Birth of items
- Death of items
- Age of items
- Flow

These characteristics can (in principle) be computed for any population of articles, including subpopulations. Special subpopulations can give good insight into the dynamics of a population, namely cohorts. These are subpopulations with elements that have some characteristic in common. In this report these characteristics are related to events, and in particular when they happen: birth, death, existence in a particular period, etc.

- Cohort structure
- Decay of cohorts, based on some characteristic of the cohorts

Subpopulations that we consider in our analyses are m-cohorts and M1-cohorts (see Section 3.4 for an explanation of these concepts). In the present report we only illustrate some of the characteristics to a few cases. Some characteristics have not been calculated at all.[3]

3.2 Articles

In the present report we apply the dynamicity characteristics to four different articles. The data used are scanner data from a Dutch retailer. The articles and the abbreviations used, are listed in Table 3.2.1.

3.2.1 Articles and their abbreviations

Article	Abbreviation
bed clothing	BC
men's T-shirts	TS
office supplies	OS
pastries	PA

The abbreviations are used in the tags for the results in the appendices.

3.3 Aggregation

In the present report we consider two levels of aggregation:

- the EAN[4] level and
- a group level (indicated by GRO).

The indication 'EAN' or 'GRO' is also used in the tags in the appendices.

The groups are formed by combining EANs in some sensible way. The EAN-level is the lowest aggregation level. All groups of items at other levels of aggregation can be built from this level.

It is interesting to study the behaviour of a dynamic population at several levels of aggregation. The higher the level of aggregation, the higher one may expect, the more stable the groups are, that is, the longer they are expected to exist. They exist

[3] Due of lack of time. However, they merit exploration.
[4] EAN = European Article Number is a 13-digit barcode symbology.

in a particular month if they generated turnover in that month. The challenge for an article is to find an aggregation that is (fairly) stable on the one hand and not too coarse (heterogeneous) on the other. If it is too coarse, it may produce inferior price indices. So there is a goal to find a right balance between continuity and detail.

3.4 Cohorts

Certain subgroups are of interest in studying the dynamics of a population. One of these are m-cohorts. An m-cohort is a group of individuals (items, in our case) that have a common defining characteristic. For instance, they all were 'born' in the same month, they all were 'alive' in a certain month, they all 'died' in the same month, etc. These are examples of cohorts that we consider in the present report. Typical for these cohorts is that they concern subpopulations of items that are closed in the following sense: no new individual can become part of it as time passes. In another way such a cohort is not closed: individuals can drop out of it because they die.

Note that m-cohorts of items that are born in the same month form a partition of the population, as each item is born in exactly one month. The same is true for subpopulations of items that 'die' in the same month. The items born in the same month need not die in the same month as well. Similarly, items that die in the same month need not be born in the same month.

Cohorts of items that were 'alive' at a certain month, say month 1 of the time window, do not have this property. Such cohorts form a cover of the population, but they are not necessarily disjoint. In the present report we consider M1-cohorts, consisting of items (within an article group) that were alive (= present) in month 1. The products present in the first month consist of products that were introduced in that month, or in earlier months. So the set of item present in the first month is actually a union of m-cohorts, say $M_1 = C_1 \cup C_0 \cup C_{-1} \cup \dots$, where C_i is the m-cohort of items that were born in month i. We shall call M_1 a composite m-cohort for that reason.

Remark. For biological subjects events such as birth and death are different from our case where we are dealing with items of articles. An item may not be present at every month in its lifespan. An item may be temporarily out of stock. This may complicate the determination whether an item is alive when it has not been observed for 1 or 2 months. The same problem is about determining when an item was born. Usually this can only be determined after some time has elapsed. The same uncertainty exists for quantities depending on these parameters, such as 'age'. ■

3.5 Variables

Sales status

For each item in an article population we can indicate for each month whether it was sold (indicator = 1) or not (indicator = 0). If for each month we count the total items

in the article group for which the indicator was 1 (i.e. that have been sold in that month), we obtain the sales status. This is a simple characteristic concerning the dynamicity of an article population. If all items of an article have been sold every month, the sales status is a constant function. If this varies greatly, the sales status also fluctuates wildly.

It should be stressed that 'sales status' is not to be confused with 'numbers of items sold'. If an item sells 150 items in a particular month m, the sales status is 1 for this item in month m. If this item was not sold in month m its sales status is 0, and the numbers sold would also be 0. So aggregates of 'sales status' and 'numbers sold' are the same when they are both 0 but typically differ when they are not zero, the latter being bigger than the former.

Economic value

The sales status as defined in section 3.1 gives equal weight to each item in the article population. This may give a somewhat distorted picture, as economically less important items have the same weight as the important ones. But it is possible, however, to give each indicator a weight equal to the value of its turnover in a particular month. Then we can get a pictures of the development of the economic value associated with the articles under consideration. Because of the additivity of economic value (turnover), we obtain the same development at the EAN level or the group level.

Of course, 'turnover' can also be directly considered as a variable of interest. There is no compelling reason to consider first sales status and then use weights to differentiate between economic importance of items, as above.

Number of items sold

Another parameter that is of interest to study the dynamic behaviour of an article population is the development of the number of copies of each item[5] sold each month. But we will also call these copies items (as 'books' and 'copies of books' are often confused in everyday usage). This can also be seen as a modification of the sales status. Instead of counting each item of an article in a particular month for 1, we can give it a weight equal to the number of (copies of) items sold in a particular month.

Age

We define the lifespan roughly as the period between the first and final appearance of an article.[6] This would be fine if the time-window that we use would be infinite. In practice it is not. As a result there typically is an effect due to censoring. At the beginning of the window, we may see articles who started life before the start of the

[5] This is a somewhat awkward expression, to avoid possible misunderstanding. Each article consists of items. This is an abstract concept. In reality copies ('realisations') of this (abstract) item are the things that are sold. Compare this to a book (an abstract concept) and a copy of a book (a concrete realisation).

[6] We have indicated in Section 3.4 why determining 'age' can be problematic to determine correctly, as a result of a similar problem with the determination of 'births' and 'deaths'.

window. At the end, articles who finish their life at some point after the end of the window. We do not know anything that happened outside this window. This knowledge may come later when, month-by-month, the window is shifting to the right. This also diminishes the effect of the censoring of some characteristics of articles that were observed close to the start of the window.

3.6 Weighting

For some analyses it is desirable that the economic significance of items is taken into account. In the present subsection we describe a method to do so. We use a set of weights based on turnover. Let V be the matrix of values (turnovers) for the items in some article group. Suppose

$$V = \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mn} \end{pmatrix}. \tag{3.1}$$

If the monetary value would be the same, we could take as a simple economic value of item i the expression

$$\widehat{w}_i = \frac{\sum_{j=1}^{n} v_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij}}. \tag{3.2}$$

This weight is simple to calculate. It neglects the variation of purchasing power of money over time. In a more refined approach this can be taken into account. But we have used (3.2) in our calculations.

As an example of the application of these weight, consider the m-cohort for month j. Suppose that this contains the items $i_1, \dots, i_h$. Unweighted, the number of items in this m-cohort equals h. Using weights, this cohort has 'size' $\sum_{j=1}^{h} \widehat{w}_{i_j}$. Similarly, if the life-span (age) of these articles is $l_{i_1}, \dots, l_{i_h}$, the weighted life-span of this cohort is $\sum_{j=1}^{h} l_{i_j}\widehat{w}_{i_j}$.

An easier way of getting some insight into the effects on the life-span of articles taking their economic importance into account is by looking only at the items that sold well. For instance we order the items according to turnover or weight (as in (3.2), which is proportional) and select the top n% for various values of n. For each of these sets we determine the life-span in the same way as for the entire set.

3.7 Censoring

Some of the results that we obtain from the data that we use are biased. This is caused by censoring, left and right, due to the fact that we consider a dynamic population of items only through a finite time window (in our case of 50 consecutive months). In this situation it impossible to know what happens outside the time window. So at the left of the window we do not know when items were 'born'. If they

are present in month 1, we do not know whether they were born in that month or before. This is the problem of left-censoring. We have similar problems at the right of the window. We do not know what happens then. When items exist at the final month of the window we do not know when they will die. Similarly we have problems in estimating the age distribution of items. Ages longer than the width of the window size (in our case 50 months) cannot be observed. We only know that if an age of 50 is observed, it means that the age in reality is at least 50. More generally, if an article has a lifespan not entirely within the window, its age is actually at least the observed age. So censoring introduces a downward bias on ages distributions.

If the time window in which we study these articles would be very big ('infinite', so to speak) then we would have no trouble in identifying the various m-cohorts. However, because in practice we always deal with a finite window, we may have problems identifying the m-cohorts, due to censoring.

We have a similar censoring problem at the other boundary (bounding the present from the future). Items available in the final month of the window may 'die' in that month, or at a later month. In both cases we have trouble dissecting the sets of items 'present in month 1' and 'present in month f' (where f is the final month of the current window) into m-cohorts. But we can consider C_1 and C_f also as m-cohorts, composite m-cohorts. They also have the property that they can only decrease in size, because of 'dying' items. But it is not only problematic for sets of items present at the beginning or end of a window, how to divide them in m-cohorts. Months close to these boundaries also pose a problem, because we cannot be sure that an item was introduced for the first time in a particular month. It may have been earlier. This uncertainty is due to the fact that once an item has been introduced to the market, it need not be present every month until it 'dies'. It may be temporarily unavailable, or escape the claws of the web scrapers.[7] But the longer an item was not in the market before a particular month it was available, the more likely it is that it was indeed first introduced then.

3.8 Flows

For (groups of) items we use two states: 'absent' (coded by '0') and 'present' (coded by '1'). If we look at the availability of items during the time window, we see a pattern of 0's and 1's. A simple statistic for each (group of) items is the number of months it is present in the time window.

Let A_i be the set of articles from the population P in month i. Let $|A_i|$ be the size of A_i, that is, the number of elements (articles in our case) it contains. The following sets are interesting to describe the dynamics of the population P for pairs of months i, j with $i < j$ in a period $W = \{1,2, \dots, n\}$:

- $A_i \cap A_j$ (flow): the set of articles present in both months i and j;

[7] In practice the distinction between these two possibilities is not to make. It would require special effort to find out. Such effort is typically not made, as it is of not worth the trouble.

- $A_j \backslash A_i$ (inflow): the set of articles present in month j, but absent in month i;
- $A_i \backslash A_j$ (outflow): the set of articles present in month i, but absent in month j.

More specifically, we apply this to successive months i and i+1 ("month-on-month", abbreviated as "mom"). This implies a considerable reduction of possible pairs.

In Table 3.8.1 the three flow situations at month i in period W are considered, and relative measures to quantify each of these flow situations.

3.8.1 Flow, inflow and outflow at month i.

Types	Set	Month i	Month i+1	Abs. meas.	Rel. meas. 1	Rel. meas. 2
Flow	$A_i \cap A_{i+1}$	present	present	$\lvert A_i \cap A_{i+1} \rvert$	$\frac{\lvert A_i \cap A_{i+1} \rvert}{\lvert A_i \rvert}$	$\frac{\lvert A_i \cap A_{i+1} \rvert}{\lvert A_i \cup A_{i+1} \rvert}$
Inflow	$A_{i+1} \backslash A_i$	absent	present	$\lvert A_{i+1} \backslash A_i \rvert$	$\frac{\lvert A_{i+1} \backslash A_i \rvert}{\lvert A_{i+1} \rvert}$	$\frac{\lvert A_{i+1} \backslash A_i \rvert}{\lvert A_i \cup A_{i+1} \rvert}$
Outflow	$A_i \backslash A_{i+1}$	present	absent	$\lvert A_i \backslash A_{i+1} \rvert$	$\frac{\lvert A_i \backslash A_{i+1} \rvert}{\lvert A_i \rvert}$	$\frac{\lvert A_i \backslash A_{i+1} \rvert}{\lvert A_i \cup A_{i+1} \rvert}$

In a static population, there is only flow for every month (no new articles, no dis-appearing ones). In the dynamic case there is inflow or outflow in at least one month, but typically in several ones. One could use these three mom flow types to characterize the dynamics of article populations. Are the inflow and outflow approximately equal, are they periodic, what is the relative size of the flow, etc.?

Note that the relative measures of type 2 in Table 3.8.1 have the attractive property that they add to unity, each month:

$$\frac{|A_i \cap A_{i+1}|}{|A_i \cup A_{i+1}|} + \frac{|A_{i+1} \backslash A_i|}{|A_i \cup A_{i+1}|} + \frac{|A_i \backslash A_{i+1}|}{|A_i \cup A_{i+1}|} = 1\,, \tag{3.3}$$

for each i in W. Identity (1) holds as the sets $A_i \cap A_{i+1}$, $A_{i+1} \backslash A_i$ and $A_i \backslash A_{i+1}$ form a partition of the set $A_i \cup A_{i+1}$. Although this additivity property is attractive, we did not use these flow measures in the present paper, but only . The remaining three measures of flow, inflow and outflow lack this property, and are for that reason not used. But they are certainly useful flow measures as well.

We can also make a binary division between flow and the rest (either inflow or outflow, representing mutation).

So far we have concentrated on unweighted methods. This results in the determination of set size by simply counting the elements it contains. For a set A this number is denoted by $|A|$. This method gives equal importance all items, irrespective of their importance, expressed by some weight. However, we could count each element with its 'economic value' (see the subsection under this name in Section 3.5), and the size of a set is then defined as the sum of the weights associated with its elements:

$$|A|_w \triangleq \sum_{i \in A} \widehat{w}_i, \tag{3.4}$$

where the $\widehat{w}_i$ are as in (3.2).

3.9 Decay

What comes to mind when thinking about the development of a cohort is a decay process like in a radioactive substance such as uranium. This is the case for several characteristics of a cohort, such as size, value (turnover), number of items sold, etc. Each of these values will eventually 'die out' or 'fade away', as the cohort has a finite life expectancy. This tendency to decrease is not deterministic but stochastic: the average of the phenomenon is likely to be a decreasing function, but the actual function typically is fluctuating with a decreasing trend. This trend is what we want to describe.

A decay model can be applied to m-cohorts. But for the four articles we consider, they tend to be relatively small. We rather would like to apply this idea to a bigger set of items, namely the M1-cohorts.

Inspired by physics, we start with a simple exponential decay model and study it a bit. As it turns out this is not the appropriate model to fit the data. The decay is typically a bit slower than this model suggests. But starting with this model we can easily switch to another model that seems to be more suited to describe the decay process we are considering here.

The model we start with is the following one:

$$y(t) = y(1)e^{-\alpha(t-1)}, \tag{3.5}$$

where y denotes the variable of interest at month t. The parameter α symbolizes the half-life. It can be estimated for each article and each aggregation level. The higher the aggregate level is that is used the higher the half-life and the more stable the group. So α can be used as a parameter to characterize the stability of a cohort.

By taking (natural) logarithms we obtain from (3.5):

$$\log y(t) = \log y(1) - \alpha(t-1), \tag{3.6}$$

or

$$\log \frac{y(t)}{y(1)} = -\alpha(t-1). \tag{3.7}$$

To estimate α we assume the restricted linear model

$$z(t) = -\alpha(t-1) + \varepsilon_t, \tag{3.8}$$

where $z(t) = \log \frac{y(t)}{y(1)}$. The model is restricted because it is only about estimating one parameter instead of two (the intercept is left out of (3.8)).

We can estimate α, for instance, by minimizing the expression

$$\sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} (\zeta_t + \alpha(t-1))^2, \tag{3.9}$$

where ζ_t is the empirical equivalent (observed value) for the theoretical $y(t)$. The result is, for each group of items, an estimate of the decay parameter:

$$\hat{\alpha} = -\frac{\sum_{t=1}^{n} \zeta_t}{\sum_{t=0}^{n-1} t} = \frac{-2\sum_{t=1}^{n} \zeta_t}{n(n-1)}. \tag{3.10}$$

In our case $n = 50$.

If there is reason to weigh the various residuals differently, we can use a weighted alternative to the object function (3.7):

$$\sum_{t=1}^{n} w_t \varepsilon_t^2 = \sum_{t=1}^{n} w_t (\zeta_t + \alpha(t-1))^2, \tag{3.11}$$

where the weights $w_t > 0$, for $t = 1, \dots, n$. A reasonable choice of weights would be: economic importance reflected by the total turnover of the (composite) m-cohort at time t. This results in a tighter fit of the model at the times when the turnover of the article was more substantial than at months that is was not. To keep things simple we shall only consider unweighted object functions like (3.9), while keeping those of type (3.11) in mind.

Looking at the data (see Appendices G, H and I) , however, it can be questioned if (3.5) is a proper model in most cases.[8] In general, a model with a linear average decay seems to be more suitable. This can be quite easily arranged using the model above as a starting point, if we take

$$z(t) = y(t) - y(1), \tag{3.12}$$

instead of

$$z(t) = \log y(t) - \log y(1), \tag{3.13}$$

as in (3.8). With this simple modification the estimation procedure sketched above can be easily adapted to the new situation.

Note. There was no time to study the decay feature more extensively when writing the current report. It remains for the future to investigate this subject more extensively and to estimate decay parameters for various article groups. They can be useful parameters to characterize the stability of groups of articles. ■

[8] The value development for the M1-cohort for the T-shirts (see TSEANMO1 value) shows rapid decline, and model (3.5) seems to be applicable.

4. Comments on the results

4.1 Entire population

Sales status

In case of the four articles, and two levels of aggregation, we have collected plots of the results in Appendix A. The pictures clearly show different dynamics for the four article populations, at each of the two levels of aggregation considered. However, the reader should take the scales of the y-axis into account before rushing to conclusions. For instance the fluctuations for TSEAN (between 300 and 800) are much more significant than for PAGRO (between 33 and 39). In some cases the phenomenon seems to fluctuate around some mean (OSEAN, BCEAN, PAGRO). In other cases there is a downward slope (PAEAN, TSEAN). In yet other cases, there is an initial increase and after some time a more or less stable situation emerges (OSGRO and TSGRO). The TSEAN case shows the most interesting behaviour: wild fluctuations until month 19 and smaller ones with a lower mean after this month.

Economic value

The results for this section can be found in Appendix B. The results presented are identical for the EAN and the GRO level (and any other level of abstraction), because value is an additive characteristic. Note that the scales of the plots differ considerably, so this may give false impressions when looking at the pictures. For OSEAN, BCEAN and PAEAN, the results seem to fluctuate about a more or less fixed mean. However in case of TSEAN there seems to be an upward trend.

Number of items sold

Appendix C contains the results for the number of items sold. As in the case of the previous section we obtain the same results for the EAN level and the group level, due to the additivity of the 'number of (copies of) items' parameter. For three of the four articles, except T-shirts, the results fluctuate around a stable mean. The development of the T-shirts results are somewhat different. After month 18 there is a qualitatively different development noticeable: higher peaks and an upward sloping trend of the number of items sold.

Age distribution

In Appendix D the age distributions of the four articles, at the EAN and group level, are presented. Most of the results concern the unweighted case. Although this give a complete overview of all 'observed' ages, it is also somewhat misleading as the items are not distinguished in economic importance. To do this, we have also computed the age distributions for the top n% economically most important items, for $n = 75, 80, 85, 90, 95$. This shows that a lot less items are involved. As is shown in a separate figure (in Appendix D) the turnover of the items in the men's T-shirt population is very skewed. Most of the items contribute very little to the total turnover. It is obvious that the numbers of these sets are much smaller than the entire set. Comparing it to the n=100% case (presented in the first table in Appendix D, the right

most column) we see that quite a few (almost 700) items have ages less than 15 months. For the top 85%, top 80% and top 75% contributors, the minimum observed age is 15 months, and for the top 95% there is one item with a shorter age (13 months).

It should be noted that the results in Appendix D have a downward bias, due to censoring, both left and right. It requires further analysis (and use of models) to unravel the effect of this censoring in the data shown. This is not attempted in the present paper, but is left as a future task.

Note. Instead of considering several important subpopulations of items for each article, one could use the method discussed at the end of Section 3.8. That allows one to give a weight to the age of an item (namely the weight associated with that item). For each age group, the total weight is equal to the sum of the weights of the items in that group, as in (3.4). It is then easy, by using various thresholds, to distinguish 'important' age groups (total weight above the threshold) from those that are less so (total weight below the threshold). ■

Flows

The flow results have been collected in Appendix E. They clearly show that there is typically more dynamics in case of articles at the EAN level than at the group level. But also comparing the results at the same level of aggregation among the four articles shows that there are differences in behaviour. In particular the dynamics of the T-shirts population is highly dynamic, 'flow-wise'. At the group level the four populations show markedly less dynamics. At this level, office supplies is the least dynamic, and T-shirts is the most dynamic population, according to this measure.

Note. We have only applied relative measure 1 of Table 3.8.1 to the data. The choice for method 1 has nothing to do with a personal preference for this method. The choice was restricted to this method only to limit the work required for the present paper. Relative measure 2 is certainly worth considering as well. It even has the nice additivity property (3.3). ■

4.2 m-cohorts: 'births' and 'deaths'

The results can be found in Appendix F. It is immediately clear that at the EAN level there are quite a lot of 'births' and 'deaths'. This indicates that the population at this level is not stable. The picture is quite different at the group level. Although still not stable, the population are notably more stable. Note also that at this level in month 1 there are quite a lot of 'births', probably meaning that a lot of items already existed in that month. Also at the end of the time window (t=50) there is a relatively large number of items. This is likely to mean that most of these items 'die' later. It should be borne in mind that the results are about censored data.

Note. The results presented in Appendix F are all about unweighted items. However, it is worthwhile to consider the age distribution of items if their economic importance

is taken into account, as in case of the age distribution. This can be done as in Appendix D by considering the top n% items in terms of turnover, for various values of n. Or we can take an entire population of items, and give each item a weight (total turnover over in the period considered), and apply formula (3.4) as a measure of the size of each 'birth' or 'death' group (of items born or died in a particular month). Instead of frequencies it gives total weights to each 'birth' or 'death' group. Using various thresholds one then can distinguish important from less important such groups. ■

4.3 M1-cohorts

M1-cohorts: sales status

Appendix G contains plots of the sales status of M1-cohorts. In all cases we see a gently decreasing trend. In case of T-shirts at the EAN level there are some big fluctuations visible, whereas in case of the other three products there is a more steady development.

M1-cohorts: value

The results are contained in Appendix H. The qualitatively most distinguishing development is that of men's T-shirt at the EANs level, with a spectacular drop in turnover after 1 year. At the group level, a similar drop is also visible, but not as pronounced as at the EAN level.

M1-cohorts: items sold

Appendix I shows the results about the total number of items sold for the M1-cohort. The items were unweighted. The results at the EAN and group level turned out to be the same. Note that for men's T-shirts there is also a sharp drop in the items sold visible.

Remarks on weighting

The results about the M1-cohorts as presented in the Appendices H, and I can be seen as weighted variants of the unweighted ones in Appendix G. In case of Appendix H the weight was 'turnover' and in case of Appendix I it is 'items sold', or more accurately, 'copies of items sold'. It is clear that the weighted results show much more variation than the unweighted ones. And the dramatic drop in turnover for T-shirts at the EAN level is absent in the sales status case. However it is interesting if the trends in the unweighted and weighted cases are similar. The peaks in the data can perhaps be interpreted mostly as noise (but including possible seasonal and other) effects. So a time series analysis would be very interesting to see how the trends behave in the unweighted and the two weighted cases, and also, how they relate to each other.

5. Discussion and conclusions

In the report some simple measures have been introduced to quantify the dynamics of a population of items comprising article groups. The ultimate goal of this exercise is to link dynamicity characteristics of such article populations to the choice of an index method. The current paper does not try to reach this goal. It is only a first exploration of some possibilities. The application of the characteristics to real data shows that they are able to pick up differences in the dynamics. In particular, the different dynamics of the men's T-shirt population was picked up by several characteristics.

For price index calculation one would like to know which sudden changes occur in the sales of certain items. Not all index formulas cope well with sudden changes. Perhaps the conclusion can be to leave a certain subgroup of items outside the data to calculate a price index, because their sudden change works disruptively. Leaving out such data might be a better choice than to spoil other data that do not show these sudden changes.

Apart from the goal to link dynamicity characteristics of populations of articles to price indices, it is also a good idea to use them for monitoring purposes in a dashboard application. For analysts it can be helpful to be informed about an article group that shows a different dynamic behaviour from other article populations. Or to get informed of a sudden change of dynamic behaviour of an article population.

Such information is not only useful for scanner data, but also for internet data, namely in the post-collection stage. It may indicate that something went wrong in the collection of internet data, perhaps because a webpage was changed. Of course, because no turnover information is available, the characteristics that can be computed in this case are a bit more limited.

The dynamicity results can also be used to find the right levels of aggregation for articles. The groups of items formed should on the one hand be coarse (enough) to guarantee continuity and on the other fine enough to give sufficient detail. The finer the groups the more likely it is that they are homogeneous enough for price index calculations. Ideally a subdivision of an article group is stable, in the sense that there are no 'births' or 'deaths' in due course, that is, at the level of aggregation considered; new items, at the lowest level, may be 'born' or 'die' as long as they do not influence the existence of groups at a higher level of aggregation. Decay parameters could be used to quantify the stability of a subgroup of items within an article group.

Chessa et al. (2016) concentrated mainly on comparing the results of different kinds of price index formulas. The effect of grouping was considered only crudely. However, this is probably an area that needs to be more thoroughly investigated. Insight into this will give guidance to a suitable partitioning of the article population for the purpose of index calculation. It is likely that a good choice is determined by taking the

dynamicity of a population of articles into account. There is a trade-off to be considered: on the one hand one would like to use groups of articles that are likely to survive for a longer period of time. On the other hand, one would like to use groups that are sufficiently homogeneous. These requirements are contrary. So the challenge is to find the right balance.

Because the observations are censored, there is a need to correct the observations – for instance of age distributions of items for the bias this introduces into the data. Survival analysis is the area to look for useful methods to correct for the 'censor bias'.

If we look at the dynamicity characteristics and the ways of looking at a dynamic population of articles that we consider in the current paper, we can remark that each throws light on some aspect. Stratifying a population into birth and death m-cohorts gives an insight in the renewal rate. Age distributions of items allow to see how many items exist for longer periods of time and how many are short-lived. Studying a particular cohort such as M1 (of items that were sold in month 1) gives a good idea of how the influence of such a group of data is diminished in due course. Weighing the results with turnover helps to get a view on what happens with the items that are really important economically. Flow characteristics give information on month-to-month changes. They allow one to distinguish quiet from hectic article populations. All this information seems to be useful to present in a dashboard, used as a tool for analysts to monitor such populations and to intervene if an anomaly occurs. Or perhaps they can be used to warn if a rather extreme phenomenon occurs that may affect a price index. Both applications require further research, however.

Some thought needs to be spent on how the dynamicity results should be presented, what kind of graphical technique to use, which information should be represented (absolute or relative, etc.)? In the right (graphical) form a few parameters are probably all that is required to characterize the dynamicity of a population.

Looking at the results that take the weight of items into account (see Appendix D) is very illuminating. It is probably very often be the case that a relatively small number of items determine a large portion of the turnover and the vast majority of items is insignificant. The question is, if such insight can be used to make the process of index making more efficient. Why put effort into what is insignificant?[9] This remark is not directly related to the dynamics of an article population per se, but it is interesting to ponder upon.

The flow measures considered in the present paper are based on transitions of states defined by two consecutive months. It is straightforward to generalize this and to

[9] Another possibility is not to discard these data but to combine them into a single, artificial, lumped item It is comparable to the category 'other' for questions. The turnover of this lumped item is the sum of the turnovers of the individual insignificant item. It has average monthly prices and quantities that are mutually consistent and also agree with aggregates calculated from the individual insignificant items. In this way no data are discarded, and the uninteresting details also have been removed. This lumped item has to be treated slightly differently from the proper items when being processed.

look for transitions between states consisting of three (or four) consecutive months.[10]

References

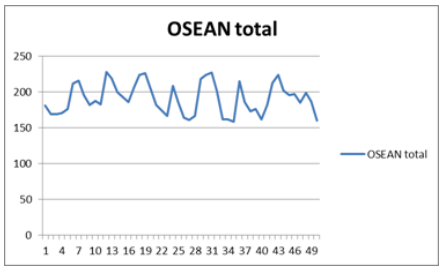
Chessa, A.G., J. de Haan & L. Willenborg (2016). Comparison of price index methods, Report, Statistics Netherlands, The Hague.

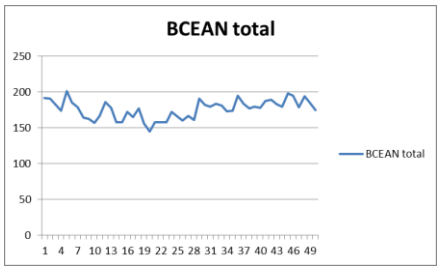
Haan, J. de, Willenborg, L., and Chessa, A.G. (2016). An overview of price index methods for scanner data, Room document for the Meeting of the Group of Experts on Consumer Price Indices, 2–4 May 2016, Geneva, Switzerland.

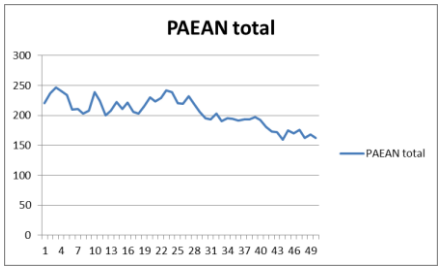
[10] Longer sequences of consecutive months are also possible, of course, but perhaps not very insightful.

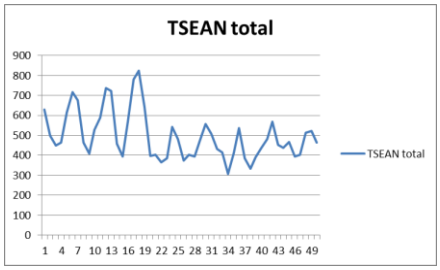
Appendix A. Results entire population: sales status

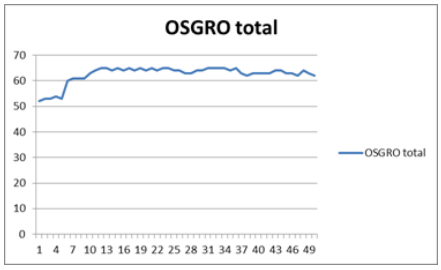
It should be stressed that the scales of the graphs of BCGRO and PAGRO are different from those of the other graphs as they do not start at 0. This may give the false impression at first sight that the results fluctuate more wildly than in the other cases. In all the graphs, the choice of the scales was made by Excel.

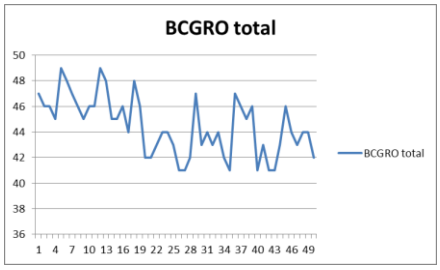


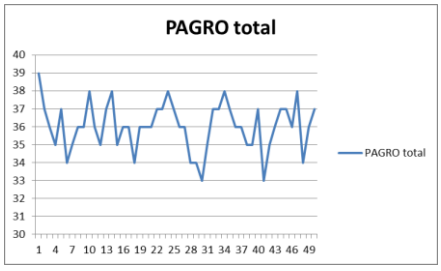


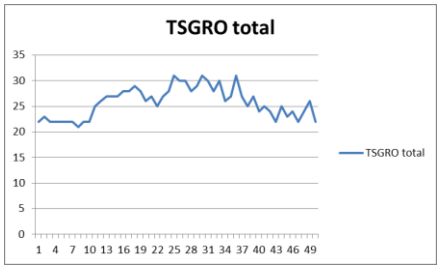






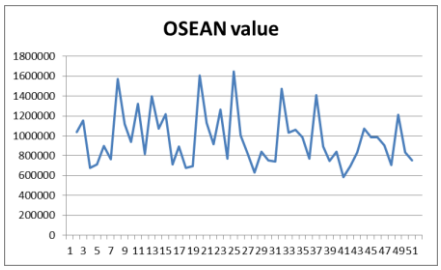


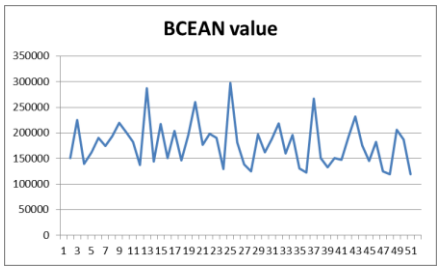


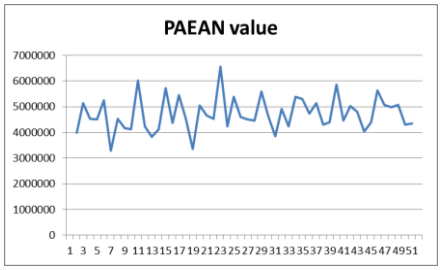


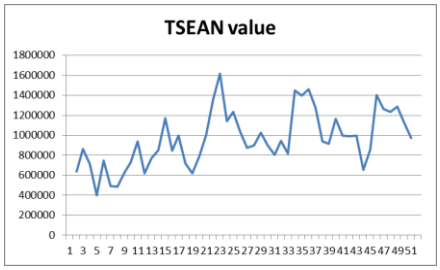
Appendix B. Results entire population: value

The total sales values of the corresponding EANs and GROs are the same, so e.g. OSEAN value and OSGRO are the same. Likewise for the other articles. The pictures below indicate 'EAN', but this is just a reminder of the files that were used to compute the results shown.



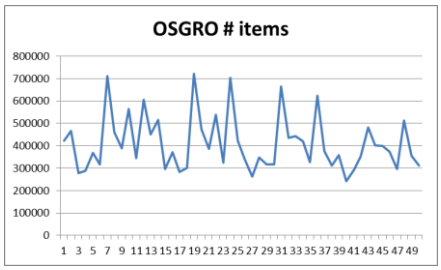


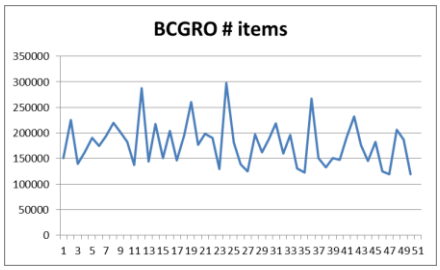


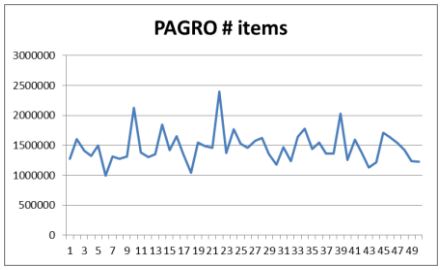


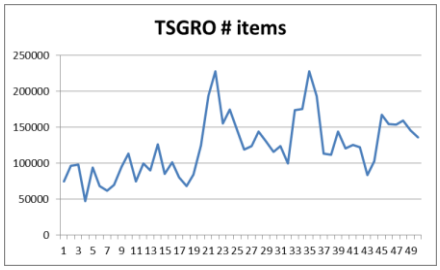
Appendix C. Results entire population: items sold

The total number of items sold is the same for articles, irrespective of the aggregation level (EAN or GRO). So e.g. OSEAN # items and OSGRO # items are the same. Likewise for the other articles. The pictures below indicate 'GRO', but this is just a reminder of the files that were used to compute the results shown.









Appendix D. Results entire population: age

The first two tables contain unweighted age distributions (at EAN and group level). The second pair of tables contains weighted results, where each item received a weight (which was 1 in the unweighted case). The weighting procedure in Section 3.6 is used. The age distributions are based on the observed data, which are biased due to censoring.

The first table in the present appendix contains unweighted age distributions of items at the EAN level for the four article groups.

OSEAN

length	freq
1	15
2	3
3	2
4	2
8	3
9	4
11	1
12	1
13	2
14	1
15	16
16	6
17	3
18	3
19	1
20	4
21	5
22	1
23	2
24	1
25	1
26	6
27	5
28	2
29	4
30	2
31	5
32	4
33	5
35	3
36	4
37	4
38	5
39	5
40	9
41	4
42	9
43	9
44	6
45	25
46	11
47	9
48	14
49	26
50	78

BCEAN

length	freq
1	9
2	1
3	2
4	9
5	2
6	4
7	3
8	1
9	4
10	1
11	12
12	5
13	2
14	2
15	3
16	4
17	2
18	1
19	11
20	12
21	7
22	2
23	6
24	2
25	3
27	5
28	3
29	19
30	12
31	7
32	3
33	1
34	4
35	2
36	4
40	1
44	1
45	4
46	5
47	1
48	8
49	8
50	113

PAEAN

length	freq
1	70
2	11
3	13
4	11
5	8
6	14
7	12
8	13
9	9
10	16
11	21
12	2
13	12
14	9
15	14
16	7
17	6
18	9
19	12
20	3
21	6
22	5
23	8
24	15
25	11
26	14
27	8
28	11
29	10
30	8
31	6
32	10
33	5
34	7
35	3
36	4
37	3
38	7
39	10
40	8
41	9
42	7
43	11
44	4
45	3
47	6
48	5
49	5
50	81

TSEAN

length	freq
1	266
2	47
3	67
4	10
5	14
6	16
7	34
8	25
9	52
10	19
11	93
12	43
13	47
14	33
15	56
16	27
17	37
18	62
19	50
20	29
21	50
22	22
23	26
24	43
25	40
26	39
27	21
28	17
29	7
30	32
31	26
32	21
33	21
34	13
35	10
36	39
37	18
38	64
39	40
40	10
41	25
42	19
43	28
44	18
45	31
46	23
47	41
48	49
49	78
50	55

The second table in the present appendix contains unweighted age distributions of items at the group level for the four article groups.

OSGRO

length	freq
3	1
16	1
36	1
40	3
41	1
42	1
44	1
45	7
48	1
49	2
50	48

BCGRO

length	freq
6	1
15	1
19	2
20	1
25	2
28	1
34	1
36	1
39	1
45	1
47	1
48	1
49	2
50	38

PAGRO

length	freq
6	1
11	1
14	1
36	1
45	1
50	36

TSGRO

length	freq
2	1
12	1
15	3
17	1
24	1
25	1
28	1
37	1
38	3
39	3
42	1
45	1
46	1
48	1
49	3
50	14

In the third and final table in the present appendix we have collected the results on the ages of the top n% of economically important items of men's T-shirts at the EAN level, for several values of n. The case n=100 can be found in the first table in Appendix D.

TSEAN_Top95%

length	freq
2	2
3	2
4	1
7	2
9	8
11	6
13	1
14	1
15	19
16	1
18	2
19	5
20	6
21	7
23	1
25	1
26	8
27	2
28	1
30	8
31	3
32	1
33	1
37	5
38	40
39	15
40	1
41	1
42	6
43	15
44	6
45	4
46	3
47	6
48	11
49	25
50	31

TSEAN_Top90%

length	freq
13	1
15	16
19	1
23	1
26	5
28	1
30	1
37	3
38	39
39	12
40	1
42	2
43	1
44	1
45	4
46	1
47	3
48	4
49	6
50	15

TSEAN_Top85%

length	freq
15	8
26	5
28	1
30	1
37	3
38	39
39	12
40	1
42	2
44	1
45	2
47	2
48	2
49	3
50	8

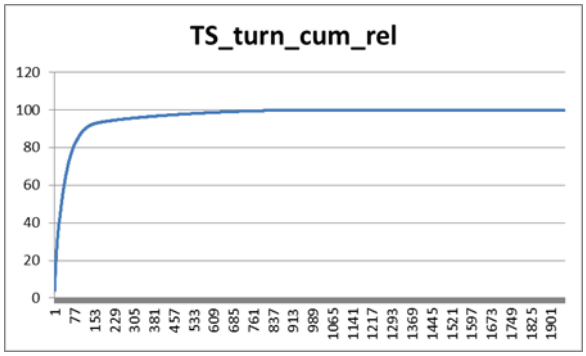
TSEAN_Top80%

length	freq
15	6
26	5
28	1
30	1
37	3
38	36
39	12
44	1
45	1
49	1
50	6

TSEAN_Top75%

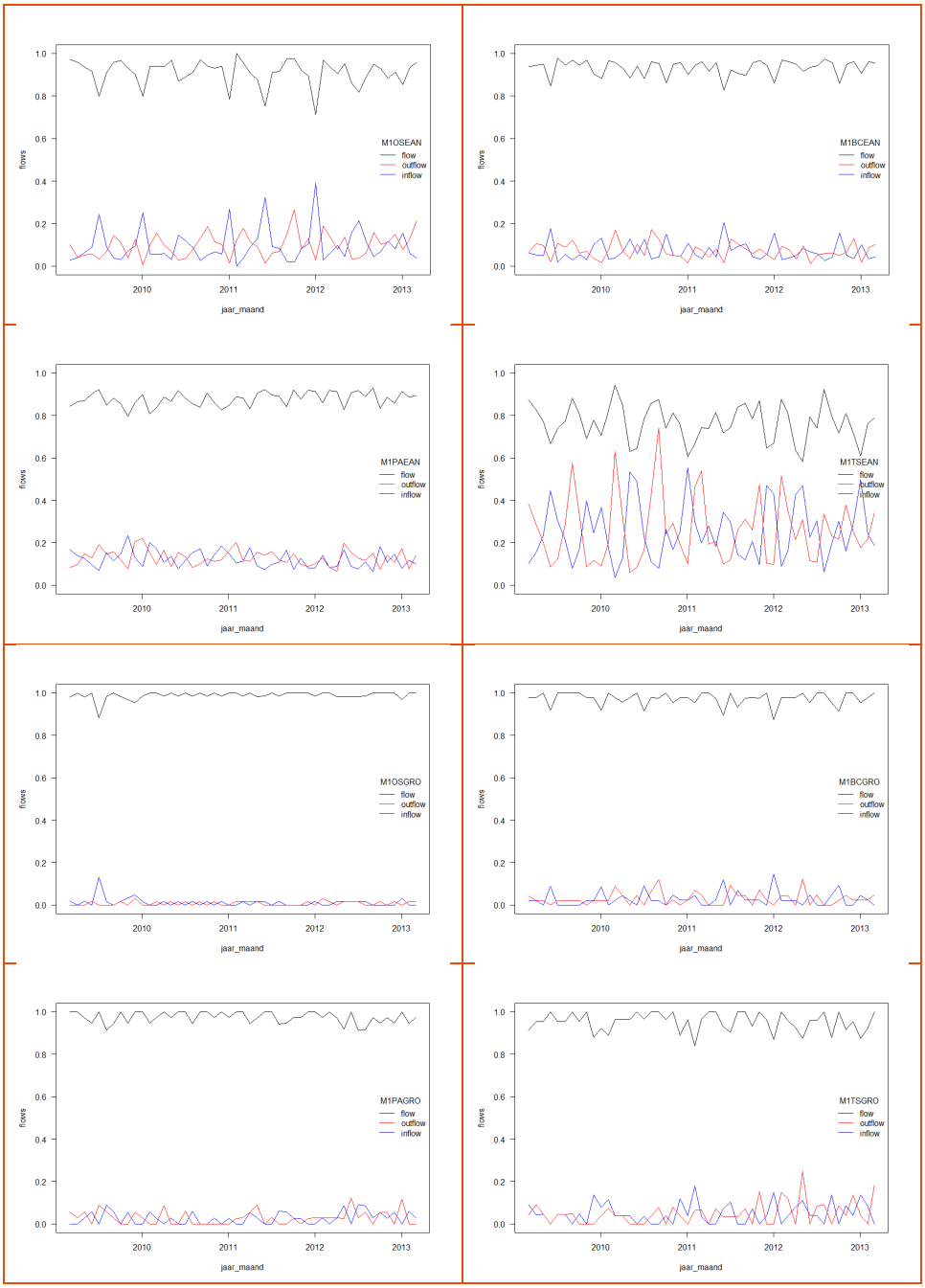
length	freq
15	4
26	4
30	1
37	3
38	31
39	12
44	1
45	1
50	4

The figure below shows the percentage of the turnover of the k largest T-shirt items, for k=1,2,…, 1953.



Appendix E. Results entire population: flows

Relative measure 1 considered in Section 3.8 is used to compute the results in the present Appendix. M1 in the tags refers to relative measure 1 defined in Table 3.8.1.



Appendix F. Results m-cohorts: ‘births’ and ‘deaths’

The results presented here are computed from the observed data. So there is no correction for censoring effects. The items are not weighted.

In the first table in the present appendix the unweighted births distributions of items at the EAN level are presented for the four article groups.

OSEAN

month	freq
1	181
2	5
3	4
4	8
5	6
6	30
7	9
8	3
9	4
10	2
11	9
12	10
13	4
14	2
15	1
17	3
18	3
19	6
21	1
22	2
24	6
28	1
30	3
31	4
34	1
35	4
36	13
37	1
42	3
48	1
50	1

BCEAN

month	freq
1	192
2	12
3	4
4	2
5	4
6	1
7	3
12	2
17	1
18	1
19	1
20	3
21	16
24	5
25	1
27	1
28	4
29	1
30	4
31	11
32	10
33	1
36	9
40	12
42	4
43	1
45	4
46	1

PAEAN

month	freq
1	220
2	37
3	28
4	18
5	7
6	5
7	7
8	12
9	5
10	10
11	9
12	3
13	13
14	6
15	5
16	2
17	4
18	4
19	6
20	11
21	7
22	6
23	15
24	3
25	4
27	4
28	2
30	2
32	13
33	6
34	5
35	4
36	3
37	8
39	3
40	20
41	6
42	2
43	3
44	3
45	8
46	5
47	5
48	1
49	5
50	7

TSEAN

month	freq
1	628
2	64
3	45
4	40
5	96
6	93
7	96
8	11
9	12
10	9
11	24
12	44
13	67
14	8
15	6
16	9
17	39
18	54
19	20
20	6
21	14
22	1
23	1
24	14
25	91
26	55
27	25
28	13
29	31
30	45
31	13
32	8
33	3
36	9
38	3
39	2
40	96
41	2
42	32
43	5
44	16
46	5
47	1
48	52
49	22
50	23

The second table contains unweighted births distributions of items at the group level for the four article groups.

OSGRO

month	freq
1	52
2	1
4	1
6	7
9	1
10	1
11	3
48	1

BCGRO

month	freq
1	47
2	1
5	1
12	1
31	1
32	1
36	1
45	1

PAGRO

month	freq
1	39
40	1
42	1

TSGRO

month	freq
1	22
2	2
4	1
6	1
9	1
12	2
13	3
14	1
23	1
25	3

The third table contains the unweighted deaths distributions of items at the EAN level for the four article groups.

OSEAN

month	freq
1	3
2	1
4	1
5	2
6	3
7	4
8	1
12	3
13	2
14	1
15	2
17	2
18	4
19	5
21	3
22	1
24	1
29	4
30	4
31	8
32	1
35	1
36	4
37	3
39	1
40	1
41	2
42	5
43	11
44	7
45	7
46	11
47	8
48	20
49	34
50	160

BCEAN

month	freq
1	1
2	1
3	2
5	5
6	1
7	3
11	1
12	2
13	3
16	2
17	4
18	2
19	6
20	1
21	2
22	1
24	3
25	2
29	13
30	7
31	6
32	3
33	1
34	3
35	2
36	6
39	6
40	1
41	1
43	2
44	1
45	6
46	7
47	1
48	10
49	18
50	175

PAEAN

month	freq
1	6
2	5
3	5
4	7
5	7
6	2
7	3
8	4
9	4
10	16
11	12
12	2
13	6
14	4
15	5
16	2
17	2
18	4
19	2
20	4
21	5
22	10
23	12
24	21
25	6
26	8
27	15
28	14
29	9
30	13
31	6
32	10
33	8
34	2
35	4
36	7
37	5
38	3
39	20
40	15
41	16
42	10
43	13
44	4
45	6
46	8
47	15
48	10
49	23
50	162

TSEAN

month	freq
1	27
2	4
3	3
4	6
5	9
6	43
7	37
8	7
9	7
10	4
11	17
12	36
13	32
14	2
15	4
16	23
17	59
18	116
19	78
20	15
21	15
22	7
23	13
24	37
25	32
26	8
27	11
28	11
29	20
30	23
31	22
32	22
33	16
34	2
35	13
36	66
37	35
38	33
39	64
40	10
41	20
42	55
43	36
44	23
45	46
46	30
47	46
48	87
49	158
50	463

The fourth and final table in the present appendix contains unweighted deaths distributions of items at the group level for the four article groups.

OSGRO

month	freq
19	1
36	1
44	1
48	1
49	1
50	62

BCGRO

month	freq
19	1
25	1
29	2
36	1
39	1
45	2
47	1
48	1
49	2
50	42

PAGRO

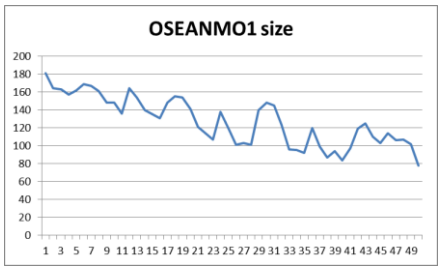
month	freq
14	1
36	1
45	1
47	1
50	37

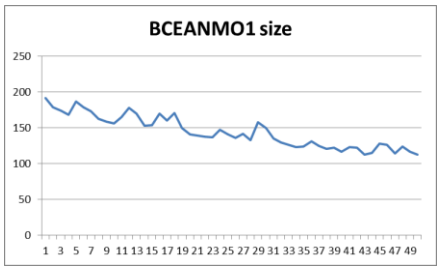
TSGRO

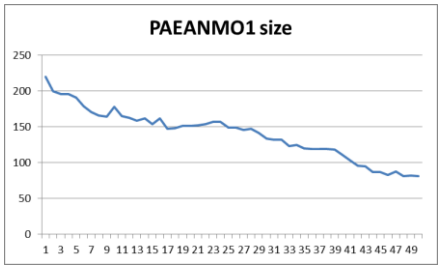
month	freq
5	1
12	1
18	1
24	1
30	1
39	4
45	1
46	1
49	4
50	22

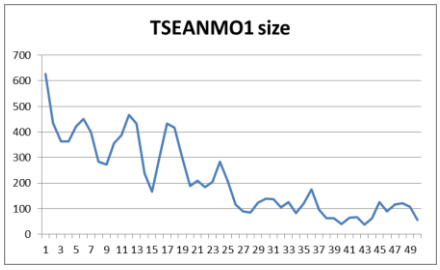
Appendix G. Results M1-cohorts: sales status

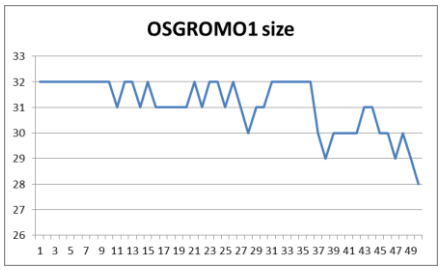
In the figures in the present appendix 'size' means 'sales status' as explained in Section 3.5.

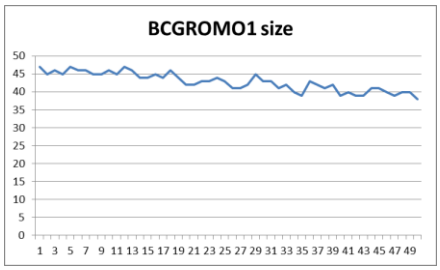


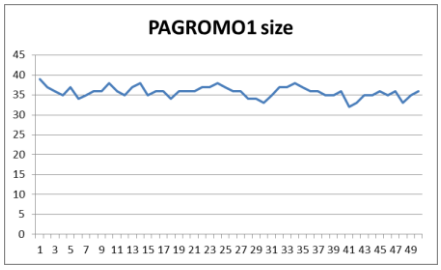


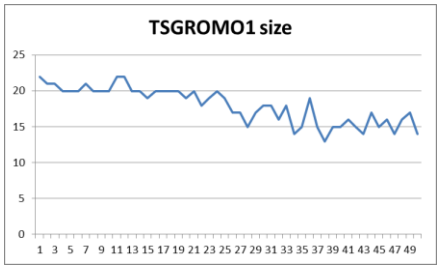






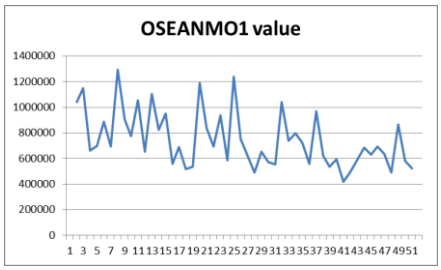


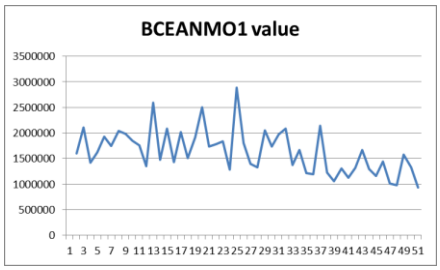


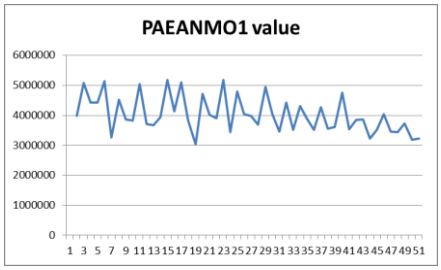


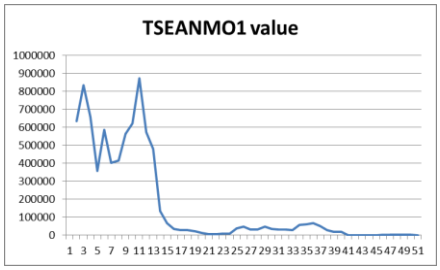
Appendix H. Results M1-cohorts: value

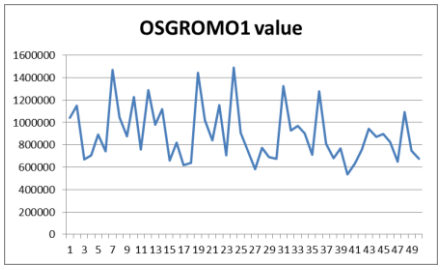
The results of the same article at different levels of aggregation are not necessarily the same in this case. This is due to the fact that a selection is made of groups of items present at t=1, but that the composition of these groups (in terms of EANs present) may be different in different months.

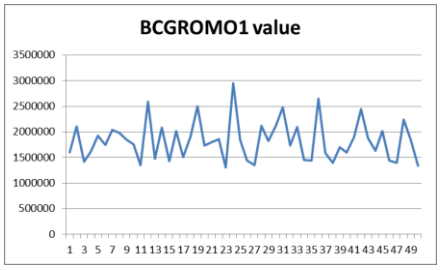


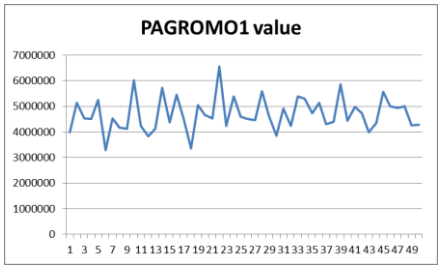


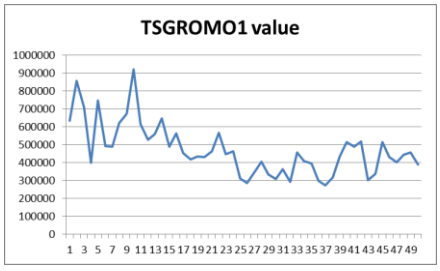






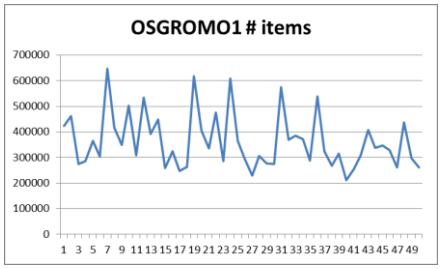


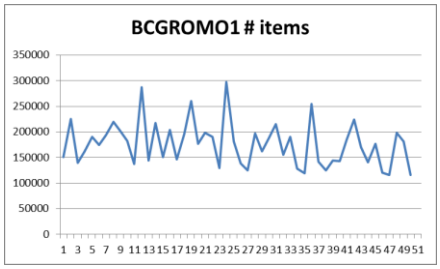


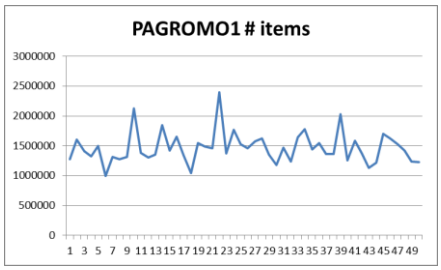


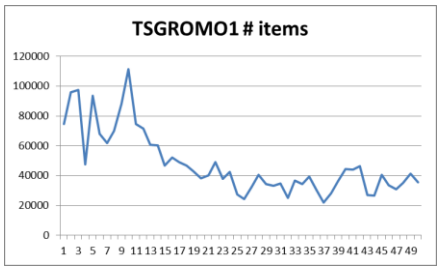
Appendix I. Results M1-cohorts: items sold

It should be noted that the results at EAN and group (GRO) level are the same. That the GRO level results are mentioned in the figures is because those data were used to compute the results.









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