Analysis of Change and Constructing Index Series for Complete Micro Data

Antti Suoperä, Satu Montonen, Kristiina Nieminen

5th May 2017

1 Introduction

In this paper we are going to study different methods or strategies for measurement of price (or quantity) changes and constructing index series in the case of complete micro data. In index number calculation the complete micro data means that we have knowledge of all prices and quantities for all commodities and time periods. In fact, the complete micro data makes possible to use of superlative index number formulas for measurement of price (and quantity) changes. According to the ILO manual (2004), superlative price indices are the best choice and an ideal framework when both detailed price and quantity information are available.

There are two main method in the index number theory for index compilation: the base and the chain method. The base method uses binary comparisons from base period 0 to observation period t and the chain method from $(t-1)$ to t, $t = 1,2,\dots$. Both methods have weak points. In the base method the price change from $(t-1)$ to t depends not only time periods (t - 1, t) prices and quantities but from (0, t - 1, t). It is not a proper index number formula and we call it as secondary one. For example, the base Laspeyres or base Log-Laspeyres belong to this class of index number formulas. The second weakness of the base method is that it excludes new and disappearing commodities from index calculations.

Chained indexes may suffer from what is known as chain drift or chain link bias. Chain drift occurs if a chained index "does not return to unity when prices in the current period return to their levels in the base period" (ILO, 2004, p. 445). When there are large period-to-period fluctuations in prices, quantities and values, all kinds of chained price indices should be avoided, because the chain drift may occur. On the contrary the base method never suffer from the chain drift whatever the index number formula is.

Ivancic, Diewert and Fox (2009) have proposed an approach that provides drift free, 'superlative-type' indexes through adapting multilateral index number theory. Nygaard (2010) and de Haan & van der Grient (2009) have already tested this index number approach to scanner data in Norway and in Netherlands. Nygaard uses the GEKS price index as a benchmark index for measuring the indications of chain drift. The GEKS price index applied to time periods, $T^0, T^1, \dots, T^k$, is 'superlative-type' index number that depends on all prices and quantities of all time periods T^k. The *adapted GEKS* index going from 0 to *k* via link period *l* calculates all possible Törnqvist indices and is unweighted geometric average of them eliminating all time specific, for example monthly, information of the data. Though, empirical studies and the results carried out by Nygaard (2010) and de Haan & van der Grient (2009) are promising.

Our aim in this paper is at first to develop practical methods to analyze existence of the chain drift and then, suggest new methods that effectively eliminates chain drift phenomenon and are reliable in construction of index series. We use Törnqvist formula in our analysis by four different ways. First, we apply *base* Törnqvist say, for commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities. In this method the chain drift never exists. Second, we select the same set of commodities as in base method and apply *chain* Törnqvist for them. This method is called *chain in isolation.* Third, we select *proper chain* Törnqvist for the data including maximum number of matched pairs. Forth, we define *mixed* method, where we combine together the base method (first method) and typical price change calculation for new and disappearing commodities by Törnqvist formula.

We visualize our index number theory by empirical analysis. The data used, is data of pharmaceutical product (i.e. over-the-counter medicines) uphold by Pharmaceutical Information Centre Ltd, Finland. It is *complete micro data* in the sense that it contains all prices and quantities for a large set commodities and periods, that are months in the paper. Instead of couple of dozens product varieties typically included into a CPI sub-index, the data we use contain complete price and quantity data of all periods (here all months during 2009-2015) for nearly 600 over-the-counter medicines. These are medicines that have a constant quality over time. This means that problems of quality change do not appear in this data set.

2 Basic concepts and notation for index numbers

Our notation for the index number calculations is the following:

Commodities:	$a_1, a_2, \dots, a_n$ are here over-the-counter medicines.
Time periods:	t = 0, 1, 2, … are the compared situations (only two in binary comparisons).
Prices:	p_i^t is the unit price of a_i in period *t*.
Quantities:	q_i^t is the quantity of a_i in period *t*.
Values:	$v_i^t = p_i^t q_i^t$ is the value of a_i in period *t*.
Total value:	$V^t = \sum v_i^t$ is the total value of all the commodities.
Total value ratio:	$V^{t/0} = V^t / V^0$ is the total value ratio from period 0 to *t*.
Price relatives:	$p_i^{t/0} = p_i^t / p_i^0$ is the price relative of a_i from period 0 to *t*.
Quantity relatives:	$q_i^{t/0} = q_i^t / q_i^0$ is the quantity relative of a_i from period 0 to *t*.
Value relatives:	$v_i^{t/0} = v_i^t / v_i^0$ is the value relative of a_i from period 0 to *t*.
Value shares:	$w_i^t = v_i^t / \sum_i v_i^t$ is the value share of a_i in period *t*.

We consider binary comparisons or time periods (or situations) T^0, T^1, ..., where $T^t = T(p^t, q^t)$ consists of its price and quantity vectors using *proper* index number formulas. These index number formulas satisfy the following minimum requirements:

Commodity reversal test (CRT):	The order of commodities has no effect on the index number.
Unit of measurement test (UMT):	The units of measurement of commodities have no effects on the index number.
Money unit test (MUT):	The money unit has no effect on the index number.
Weak proportionality test (Vartia (1976):	The price index equals *k*, whenever $p^1 = kp^0$ and $q^1 = mq^0$, where *k, m* are positive numbers

These four minimum requirements are represented accurately in Vartia (1976) and discussed in Vartia (2010).

In this paper we consider the base, the chain and the mixed methods (or strategies) in constructing the index series. The base method is based on binary comparisons between T^0 and T^t (for $t = 1,2, \dots$) whereas the chain method is based on binary comparisons between T^{t-1} and T^t (for , $t = 1,2, \dots$). In general, we say that all these binary comparisons use *proper* index number formulas, whatever the formula happens to be. The index numbers based on these binary comparisons are referred as *measured* changes. All other changes included in the index series are called *derived* changes – they are not actual binary comparisons between the situations or measured changes, because they depend on prices and quantities from at least three periods. The derived changes do not correspond to *proper* index number formulas, but are called here as *secondary* index number formulas.

For instance, the change $T^1 \rightarrow T^2$ is a *measured* change for the chain method, but a *derived* change for the base method (i.e. base period 0). The base method is not based on a binary comparison and on a *proper* index number formula, but a *secondary* index number formula. In case of base method, only the comparisons $T^0 \rightarrow T^t, t = 1,2, \ldots$ are measured changes, while all the other changes are *derived changes* based on *secondary* index numbers. Similarly, in the case of chain method, only the comparisons $T^{t-1} \rightarrow T^t, t = 1,2, \ldots$ are measured changes, while all other changes are *derived* changes based on *secondary* index numbers.

We have noticed that the estimate of price change (measured or derived) is conditional on used method (e.g. base, chain) and index number formula (e.g. Laspeyres, Log-Laspeyres, Paasche, Log-Paasche,…). The following table shows example of that for commodity set $\{a_1, a_2, \ldots, a_n\}$ and Log-Laspeyres formula *l*.

Table 1: Example of measured and derived change for base and chain methods in the case of Log-Laspeyres formula for commodity set $\{a_1, a_2, \ldots, a_n\}$.

Method	Index number formula	Measured change	Derived change
Base	$l_{Base}^{t/0} = \prod (\frac{p_i^t}{p_i^0})^{w_i^0}$	$l_{Base}^{t/0} = \prod (\frac{p_i^t}{p_i^0})^{w_i^0}$	$l_{Base}^{t/(t-1)} = \prod (\frac{p_i^t}{p_i^{t-1}})^{w_i^0},\ t = 0, t-1, t$
Chain	$l_{Chain}^{t/(t-1)} = \prod (\frac{p_i^t}{p_i^{t-1}})^{w_i^{t-1}}$	$l_{Chain}^{t/(t-1)} = \prod (\frac{p_i^t}{p_i^{t-1}})^{w_i^{t-1}}$	$l_{Chain}^{t/0} = exp\left\{\sum_{i=1}^{n} w_i^0 \log\left(\frac{p_i^1}{p_i^0}\right) + \sum_{i=1}^{n} w_i^1 \log\left(\frac{p_i^2}{p_i^1}\right) + \cdots + \sum_{i=1}^{n} w_i^{t-1} \log\left(\frac{p_i^t}{p_i^{t-1}}\right)\right\}$

In the base and the chain methods, measured change depends on prices and quantities of two time periods. They are *proper* index number formulas based on binary comparison. The change $T^{t-1} \rightarrow T^t$ for the base method depends on prices and quantities form three periods and the change $T^0 \rightarrow T^t$ for the chain method depends on prices and quantities from all 0, 1,…,*t* time periods. That's why these formulas are not proper index number formulas and we call them as secondary ones.

In the base index number formulas no chain-drift phenomena takes place. Let's look at the chain method by trivial example. We analyze the price changes $p_i^0 \rightarrow p_i^1 \rightarrow p_i^2 = p_i^0$, for all *i*. Prices first decline (or increase) and then return back to their original level. We can analyze this phenomena by writing the chain Törnqvist in a logarithmic form

$$\sum \tfrac{1}{2}(w_i^0 + w_i^1)\log\left(p_i^1/p_i^0\right) + \sum \tfrac{1}{2}(w_i^1 + w_i^2)\log\left(p_i^2/p_i^1\right) = 0$$

and after simple derivation we get

$$\sum \tfrac{1}{2}(w_i^0 - w_i^2)\log\left(p_i^1/p_i^0\right) = 0.$$

In this trivial example the equation holds if all quantities are exactly the same or have changed proportionally for all $i = 1,\ldots, n$ and $T^0 \rightarrow T^2$ or by accident. In general we may say that when prices and quantities have changed proportionally between time periods 0 and 2, the superlative chain and base methods capture the same change, otherwise by accident. If the equation do not hold, 'reproaching finger' points to the chain method and existence of the chain drift. This is trivial theoretical method to reveal the existence of chain drift, but it has no use in practice. We have to develop practical (scientific) methods to reveal it. For that, we select Törnqvist index number formula.

We define the base Törnqvist index and its *measured* price change for commodity set $\{a_1, a_2, \dots, a_n\}$ by

(1) $$t_{Base}^{t/0} = exp\{\sum \tfrac{1}{2}(w_i^0 + w_i^t)\log(p_i^t/\bar{p}_i^0)\},$$

where the base period 0 is previous year and *t* is observation month in current year. The base Törnqvist index is *proper* superlative index number formula satisfying minimum requirements applied for index number formula. In (1) we compare arithmetic mean prices $\bar{p}_i^0$ (i.e. quantity weighted) and current month prices p_i^t and the index weights are arithmetic mean of base and observation periods value shares w^0 and w^1. We change the base year every January. Index series with a different base year are chained by transforming each single base year index series to begin from December in base year (i.e. $t_{Base,Dec} = 1$). In base method, like in this example, chain drift never exists.

In the second method we calculate chained Törnqvist for the same set of commodities than in the first method. The Törnqvist index for set $\{a_1, a_2, \dots, a_n\}$ for time periods *t*-1 and *t* is

(2) $$t_{Chain}^{t/(t-1)} = exp\{\sum \tfrac{1}{2}(w_i^{t-1} + w_i^t)\log(p_i^t/p_i^{t-1})\}$$

These *measured* price changes can be easily chained to get index series based on chained method.

In the third method, the *proper chain* Törnqvist is calculated including maximum number of matched pairs in base and observation periods. The price changes are calculated by (2) and they are chained together to get the index series for self-care and recipe drug commodities. In this method we include new and disappearing commodities in the most dynamic way. The $t_{Proper\ Chain}$ includes set $\{a_1, a_2, \dots, a_n\}$ and new and disappearing commodities.

For mixed method, we need to derive the price change for the base Törnqvist for all $i, t = 1,2, \dots$ For simplicity we analyze only base (i.e 0) and two other periods, say *t* =1, 2 and the *derived* price change $logt_{Base}^{2/1} = logt_{Base}^{2/0} - logt_{Base}^{1/0}$ in the logarithmic form is

$$logt_{Base}^{2/0} - logt_{Base}^{1/0} = \sum \tfrac{1}{2}(w_i^0 + w_i^2)\log(p_i^2/\bar{p}_i^0) - \sum \tfrac{1}{2}(w_i^0 + w_i^1)\log(p_i^1/\bar{p}_i^0)$$

(3a) $$= \tfrac{1}{2}\{logl^{2/1} + \sum(w_i^0 - w_i^1)\log(p_i^2/p_i^1) +$$

(3*b*) $$logp^{1/2} + \sum(w_i^2 - w_i^1)\log(p_i^1/\bar{p}_i^0)\},$$

where $l^{2/1}$ is measured price change for Log-Laspeyres and $p^{1/2}$ is measured price change for Log-Paasche for the change $T^1 \rightarrow T^2$. The first term, i.e. (3a), is more often expressed by $\sum w_i^0 \log(p_i^2/p_i^1)$ and it is widely used *secondary* index number formula in CPI calculations all over the world. In general, change $T^1 \rightarrow T^2$ for base Törnqvist (i.e. $exp\left(logt_B^{2/1}\right)$) depends on prices and quantities from three time periods. So it is not based on binary comparisons and is not a *proper* index number formula. In fact, (3a) and (3b) satisfy four minimum requirements (CRT, UMT, MUT and WPT) and so the *derived* price change $t_{Base}^{2/1}$ is an index number formula, but belongs to class of *secondary* ones.

In the mixed method, we combine the *derived* price change $t_{Base}^{2/1}$ and the *measured* price change $t_{Chain,N\&D}^{2/1}$ estimated by (2), that is

(4) $$t_{Mixed}^{2/1} = exp\left\{\tfrac{1}{2}(w_{Base}^1 + w_{Base}^2)logt_{Base}^{2/1} + \tfrac{1}{2}(w_{N\&D}^1 + w_{N\&D}^2)logt_{Chain,N\&D}^{2/1}\right\}.$$

w_{Base}^t and $w_{N\&D}^t$ are value shares for commodity set $\{a_1, a_2, \dots, a_n\}$ and for new and disappearing commodities respectively. These value shares sum up to unity i.e. $\tfrac{1}{2}(w_{Base}^1 + w_{Base}^2) + \tfrac{1}{2}(w_{N\&D}^1 + w_{N\&D}^2) = 1$.

The mixed Törnqvist is *derived* price change and does not corresponds to a *proper* index number formula, but is *secondary* one satisfying the minimum requirements CRT, UMT, MUT and WPT. The index series for mixed Törnqvist is easily derived by chaining the price changes (4) properly.

3 Empirical results

Prices of prescription and over-the-counter medicines are equal in all pharmacist thus no price aggregation is needed. Hence the compensation of the health insurance has effect on the actual prices, paid by the consumer for reimbursed prescription medicines in Finland, we will only use the data on over-the-counter medicines in following analysis.

The existence of chain drift is linked to changes of weights used in index number formula. Because we have no theoretical methods that may be used universally to reveal the existence of chain drift, we have to develop practical (scientific) methods to reveal it. Nygaard (2010) and de Haan & van der Grient (2009) both use the so called GEKS 'index number formula' to do that.

We suggest four different methods or strategies to reveal the existence of chain drift. We also provide a comparative discussion on their strengths and weaknesses. Let us compare these methods in the following order: First we compare base to chain Törnqvist in isolation for the same set of commodities. Differences between base and chain index series for the same set of commodities reveal the chain drift. After that we compare the *chain Törnqvist in isolation* (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$) to *the proper chain* Törnqvist allowing maximum number of matched price pairs in proper chain analysis. This will give information about price changes between commodity set $\{a_1, a_2, \dots, a_n\}$ and maximum number of matched price pairs. In the case of these two chained methods having different price developments, these differences will be accounted to new and disappearing commodities. If base and both chain methods differ from each other, we need a *mixed* method or strategy to solve weak points of other three methods. Especially when the value shares of new and disappearing commodities are stable in time, the mixed method is superior compared to the other methods.

As we can see in figure 1, for any base year the value shares (Törnqvist) are in the beginning of year about 2 – 4 percent and they increases during each year about 2 – 9 percent. The year 2011 is unusual, but in general the value shares of the new and disappearing commodities changes quite moderately compared to for example Nygaard (2010) and de Haan & van der Grient (2009) cases. During the years 2014 and 2015 the value shares are about 6 percent almost all the time. Increase of the value shares are mainly explained by increase of the number of new commodities and not by increase of the values (i.e. expenditures) of a single commodity. In de Haan & van der Grient (2009) –study problem arise from very dramatic changes in the quantities (i.e. more than 10 times increases and declines), the expenditures and the value shares of a single commodity.

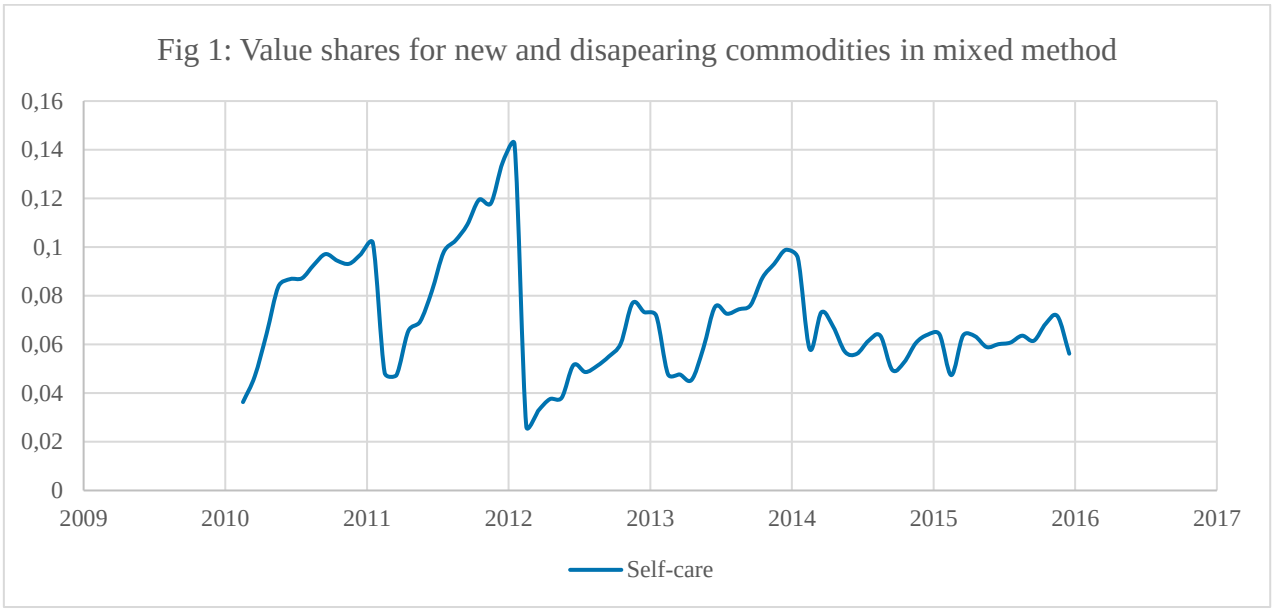
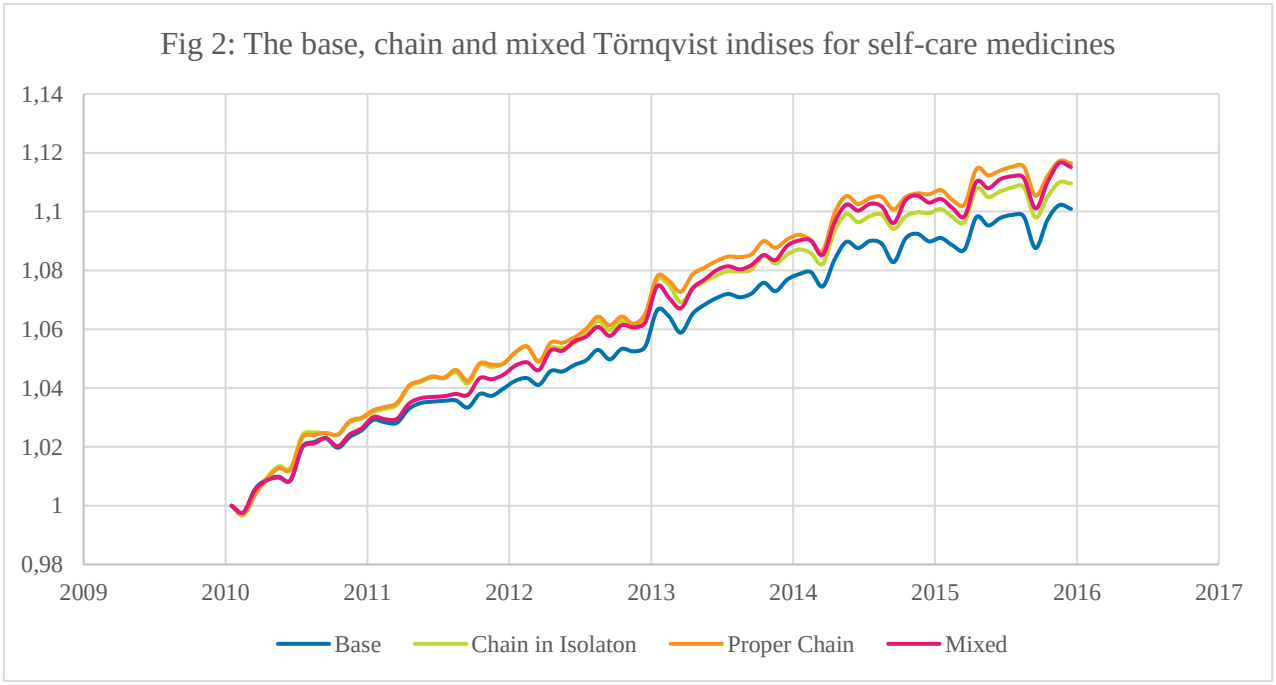


Figure 2 represent index series for base (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$), chain (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$), proper chain (i.e. maximum number of matched pairs) and mixed methods.



The difference between the base and the chain in isolation may be seen in mid of the years 2010 and 2011. Otherwise the price changes are almost the same. It's true – a slight chain drift happens. It corresponds about one percent increase of the prices during six years compared to the base method. Comparing the chain in isolation with the proper chain series, we notice that the new and the disappearing commodities increase the price changes about one percent during six years. There exists 72 periods where the price changes are almost equal in four methods.

The Figure 2 tells also that the base Törnqvist is downward biased. That happens because it excludes the new and the dissapearing commondities. We suggest using a *superlative mixed method*, because the base Törnqvist under estimates the price changes and both chain methods may include a slight chain drift.

4 Conclusion

Vartia and Suoperä (2017) have proved that the index number formulas based on only old or new weights are contingently biased compared to the superlative indices. So, for example '*Laspeyres-type*' indices, whatever they are, should be avoided in the case of complete micro data. The only selected index formula should be *superlative* index number formula. The superlative index number formulas are generally interrelate with the *chain* method. For example, according to the ILO manual (2004), superlative price indices are the best choice, when large period-to-period fluctuations in prices and quantities do not take place, other while they should be avoided. This is of course included in the *proper* chained superlative index number formulas and the possible existence of chain drift. In general, the superlative index number formulas should be preferred. For example, use of the *base* Törnqvist, the chain drift is impossible. We should always prefer superlative indices over '*Laspeyres-type*' indices and pay attention to different strategies in the constructing of index series by some superlative index number formula.

The two main strategies in construction of index series are the *base* and *chain* methods. Both methods have weak points: the *base* method excludes the new and the disappearing commodities and the *chain* method suffers from possible chain drift. In this paper we suggest the third strategy in construction of index series that eliminates disadvantages of the pure base and the chain methods. We call it as *mixed* method. The mixed method consist of the base method, applied to commodity set $\{a_1, a_2, \dots, a_n\}$ that will be selected every January and the chain method applied to new and the disappearing commodities. These two methods, or more precisely their price changes, are weighted together by the value shares and then used in the construction of index series. In our data, the value shares of the base method varies from 86 to 98 percent and it clearly dominates in the mixed method. We use the base and the chain Törnqvist as an example.

Some question related to the price changes surely arises especially from the base method. The price changes $T^0 \rightarrow T^t, t = 1,2,$ by Törnqvist formula are surely *measured* changes and are *proper* superlative index number formulas. The price changes $T^{t-1} \rightarrow T^t, t = 1,2,$ are *derived* changes that depends on prices and quantities from three periods (i.e. T^0, T^{t-1}, T^t).

Following Vartia (1976, 2010) and Pursiainen (2005), *derived* changes should satisfy some minimum requirements to be an index number. The *derived* changes for base Törnqvist satisfy these minimum requirements - Commodity reversal test, Unit of measurement test, Money unit test and Weak proportionality test – being, as we call, *secondary* index number. All the base index number formulas and their derived price changes, like base Laspeyres or Log-Laspeyreas, are *secondary* index numbers – using these two index formulas is a common practice, in particular when it comes to producing the official statistics.

We require that our mixed method satisfy at least minimum requirements and the rest are left to data in question. In our empirical analysis we have compared *base*, *chain in isolation* (same set of commodities as in base), *proper chain* (maximum number of matched pairs) and *mixed* methods for Törnqvist formula. We notice, that base method under estimates price development and the chain methods include some chain drift. The best method for pharmaceutical products data seem to be the mixed method.

References

ILO, International Labour Office (2004): "Consumer Price Index Manual: Theory and Practice", expanded version of "Consumer Price Indices: An ILO manual" (1989).

Ivancic, L., Fox, K.J. and Diewert, W.E. (2009): "Scanner Data, Time Aggregation and the Construction of Price Indexes", article presented at the Ottawa meeting, Neuchâtel, Switzerland, May 2009.

Haan, J. de and Grient, H. van der (2009): "Eliminating Chain Drift in Price Indexes Based on Scanner Data", article presented at the Ottawa meeting, Neuchâtel, Switzerland, May 2009

Nygaard, Ragnhild (2010): Chain Drift in a Monthly Chained Superlative Price Index, Statistics Norway.

Pursiainen, Heikki (2005): **Consistent aggregation methods and index number theory,** Research Reports, Kansantaloustieteen laitoksen tutkimuksia (Research of the Department of Economics in Helsinki) No. 106:2005. Dissertationes Oeconomicae

Vartia, Yrjö (1976): **Relative changes and index numbers,** Research Institute of the Finnish Economy ETLA, series A4 (dissertation in statistics).

Vartia, Yrjö (2010): Principles of defining index numbers and constructing index series, *HECER Discussion Papers No. 283, January 2010.*

Vartia, Yrjö and Suoperä, Antti (2017): Index number theory and construction of CPI for complete micro data, unpublished manuscript.