# Scanner Data in the CPI: The Imputation CCDI Index Revisited

Jan de Haan and Jacco Daalmans

Statistics Netherlands

# Outline

- Introduction
- Imputation Törnqvist price index
- Hedonic regression
- Imputation CCDI index
- Item definition and relaunches
- Example using TV scanner data
- Discussion

# Introduction

With scanner data, prices and quantities known: superlative index numbers possible

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

To maximize matches in the data: chaining required

High-frequency chaining can lead to drift due to sales or discounts

Chain drift is usually downward

# Introduction

Ivancic, Diewert and Fox (2011) proposed using a multilateral method, in particular GEKS-Fisher

Multilateral methods 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

Compared to (most) other multilateral methods, GEKS is preferred from economic approach to index number theory (Diewert and Fox, 2017)

GEKS-Törnqvist, referred to as CCDI, assists decomposition analysis

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

- Based on CCDI
- Explicit quality adjustment through hedonic 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*

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 a bilateral comparison between 0 and *t*

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

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

$$\vdots$$

$U_M^{0t} = U^0 \cap U^t$ 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

(Single) 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

# The use of hedonic regression

Log-linear hedonic model

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

All parameters allowed to 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 single imputation approach: “ITGEKS” (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

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

# The use of hedonic regression

Double imputation: observed 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 in predicted prices for price relatives of unmatched items may cancel out

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

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

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

If R squared is high, difference is expected to be 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

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

Without making a distinction between new and disappearing items, the index can be decomposed as

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

$\Omega_{SI}^{0t}$ is a quality-adjustment factor

# The imputation CCDI index

Similarly, DICCDI (Double Imputation CCDI) index can be decomposed as

$$P^{0t}_{DICCDI} = P^{0t}_{CCDI}\Omega^{0t}_{DI}$$

Decompositions shows how the quality-adjusted CCDI index compares to the standard matched-item CCDI index

Revisions when new data is added – extension method required, e.g. mean splice (Diewert and Fox, 2017)

# Item definition and relaunches

Barcode/GTIN (EAN, UPC)

- Available in scanner data sets from retailers
- Natural key to define homogeneous items
- Straightforward calculation of unit values at barcode level (for a particular store or retail chain)

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

Price changes during relaunches not captured in matched-item index

# Item definition and relaunches

## Stratification approach (Netherlands)

Broadening item definition by grouping GTINs that are similar in terms of a small number of price-determining characteristics

## Why stratify?

E.g., Dutch approach (Geary-Khamis) does not depend on imputations for “missing prices” – grouping needed to address relaunch issue

Trade-off between increase in heterogeneity and loss of matches (MARS; Chessa, 2018)

# Item definition and relaunches

Potential problems with stratification

- Heterogeneous items – not comparing like with like
- Unit value bias

(D)ICCDI method – no trade-off

- Items identified by barcode/GTIN or SKU
- Item characteristics used as explanatory variables in hedonic model

Resulting index is free of unit value bias; hedonic imputations deal with unmatched items, including relaunches

# Example using scanner data on TVs

- Scanner data from a major Dutch retail chain; online sales excluded
- January 2015 – May 2016; 17 months of data
- Prices at barcode level calculated as unit values across all stores
- Categorical characteristics (from web scraping):

  brand, screen size, screen type, screen resolution, screen curvature, processor type, energy class, Internet access, video on demand, 3D, DLNA, satellite receiver

# Example: TVs

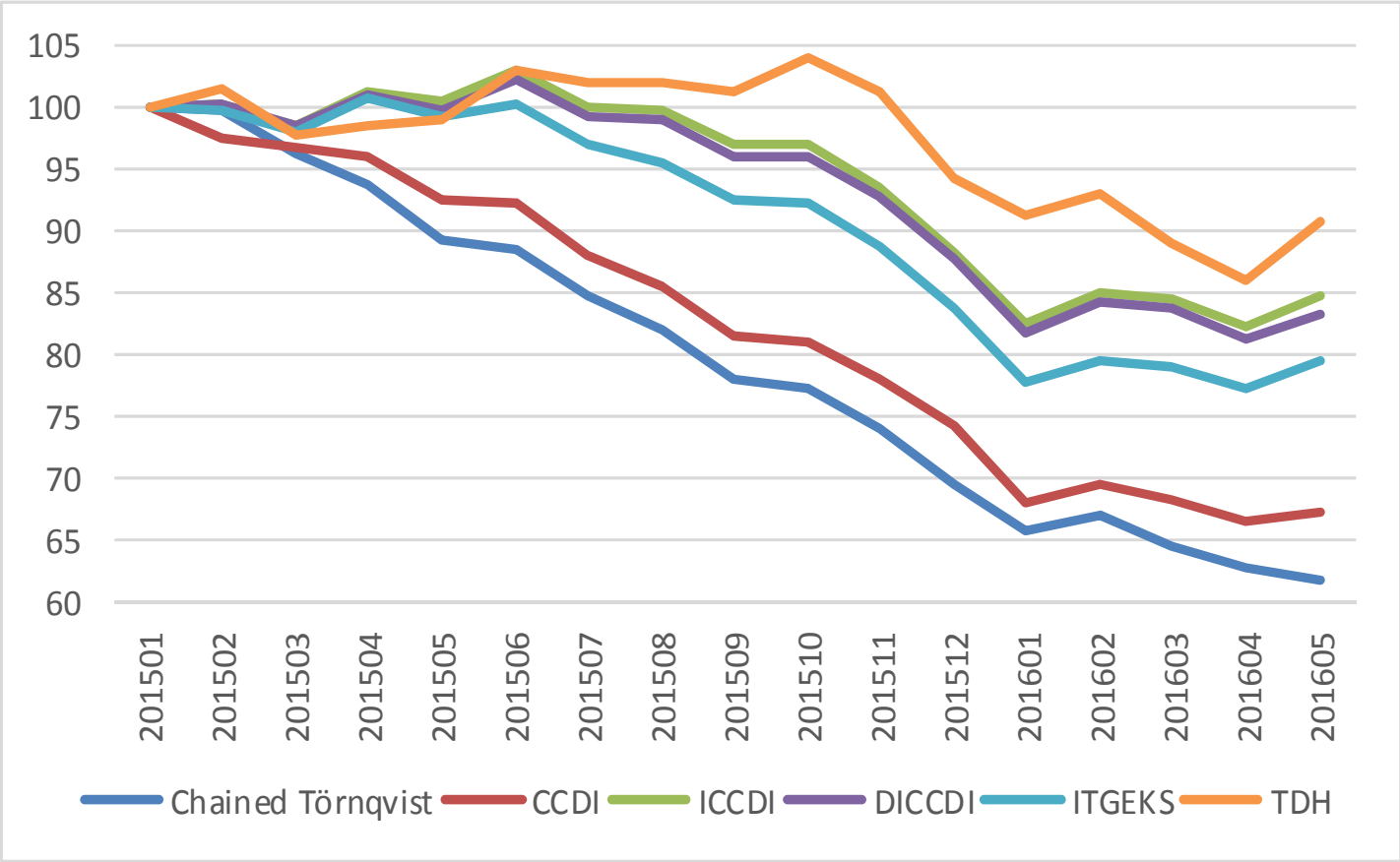
Six different price indexes, coded in R

- Chained Törnqvist
- (matched-item) CCDI

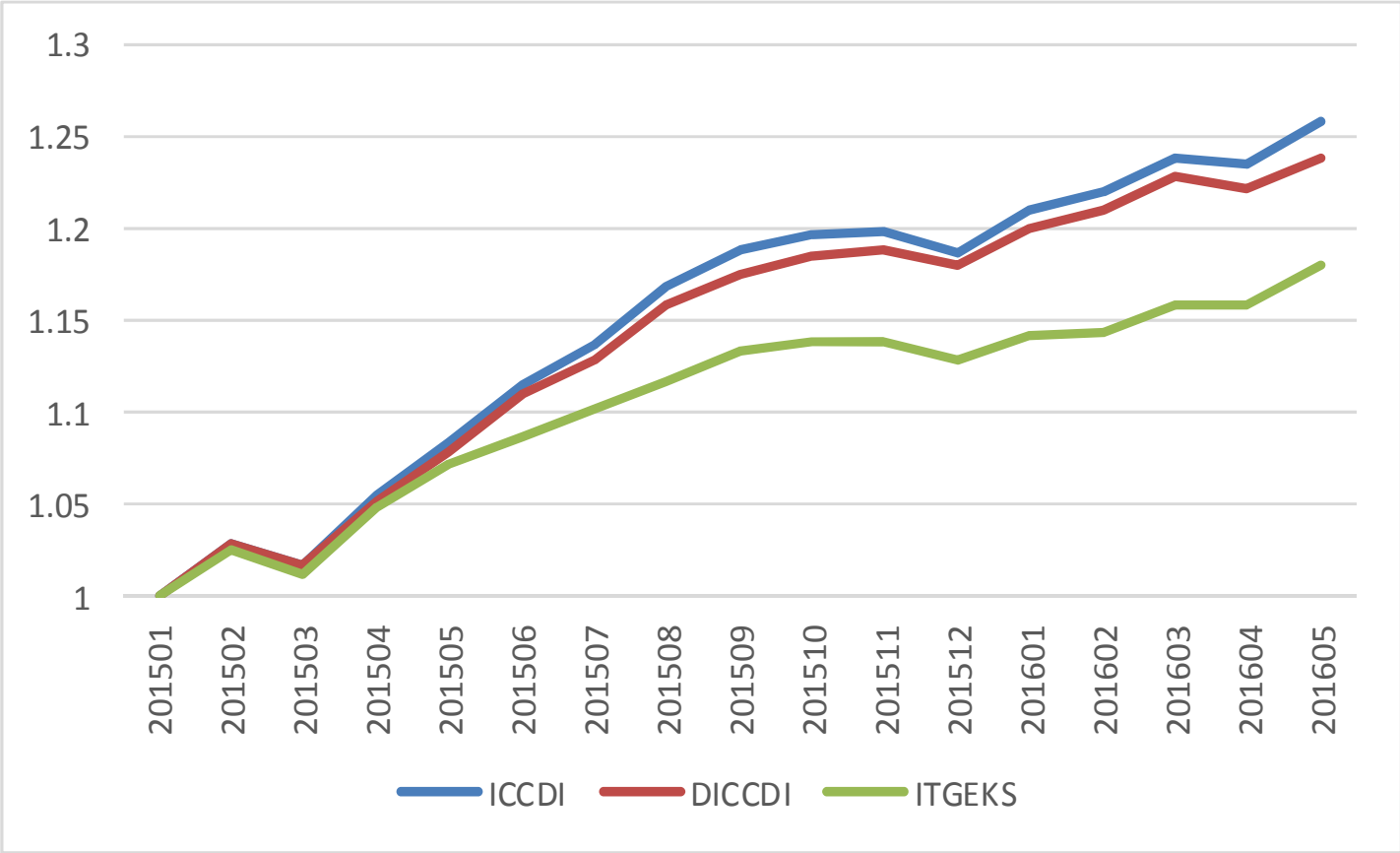
- ICCDI
- DICCDI
- ITGEKS

- Weighted multi-period Time Dummy Hedonic (TDH)

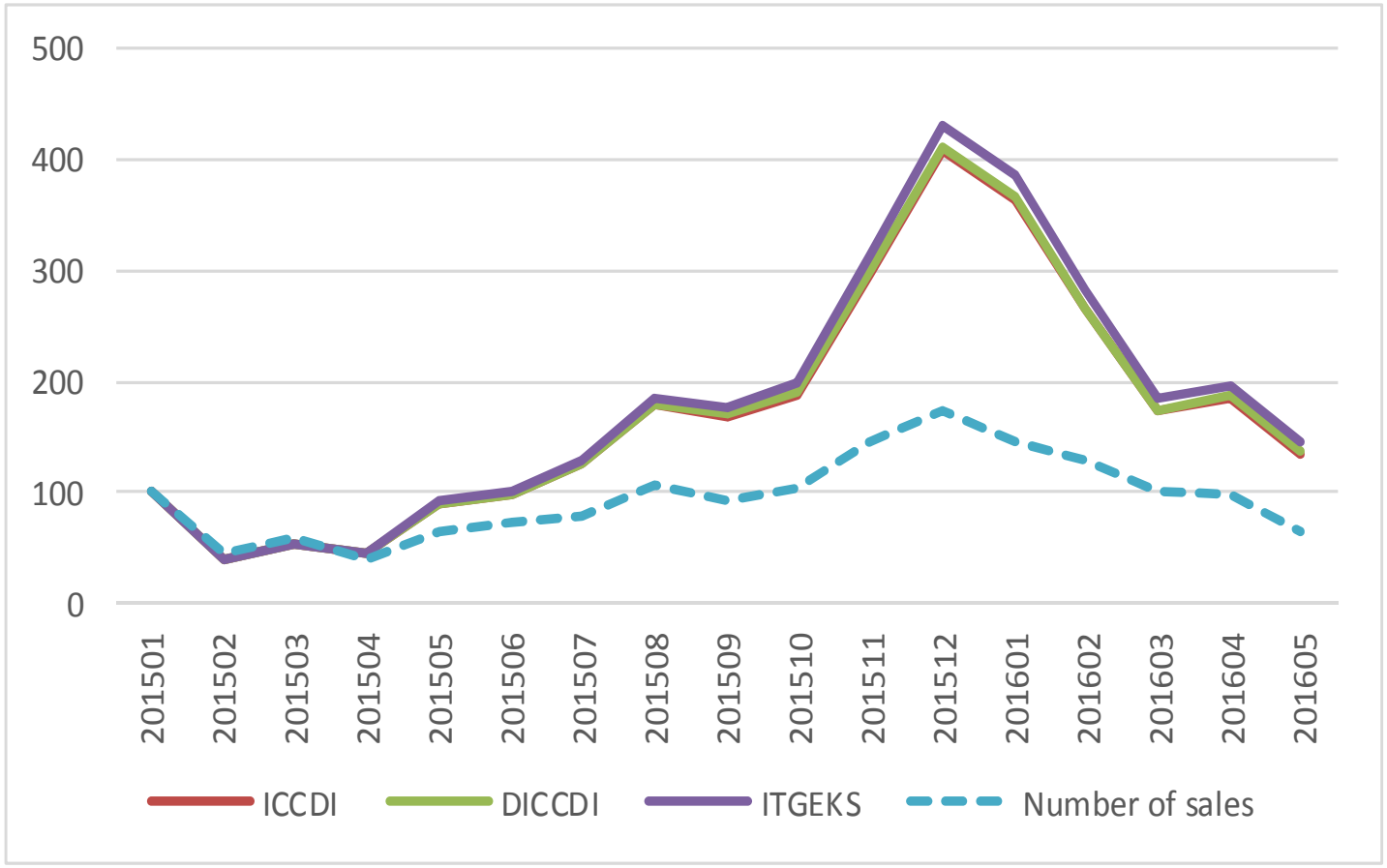
# Example: price indexes



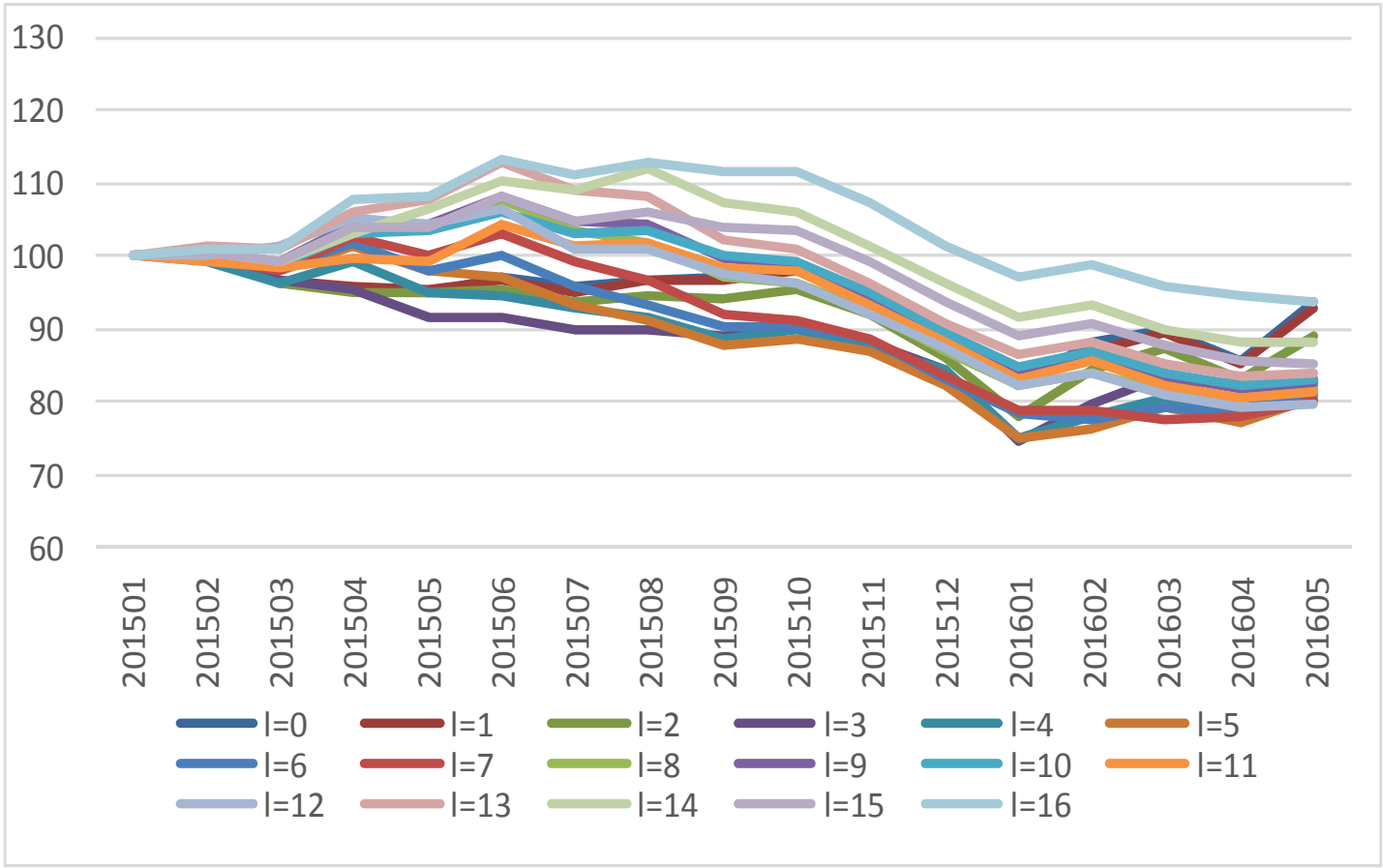
# Example: quality-adjustment factors



# Example: implicit quantity indexes



# Example: constituent indexes of ICCDI



# Potential problems

- Violation of multi-period identity test

  Diewert (2018) proposed “similarity linking” as alternative to GEKS/CCDI
- Hedonic methods depend on choice of functional form and characteristics included
- New characteristics

  Imputations in (D)ICCDI not possible; double imputation may not fully adjust
- Interpretation of hedonic imputations

  Supply restrictions (strategic choices of manufacturers or retailers; models being temporarily out of stock)?

# Reservation prices?

Lecture Erwin Diewert: missing prices treated 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"

CPI Manual

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

Thank you