Scanner Data in the CPI: The Imputation CCDI Index Revisited

Jan de Haan

Statistics Netherlands

EMG, chain drift and multilateral methods

Lorraine Ivancic (2007)

“Scanner Data and the Construction of Price Indices”

PhD thesis, School of Economics, The University of New South Wales

Evidence of chain drift in superlative price indexes

Jan de Haan (2008)

“Reducing Drift in Chained Superlative Price Indexes for Highly Disaggregated Data”, Unpublished paper

Presented at EMG Workshop 2008

“Flawed paper” ….

EMG, chain drift and multilateral methods

Lorraine Ivancic, Erwin Diewert and Kevin Fox (2011)

“Scanner Data, Time Aggregation and the Construction of Price Indexes”, *Journal of Econometrics* 161, 24-35.

Presented at EMG workshop 2009

Jan de Haan and Heymerik van der Grient (2011)

“Eliminating Chain Drift in Price Indexes Based on Scanner Data”, *Journal of Econometrics 161*, 36-46.

Results for seasonal goods presented at EMG workshop 2009

CCDI index implemented in December 2017 by the ABS

EMG, chain drift and multilateral methods

Jan de Haan and Frances Krsinich (2014)

“Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes”, *Journal of Business & Economic Statistics* 32, 341-358.

Presented at EMG workshop 2012

Quality-adjusted multilateral method

Implemented in 2014 by Statistics New Zealand for consumer electronics

Abstract of the paper

The imputation CCDI index combines the multilateral GEKS-Törnqvist, or CCDI, method with hedonic imputations for the “missing prices” of unmatched new and disappearing items. This index is free of chain drift, uses all of the matches in the data and is quality-adjusted.

We revisit the imputation CCDI index and show how it can be decomposed into the matched-item (maximum overlap) CCDI index and a quality-adjustment factor.

Outline

- Introduction
- The imputation Törnqvist price index
- The use of hedonic regression
  - Single and double imputation
- The imputation CCDI index
- Item definition and re-launches
- Concluding remarks
  - Reservation prices

(Appendix: Treatment of revisions)

Introduction

Prices and quantities known: superlative price index possible

Item churn can be significant in scanner data, especially when items are identified by barcode/GTIN

To maximize matches in the data: chaining required

High-frequency chaining of superlative price indexes often leads to drift due to sales or discounts

Chain drift is usually downward (Feenstra and Shapiro, 2003; Ivancic, 2007, Diewert, 2018)

Introduction

Ivancic, Diewert and Fox (2011) proposed the use of a multilateral method, in particular GEKS

Multilateral methods were originally developed for spatial price comparisons

When adapted to comparisons across time, these methods

- are estimated simultaneously on all the data for a given sample period or “window”
- lead to transitive indexes that are free of chain drift

Introduction

Two basic rules for good practice in price measurement

- Compare like with like (and maximize matching)
- Use an appropriate index number formula

GEKS is preferred method from economic approach to index number theory (Diewert and Fox, 2017)

GEKS-Törnqvist (CCDI) assists decomposition analysis

The CPI section at Statistics Netherlands found GEKS “too complex” to implement

Introduction

Later I proposed using weighted Time Product Dummy or, when sufficient characteristics information is available, weighted Time Dummy Hedonic (De Haan, 2015)

Statistics Netherlands has recently implemented Geary-Khamis (perhaps because they wanted an additive method)

This paper follows up on De Haan and Krsinich (2014):

- GEKS-Törnqvist (CCDI)
- Explicit quality adjustment through imputations for missing prices

Imputation Törnqvist price index

Törnqvist price index for a constant set of items *U*

$$P_T^{0t} = \prod_{i \in U} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}}$$

p_i^0 : price of item *i* in base period 0

p_i^t : price of item I in comparison period *t*; *t*= 1,…,*T*

s_i^0 : expenditure share of *i* in period 0

s_i^t : expenditure share of *i* in period *t*

The Törnqvist price index satisfies time reversal test

Imputation Törnqvist price index

Dynamic universe – new and disappearing items

Every item purchased in period 0 and/or period *t* should be included in (quantity and) price comparison between 0 and *t*

Index must be defined on the union of the item sets in 0 and *t*:

$U^0 \cup U^t = U_M^{0t} \cup U_D^{0t} \cup U_N^{0t}$

$U_M^{0t} = U^0 \cap U^t$: subset of matched items

U_D^{0t} : subset of disappearing items (available in 0, not in *t*)

U_N^{0t} : subset of new items (available in *t*, not in 0)

Imputation Törnqvist price index

- Period *t* prices for $i \in U_D^{0t}$ and period 0 prices for $i \in U_N^{0t}$ are unavailable or "missing" - requires imputations $\hat{p}_i^t$ and $\hat{p}_i^0$
- By definition: $s_i^t = 0$ for $i \in U_D^{0t}$ and $s_i^0 = 0$ for $i \in U_N^{0t}$

Leads to (single) imputation Törnqvist price index

$$P_{IT}^{0t} = \prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} \prod_{i \in U_D^{0t}} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in U_N^{0t}} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}}$$

Satisfies time reversal test if same imputed values are used for calculating index going backwards

Imputation Törnqvist price index

Imputation Törnqvist price index can be decomposed as

$$P_{IT}^{0t} = \prod_{i\in U_M^{0t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{s_{iM(0t)}^0 + s_{iM(0t)}^t}{2}} \left[\frac{\prod_{i\in U_D^{0t}} \left(\frac{\hat{p}_i^t}{p_i^0}\right)^{s_{iD(0t)}^0}}{\prod_{i\in U_M^{0t}} \left(\frac{p_i^t}{p_i^0}\right)^{s_{iM(0t)}^0}}\right]^{\frac{s_{D(0t)}^0}{2}} \left[\frac{\prod_{i\in U_N^{0t}} \left(\frac{p_i^t}{\hat{p}_i^0}\right)^{s_{iN(0t)}^t}}{\prod_{i\in U_M^{0t}} \left(\frac{p_i^t}{p_i^0}\right)^{s_{iM(0t)}^t}}\right]^{\frac{s_{N(0t)}^t}{2}} = P_{MT}^{0t} D^{0t} N^{0t}$$

P_{MT}^{0t} : matched-model (maximum overlap) Törnqvist price index

D^{0t} : effect of disappearing items

N^{0t} : effect of new items

Imputation Törnqvist price index

Similar (identical?) decomposition in Erwin Diewert, Kevin Fox and Paul Schreyer

“The Digital Economy, New Products and Consumer Welfare”, Discussion Paper 17-09, Vancouver School of Economics, UBC

Reservation prices as imputed prices (explained later)

Two slides from presentation by Kevin Fox at ESCoE conference, 16-17 May 2018, London:

Törnqvist Price Index

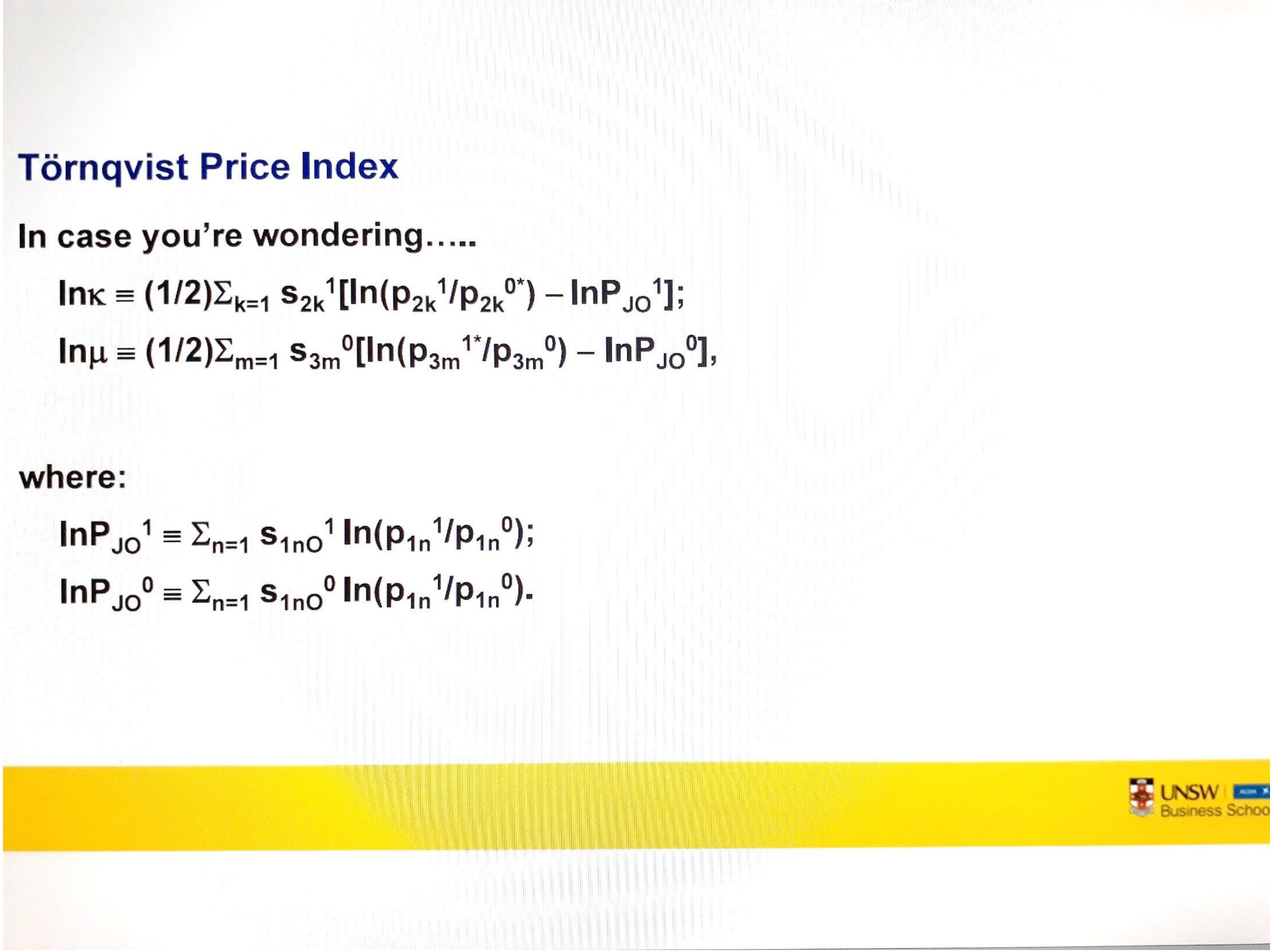
Törnqvist index is the target index for the US CPI.

Log of the Törnqvist maximum overlap index:

$$\ln P_{TO} \equiv \Sigma_{n=1}\ (1/2)(s_{1nO}{}^{0} + s_{1nO}{}^{1})\ln(p_{1n}{}^{1}/p_{1n}{}^{0})$$

Log of the true Törnqvist maximum overlap index:

$$\begin{aligned}\ln P_{T} \equiv\ & \Sigma_{n=1}(1/2)(s_{1n}{}^{0} + s_{1n}{}^{1})\ln(p_{1n}{}^{1}/p_{1n}{}^{0}) + \Sigma_{k=1}\ (1/2)(s_{2k}{}^{0} + s_{2k}{}^{1})\ln(p_{2k}{}^{1}/p_{2k}{}^{0*}) \\ & + \Sigma_{m=1}(1/2)(s_{3m}{}^{0} + s_{3m}{}^{1})\ln(p_{3m}{}^{1*}/p_{3m}{}^{0}) \\ =\ & \ln P_{TO} + \ln\kappa + \ln\mu\end{aligned}$$



Törnqvist Price Index

In case you're wondering…..

$$\ln\kappa \equiv (1/2)\Sigma_{k=1}\, s_{2k}^{1}[\ln(p_{2k}^{1}/p_{2k}^{0*}) - \ln P_{JO}^{1}];$$

$$\ln\mu \equiv (1/2)\Sigma_{m=1}\, s_{3m}^{0}[\ln(p_{3m}^{1*}/p_{3m}^{0}) - \ln P_{JO}^{0}],$$

where:

$$\ln P_{JO}^{1} \equiv \Sigma_{n=1}\, s_{1nO}^{1} \ln(p_{1n}^{1}/p_{1n}^{0});$$

$$\ln P_{JO}^{0} \equiv \Sigma_{n=1}\, s_{1nO}^{0} \ln(p_{1n}^{1}/p_{1n}^{0}).$$

The use of hedonic regression

“What the hedonic approach attempted was to provide a tool for estimating “missing prices”, prices of particular bundles not observed in the original or later periods. […..] Because of its focus on price explanation and its purpose of “predicting” the price of unobserved variants of a commodity in particular periods, the hedonic hypothesis can be viewed as asserting the existence of a reduced-form relationship between prices and the various characteristics of the commodity.”

(Ohta and Griliches, 1976)

The use of hedonic regression

Log-linear (semi-log) model

$$\ln p_i^t = \alpha^t + \sum_{k=1}^{K} \beta_k^t z_{ik} + \varepsilon_i^t$$

(item characteristics are fixed; parameters vary over time)

Estimated on data for each period separately

WLS regression - expenditure share weights

Predicted prices serve as imputed values for “missing prices” of unmatched items

The use of hedonic regression

Alternative approach (De Haan and Krsinich, 2014)

Bilateral Time Dummy Hedonic method

$$\ln p_i^t = \alpha + \delta^t D_i^{0t} + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t$$

Fixed characteristics parameters (may be too restrictive ….)

Specific type of WLS regression: $P_{TDH}^{0t} = \exp(\hat{\delta}^t)$ can be written as a single imputation Törnqvist price index

(De Haan, 2004)

The use of hedonic regression

Double imputation: observable prices of unmatched new and disappearing items replaced by predicted values

$$P_{DIT}^{0t} = \prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} \prod_{i \in U_D^{0t}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in U_N^{0t}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}}$$

$$P_{DIT}^{0t} = \prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_{iM(0t)}^0 + s_{iM(0t)}^t}{2}} \left[\frac{\prod_{i \in U_D^{0t}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0} \right)^{s_{iD(0t)}^0}}{\prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{s_{iM(0t)}^0}} \right]^{\frac{s_{D(0t)}^0}{2}} \left[\frac{\prod_{i \in U_N^{0t}} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0} \right)^{s_{iN(0t)}^t}}{\prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{s_{iM(0t)}^t}} \right]^{\frac{s_{N(0t)}^t}{2}} = P_{MT}^{0t} D_{DI}^{0t} N_{DI}^{0t}$$

The use of hedonic regression

Omitted variables bias of predicted prices in price relatives of unmatched items are likely to (partially) cancel out

(De Haan, 2004; Hill and Melser, 2008)

Relation between expenditure-share weighted single and double imputation Törnqvist price indexes

$$\frac{P_{IT}^{0t}}{P_{DIT}^{0t}} = \exp\left[\frac{s_{M(0t)}^{t}}{2} \bar{e}_{M(0t)}^{t} - \frac{s_{M(0t)}^{0}}{2} \bar{e}_{M(0t)}^{0} \right]$$

(Weighted) average residuals expected to be close to 0, so difference probably small

The imputation CCDI index

CCDI index: geometric mean of the ratios of all possible bilateral matched-item Törnqvist price index, where each link period *l* $(0 \le l \le T)$ serves as the base

(note that *l* can be greater than *t*)

$$P_{CCDI}^{0t} = \prod_{l=0}^{T} \left[P_{MT}^{0l} / P_{MT}^{tl} \right]^{1/(T+1)} = \prod_{l=0}^{T} \left[P_{MT}^{0l} P_{MT}^{lt} \right]^{1/(T+1)}$$

- Independent of choice of base period; transitive, hence free of chain drift
- Satisfies time reversal test

The imputation CCDI index

ICCDI index: bilateral single imputation rather than matched-item Törnqvist price indexes in GEKS procedure

$$P_{ICCDI}^{0t} = \prod_{l=0}^{T} \left[P_{IT}^{0l} / P_{IT}^{tl} \right]^{1/(T+1)} = \prod_{l=0}^{T} \left[P_{IT}^{0l} P_{IT}^{lt} \right]^{1/(T+1)}$$

Can be decomposed as

$$P_{ICCDI}^{0t} = P_{CCDI}^{0t} D_{SI}^{0t} N_{SI}^{0t}$$

Notions of “new” and “disappearing” become blurred in multilateral context. This impedes the interpretation of

$D_{SI}^{0t} = \prod_{l=0}^{T} [D^{0l} D^{lt}]^{1/(T+1)}$ and $N_{SI}^{0t} = \prod_{l=0}^{T} [N^{0l} N^{lt}]^{1/(T+1)}$

The imputation CCDI index

No distinction between effects of “new” and “disappearing” items:

$$P_{ICCDI}^{0t} = P_{CCDI}^{0t} \Omega_{SI}^{0t}$$

$\Omega_{SI}^{0t} = \prod_{l=0}^{T} [D^{0l} N^{0l} D^{lt} N^{lt}]^{1/(T+1)}$ measures the impact of unmatched items across estimation window 0,…,*T*;

quality-adjustment factor [no need to estimate it separately]

Similarly, DICCDI (Double Imputation CCDI) index decomposed as

$$P_{DICCDI}^{0t} = P_{CCDI}^{0t} \Omega_{DI}^{0t} \qquad \Omega_{DI}^{0t} = \prod_{l=0}^{T} [D_{DI}^{0l} N_{DI}^{0l} D_{DI}^{lt} N_{DI}^{lt}]^{1/(T+1)}$$

The imputation CCDI index

Decomposition: simple tool that shows how quality-adjusted CCDI index compares to standard matched-item CCDI index; useful for CPI compilers

Window length of T+1 periods requires estimation of $T(T+1)/2$ different bilateral Törnqvist price indexes

(e.g. 13-month window requires estimation of 72 different bilateral indexes)

Revisions when new data is added – “mean splice” (Diewert and Fox, 2017)

Item definition and re-launches

Barcode/GTIN (EAN, UPC)

- Always available in scanner data sets
- Natural key to define homogeneous items
- Calculation of unit values at barcode level (for a particular store or retail chain) straightforward

“Re-launch”: change in barcode for the “same” item, e.g. in case of slight change in type of packaging

Price changes during re-launches not captured in matched-item index (Reinsdorf, 1999; de Haan, 2003)

Item definition and re-launches

Group approach (Chessa, 2016): broadening item definition by grouping GTINs that are similar in terms of price-determining characteristics

[Use of Stock Keeping Unit (SKU) is essentially a detailed group approach]

Potential problems when only few characteristics are available:

- Defines heterogeneous items
- Causes unit value bias
- Overestimates “true” fraction of matched items

Item definition and re-launches

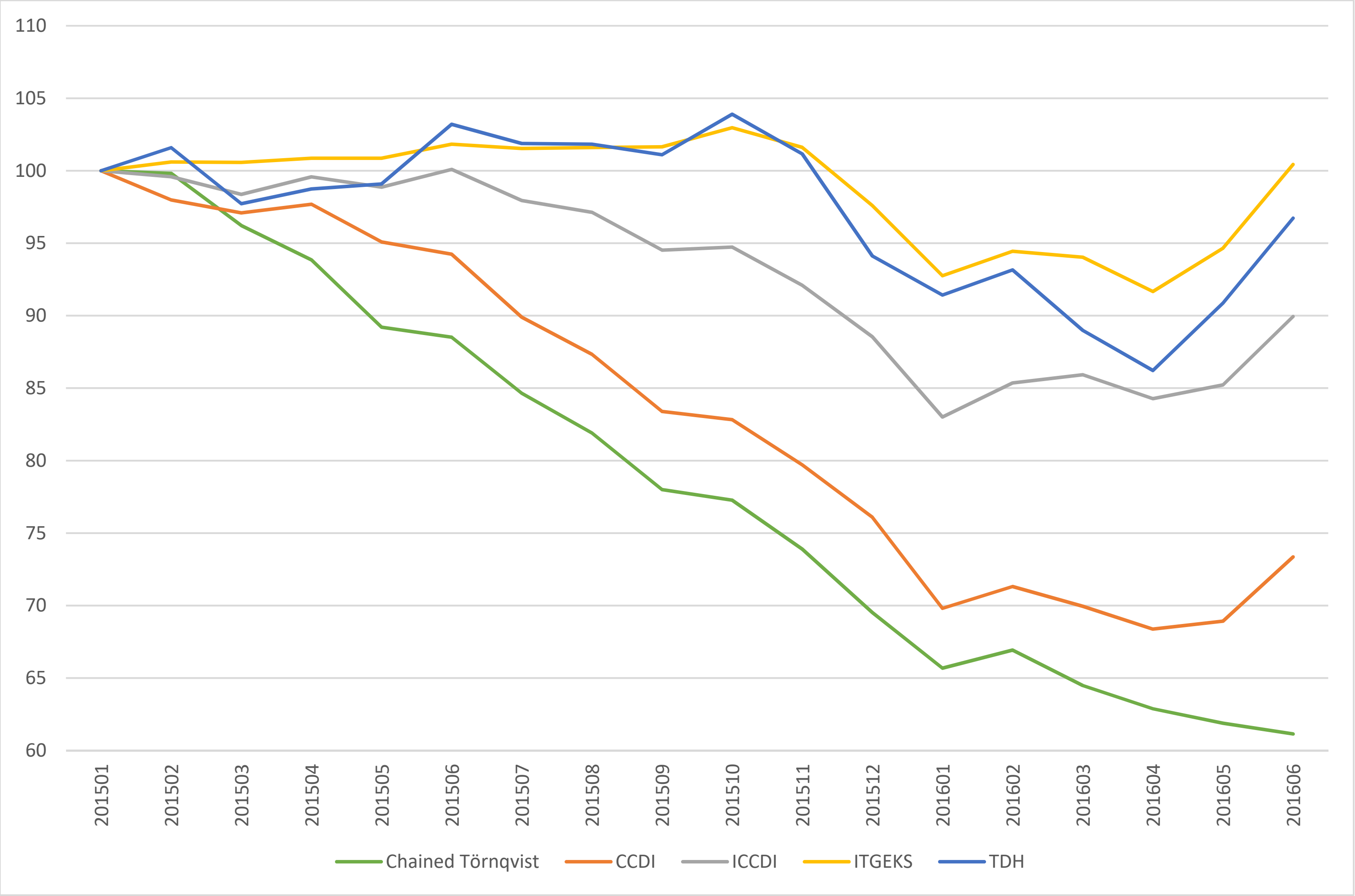
Why did Chessa (2016) used a group approach? Geary-Khamis method does not depend on imputations for “missing prices” – grouping is the only way to include characteristics information to address re-launch issue

Group approach should be avoided when using (D)ICCDI

- Identify items by barcode/GTIN or SKU
- Use characteristics that would have defined the groups as explanatory variables in hedonic model

Resulting index is free of unit value bias; hedonic imputations deal with re-launches

Example: scanner data on TVs



Concluding remarks

Diewert, Fox and Schreyer (2017), Diewert and Feenstra (2017) and Diewert (2018)

Missing prices interpreted as Hicksian reservation prices: “The reservation price for a missing product is the price which would induce a utility maximizing potential purchaser of the product to demand zero units of it”

Reinsdorf and Schreyer (2017)

Reservation prices approach relates to entirely new goods (CPI manual: evolutionary goods) rather than new variants of existing goods (evolutionary goods)

Concluding remarks

Econometric estimation of reservation prices very complicated

Alternative approach proposed by Diewert (2018)

carry forward (disappearing items) and carry backward (new items) plus inflation adjustment

- Form of implicit quality adjustment, similar to what statistical agencies are doing
- useful for temporarily missing items
- Depends on choice of measure of inflation
- Cannot resolve problem of re-launches (because of the matched-item measure for inflation adjustment)

Thank you