Statistics Finland

Small scale “big data” in the Finnish pharmaceutical product index compilation

Ottawa Group –conference / Eltville, Germany
Kristiina Nieminen
10th May 2017

Content

1. Background and introduction of the data
2. The practices
    1. Define the index compilation strategy
    2. Standardise data collection with metadata
3. The test calculations and the results
    1. Results from current calculation
    2. Index formula tests by Vartia & Suoperä
    3. The chain-drift –test
4. Conclusions

1. Background

- First attempt to utilise the transaction data in year 2000
  - Daily products from selected commodity groups
- Eurostat's venture on "Modernisation of price collection and compilation"
  - Recommendations for obtaining and processing the scanner data
  - Facilitates the EU-members in the introduction of scanner-data
- New project in 2014-2016
  - Re-design of data collection >> scanner-data and web-scraping
  - Re-design of the index compilation
- Results of the project
  - Pharmaceutical products data implemented into production in the beginning of year 2017
  - Test calculations with superlative index formulas

1. Introduction of the data

- Source: Pharmaceutical Information Centre
- Pharmaceutical products for eCOICOP-groups >>
- Medicine prices are regulated
  - No discounts
- All products are identified with VNR-code
  - No relaunches
- Monthly delivery of prices, quantities and descriptive information by product
  - 10 000 individual product in a month, 32 variables
- Aim is to utilise as much of the data as possible

06	HEALTH
06.1	Medical products, appliances and equipment
06.1.1	Pharmaceutical products
06.1.1.0	Pharmaceutical products
06.1.1.0.1	Prescription medicines
06.1.1.0.1.1	Refundable prescription medicines
06.1.1.0.1.2	Non-refundable prescription medicines
06.1.1.0.2	Over-the-counter medicines
06.1.1.0.2.1	Over-the-counter medicines
06.1.1.0.3	Nicotine replacement therapy preparations
06.1.1.0.3.1	Nicotine gum
06.1.1.0.4	Vitamins
06.1.1.0.4.1	Multivitamins
06.1.1.0.5	Oral contraceptives
06.1.1.0.5.1	Oral contraceptives

2.1 Practices: The definition of compilation strategy

The purpose for using the index :

- 1. the characterisation of the commodities >>*described in slide 4*
- 2. the reference group of economic actors >> *consumers*
- 3. the length of the time periods >> *one month*

The technical problems of index calculation :

- 4. the classification applied to the commodities >> *COICOP*
- 5. the collection method >> *complete microdata collected*
- 6. the appropriate weight structure >> *relative value shares of the previous year by commodity*

The index calculation methods should be decided by specifying:

- 7. the index formula >> *Log-Laspeyres (elementary aggregates)*
- 8. the strategy for constructing the index series >> *Chain method where relative price changes of consecutive months are calculated for each VNR-commodity. These changes are aggregated together with value share weights. Price comparison is made for those commodities that belong to the two year panel data*

The special challenges

- 9. Quality changes in commodities >> *no quality change*
- 10. New and disappearing commodities >> *price for disappearing commodities is estimated by calculating the average change by strata*
  >> new commodities are introduced in the next update of panel data

2.2 Practices: The utilisation of metadata in data collection

```
VNR;Date;status;PriceNoTax;PriceTax;PriceWholeSale;SubstitutionGroup;SubstitutionCode;ReferencePrice;PriceUpperLimit;
421180;2017-02-01;5;8,61;9,47;5,94;;;;;207;1;AEK, PK;1
137340;2017-02-01;5;2,25;2,48;1,55;0849;0008490100;3,48;3,48;110;1;PK, YEK;1
521789;2017-02-01;5;8,61;9,47;5,94;1082;0010820100;8,32;8,32;110;1;PK, YEK;1
558709;2017-02-01;5;17,31;19,04;12,14;1069;0010690001;19,04;19,04;;1;PK;1
421495;2017-02-01;5;3,81;4,19;2,63;0322;0003220020;4,69;4,69;115, 116, 117, 128, 130;1;PK, YEK;1
520647;2017-02-01;5;23,44;25,78;16,68;1069;0010690003;25,79;25,79;;1;PK;1
421636;2017-02-01;4;120,65;132,72;92,09;0224;0002240100;;;;0;;1
173653;2017-02-01;5;567,95;624,75;483,00;;;;;;0;EK;1
```

Take original data and complement it with metadata. Utilise this information in design of data processing.

Hinnat ja kustannukset /Kuluttajahintaindeksi

Tiedoston nimi — Dataset name

/TKSAS/SASDATA/Tilastot/khi/Import//DWFIN_Prices.csv

Dataset format — Tiedoston formaatti: sequential

Delimiter — Erotinmerkki: ;

Tiedostokommentti

Muuttujia: 14 — Variable quantity

Havaintoja — Observation quantity

Technical name / Tekninen nimi	Label / Muuttujan nimi	Group / Muuttujaryhmät	Data type / Tietotyyppi	Format / Esitysmu	Min / Minimiarv	Max / Maksimia	Values / Arvot-list	Length / Pituus	Alkupo	Puuttuva tieto (sallittu)	Puuttuv tietojen lkm
VNR	Product ID-number	Register;Prices;	character				"	6		no	
Date	Date	Register;Prices;Quantities;;	dateandtime	yymmdd10.			"	10		no	
Status	Status	Prices;	numeric					8		no	
PriceNoTax	Price without VAT	Prices;	numeric				"	8		no	
PriceTax	Price with VAT	Prices;	numeric				"	8		no	
PriceWholeSale	Wholesaleprice	Prices;	numeric				"	8		no	
SubstitutionGroup	Substitution Group	Prices;	numeric				"	8		no	

Statistics Finland

Pre-analysis report

Source Data: /TKSAS/SASDATA/Tilastot/khi/Import//DWFIN_Prices.csv
Pre-analysis report based on the data description:

Observation count
10 106

Key figures for numerical variables

Obs	variable	variablename in Finnish	obs	missing	mean
1	date	Tietueen päivämäärä	10 106	0	20 910.00
2	pricenotax	Vähittäismyyntihinta, veroton	9 998	108	237.03
3	…		9 998	108	260.74
10	substitutiongroup	Substituutioryhmä	5 582	4 524	968.79

Character variable frequencies

Obs	variable	variablename in Finnish	obs	missing
1	compensation	Tieto korvattavuudesta	10 106	0
2	reimbursementcodes	Kela-korvattavien läkkeiden korvausnumerot koodeina	9 788	318
3	reimbursementnumber	Kela-korvattavien läkkeiden korvausnumerot	3 513	6 593
4	vnr	Tuotteen yksilöintitunnus	10 106	0

Check of classification values

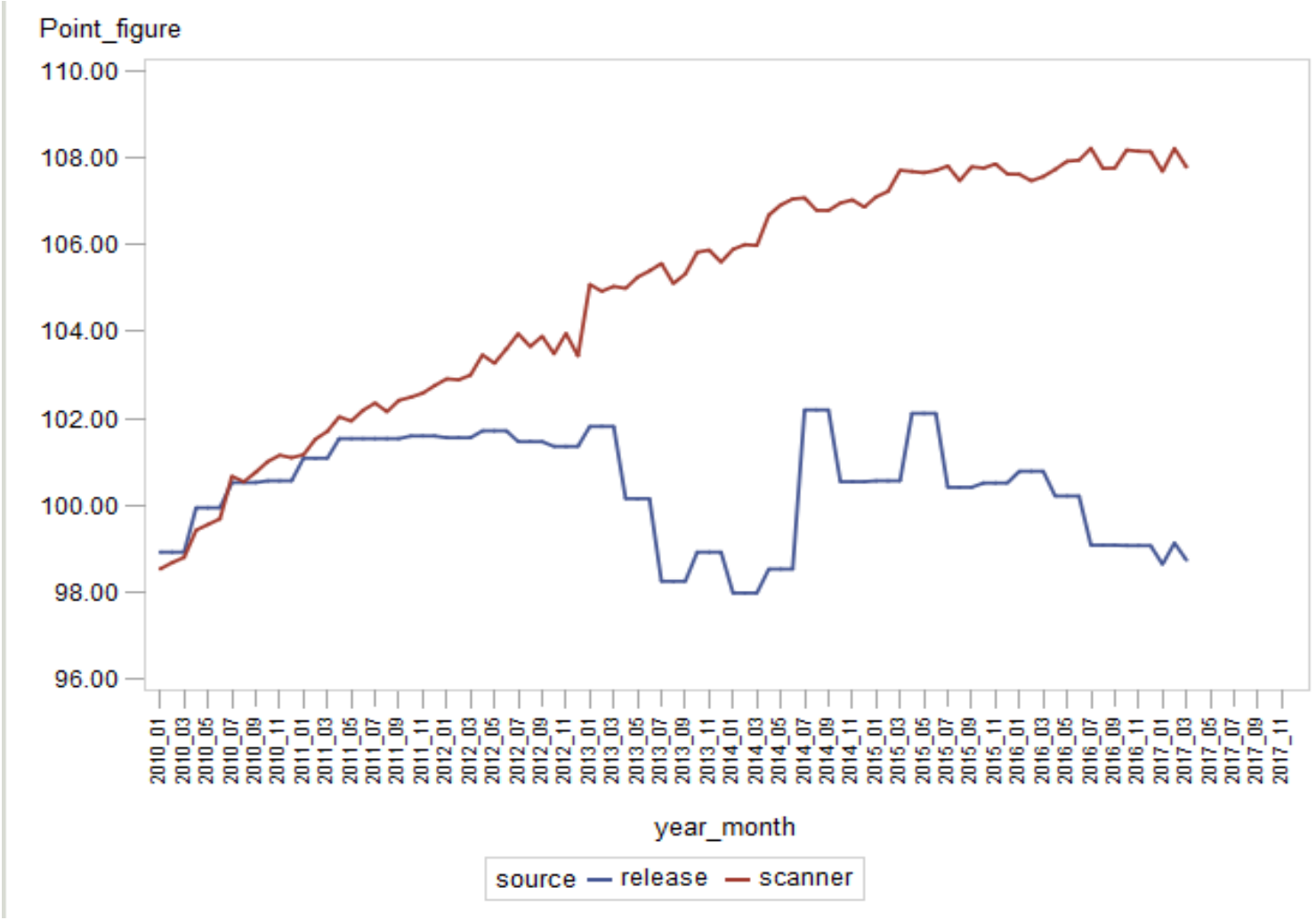
	Compensation code			
reimbursementcodes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AEK. LRPK	38	0.39	38	0.39
AEK. PK	1372	14.helmi	1410	14.41
AEK. PK. YEK	86	0.88	1496	15.28
EK	4805	49.09	6301	64.37

3.1 Results from current calculation

Compilation of elementary indices

- According to the strategy definition (slide 5)
  - Two year panel
  - Paired comparison of the prices of base and comparison periods
  - relative change in prices is estimated for each commodity
  - Laspeyres used in aggregation
- Results:
  - over-the-counter medicine prices have grown by almost 12.5 per cent between 2009/1 and 2016/12
  - comparison between new index series and the published index series tells another story

3.1 Results from current calculation

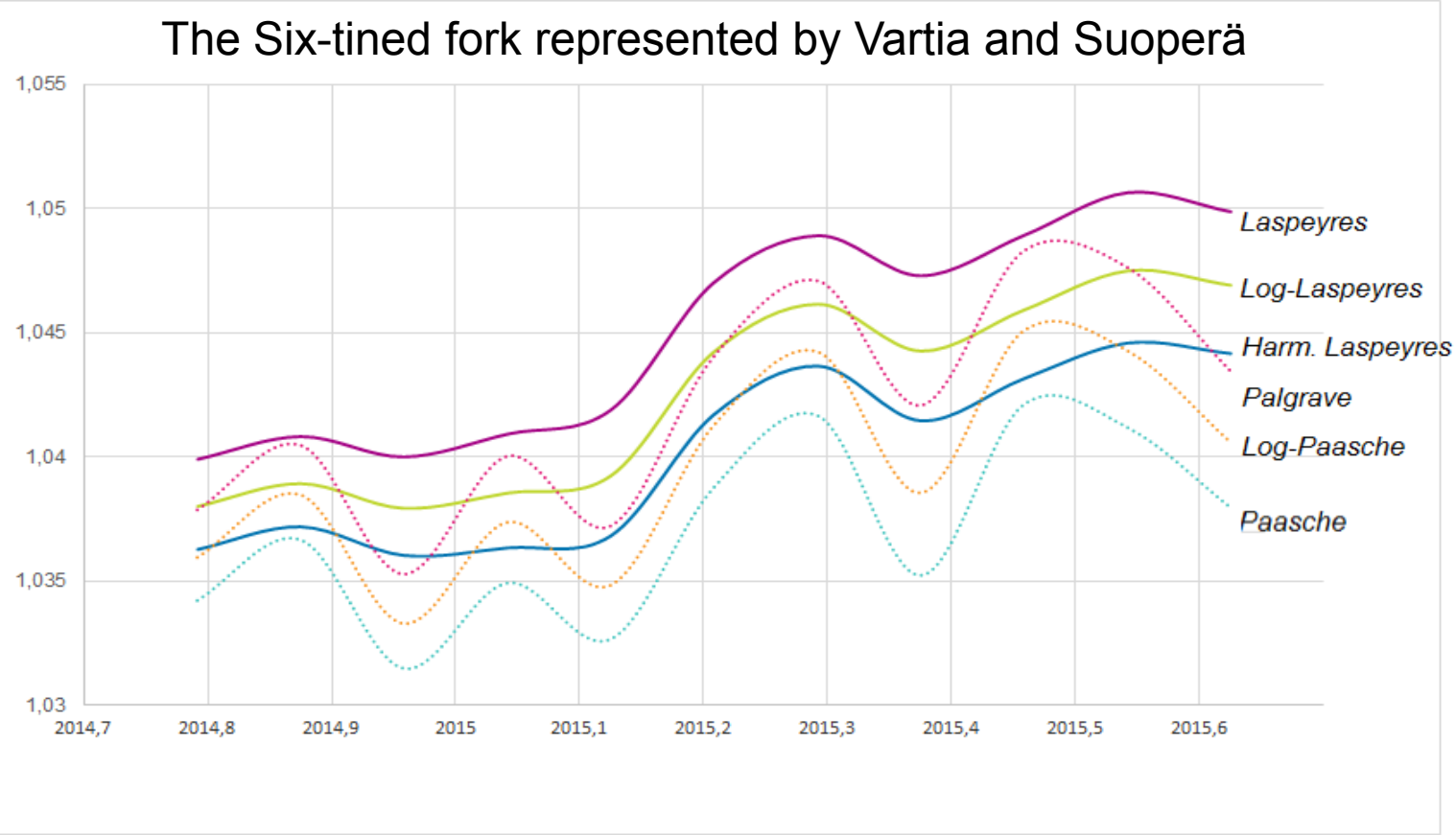


Statistics Finland

3.2 Index formula tests by Vartia & Suoperä

- Tests were accomplished in joint-work of professor Yrjö Vartia and methodologist Antti Suoperä
- Most popular index numbers were analysed
  - At first comparison between old and new weights: Laspeyreys, Paasche etc.
    >> so called Fisher-Five-tined fork
  - Then superlative index formulas : Fisher, Törnqvist, Stuvel, Diewert, Sato & Vartia, and Montgomery & Vartia
- Aim was to treat new and disappearing commodities in systematic and simple way
- Before calculations data was split in two groups:
  - 5S – commodities with larger relative change in values
  - 5N – commodities where values stay constant

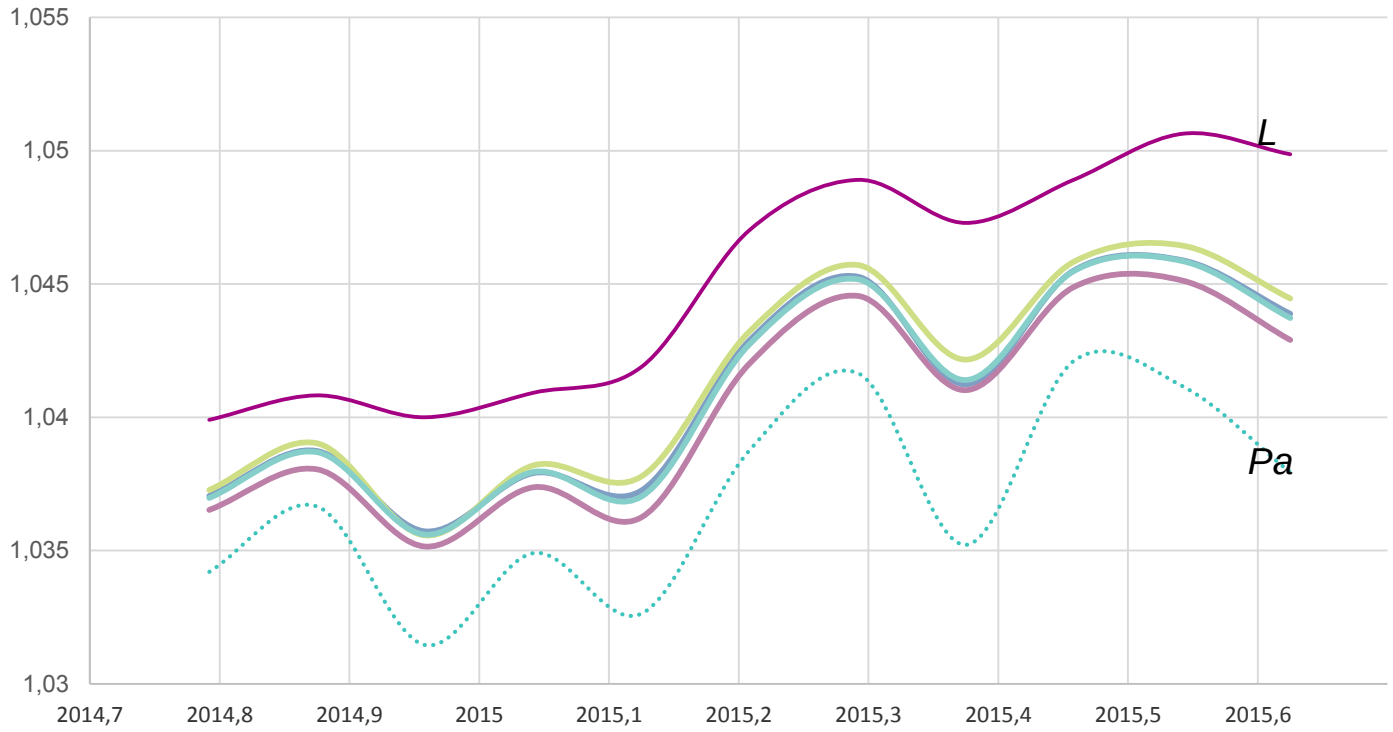
3.2 Index formula tests by Vartia & Suoperä



Statistics Finland

3.2 Index formula tests by Vartia & Suoperä

Results from the tests of superlative index formula by Vartia and Suoperä



Statistics Finland

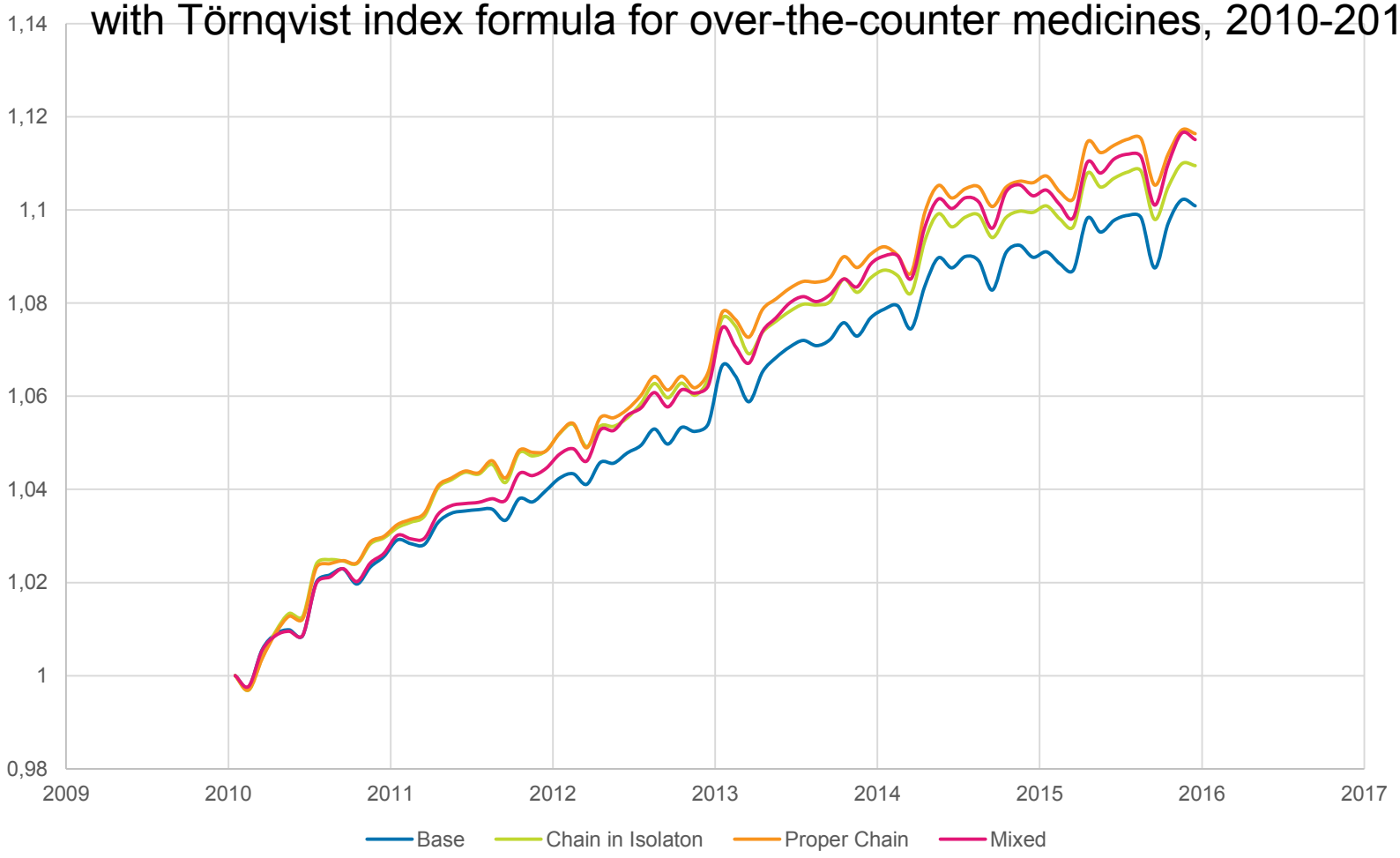
3.3 The test of chain-drift

- Aim was to analyse existence of the chain-drift and to construct new method that eliminates the chain drift phenomenon
- Following strategies were used:

Method	Formula	Sample strategy
Base Törnqvist (1)	$t_{Base}^{t/0} = exp\left\{\sum \frac{1}{2}(w_i^0 + w_i^t)\log(p_i^t/\bar{p}_i^0)\right\}$	commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities
Chain Törnqvist (2)	$t_{Chain}^{t/(t-1)} = exp\left\{\sum \frac{1}{2}(w_i^{t-1} + w_i^t)\log(p_i^t/p_i^{t-1})\right\}$	commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities
Chain Törnqvist (3)	$t_{Proper\ chain}^{t/(t-1)} = exp\left\{\sum \frac{1}{2}(w_i^{t-1} + w_i^t)\log(p_i^t/p_i^{t-1})\right\}$	Maximum number of matched pairs in base and observation periods
Mixed Törnqvist (4)	In next row, below	All commodities except new and disappearing (base Törnqvist) + new and disappearing (price ratio)
$t_{Mixed}^{2/1} = exp\{\frac{1}{2}(w_{Base}^1 + w_{Base}^2)logt_{Base}^{2/1} + \frac{1}{2}(w_{N\&D}^1 + w_{N\&D}^2)logt_{Chain,N\&D}^{2/1}\}$		

3.3 Existence of chain-drift -test

Comparison between alternative methods used
with Törnqvist index formula for over-the-counter medicines, 2010-2016



Statistics Finland

Conclusions

A lot of experience and competence achieved

When complete datasets (e.g. scanner-data) are available

- new approaches in CPI compilation may be taken
- accuracy and reliability of CPI is improved
- superlative index formulas produce more accurate index series
  - chain-drift must be controlled

Pharmaceutical products were implemented into CPI-production in the beginning of year 2017

Finland continues the tests with new data sources :

1) the daily products data obtained from the major retail chain,
2) the alcoholic beverages obtained from monopoly owner and
3) the hardware store data obtained by web-scraping

Thank you for your attention

Statistics Finland

Kristiina Nieminen / Statistics Finland, CPI-team
Kristiina.nieminen@stat.fi