18th Meeting of the Ottawa Group on Price Indices.

Session 3: Alternate data sources, quality adjustment methods and index number formulas.

Product Churn and Quality Adjustment: Using Scanner data of Laptop in Japan.

May 13, 2024

Ottawa, Bank of Canada and Statistics Canada.

Erwin W. Diewert
(The University of British Columbia)
and
Chihiro Shimizu
(Hitotsubashi University)

1. Introduction.

- **Quality Adjustment: CPI Manual Chapter 8 (2020).**

- ***Scanner data***.
  - An increasing number of business firms are willing to share their price and quantity data on their sales of consumer goods and services to a statistical office.
- Scanner data involves ***high technology products*** which are characterized by ***product churn***; i.e., the rapid introduction of **new models and products and the short time** that these new products are sold on the marketplace.

- This presentation will look at possible methods that statistical offices could use for ***quality adjusting on Laptop or high technology products***.

Contents.

- 1. Introduction.
- 2. Hedonic Regressions and Utility Theory: The Time Product Dummy Hedonic Regression Model.
- 3. The Time Dummy Hedonic Regression Model with Characteristics Information.
- 4. Laptop Data for Japan and Hedonic Regressions Using Characteristics.
- 5. Time Product Dummy Regression Models.
- 6. Traditional Indexes and Similarity Linked Price Indexes for Laptops.
- 7. Expanding Window Weighted Time Product Dummy Indexes.
- 8. Conclusion.

2. Hedonic Regressions and Utility Theory:

The Time Product Dummy Hedonic Regression Model.

3. The Time Dummy Hedonic Regression Model with *Characteristics* Information.

4. Laptop Data for Japan and Hedonic Regressions Using Characteristics.

The Laptop Data and Some Preliminary Price Indexes

- **Data:**
  - We obtained data from a private firm that collects **price, quantity and characteristic information on the *daily sales* of digital device across Japan, 2015-2024.**
  - ***Scanner data and amazon.***
  - Laptop computer and the **24 months in the years 2020 and 2021**.
  - **366 products, average life term: 24.7 months(Std.Dev 11.98).**
- Thus the prices and quantities **are p_{tn} and q_{tn}** where **p_{tn} is the average monthly (unit value) price for product n in month t and q_{tn} is the number of product n units sold**.

- **CLOCK** is the clock speed of the laptop. The mean clock speed was 1.94 and the range of clock speeds was 1 to 3.4. The larger is the clock speed, the faster the computer can make computations.
- **MEM** is the memory capacity for the laptop. The mean memory size was 8188.9. There were only 4 clock speeds listed in our sample: 4096, 8192 and 16,384.
- **SIZE** is the screen size of the laptop. The mean screen size (in inches) was 14.49. There were 10 distinct screen sizes in our sample: 11.6, 12, 12.5, 13.3, 14, 15.4, 15.6,16, 16.1 and 17.3.

- **PIX** is the number of pixels imbedded in the screen of the laptop. The mean number of pixels was 24.82. There were only 10 distinct number of pixels in our sample: 10.49, 12.46, 12.96, 20.74, 33.18, 40.96, 51.84, 55.30, 58.98 and 82.94.
- **HDMI** is the presence (HDMI = 1) or absence (HDMI = 0) of a HDMI terminal in the laptop. If HDMI =1, then it is possible to display digitally recorded images without degradation.

- **BRAND** is the name of the manufacturer of the laptop. In the data file, BRAND takes on the values 1-12 but the second brand is not present in 2020-2021 so we have only 11 brands in our sample.
  - BRAND is frequently used as an explanatory variable in a hedonic regression as a proxy for company wide product characteristics that may be missing from the list of explicit product characteristics that are included in the regression.
- **CPU** is Central Processing Units (CPUs) . The numbers of observations in each of the 10 CPU categories: 245, 702, 766, 39, 66, 87, 255, 11, 462, 6.
  - Construct the 10 vectors of dummy variables for the 10 CPU categories and denote these vectors of dimension 2639 by DU1-DU10.
- **Laptop weight** is the weight of the laptop in kilograms. Laptop weights ranged from 0.747 to 2.9 kilos.

A Hedonic Regression with *Clock Speed* as the Only Characteristic

- The price indexes P_A^t and P_{UV}^t make no adjustments for changes in the average quality of laptops over time. Thus we now consider hedonic regression models of the type.

- We start our analysis by **regressing the price vector P on the time dummy variables D_1.,,,,D_{24}** ***and dummy variables for the clock speed*** of each laptop that was sold during the sample period.

- ***Our first hedonic regression*** **sets the dependent variable vector equal to the logarithms of the product price vector P (which we denote by lnP) and the vectors in the matrix of independent variables are the time dummy variable vectors D_2, D_3,…,D_{24} and the *new 7 clock speed dummy* variable vectors D_{C1}, D_{C2}, …, D_{C7}.**

- $$\ln P = \boxed{\Sigma_{t=2}^{24}\, \rho_t D_t} + \boxed{\Sigma_{j=1}^{7}\, b_{Cj} D_{Cj}} + e$$

  where e is an error vector of dimension 2639.

A Hedonic Regression that Added *Memory Capacity* as an Additional Characteristic

- We add memory capacity as another price determining characteristic of a laptop. There were **only 3 sizes of memory capacity** (the variable MEM in the Data Appendix): 4096, 8192 and 16384. Construct dummy variable vectors of dimension 2639 for each value of MEM.
- Denote these vectors as D_{M1}, D_{M2} and D_{M3}. The new log price time dummy characteristic hedonic :
- $\mathbf{\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \boxed{\Sigma_{j=2}^{3} b_{Mj} D_{Mj}} + e.}$

A Hedonic Regression that Added *Screen Size* as an Additional Characteristic.

- There were **10 different screen sizes** (in units of 10 inches) in our sample of laptop observations. The usual commands were used to generate **10 dummy variables** for this characteristic.
- **New Groups 1 to 7 aggregated old groups 1-3, 4-8, 8-9, 10-12, 13-15, 16-18 and 19-25** respectively.
- The new log price time dummy characteristic hedonic regression:
- $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \boxed{\Sigma_{j=2}^{7} b_{Sj} D_{Sj}} + e.$

A Hedonic Regression that Added *Pixels* as an Additional Characteristic.

- There were **10 different numbers of pixels** in our sample of laptop observations. A larger number of pixels per unit of screen size will lead to clearer images on the screen and this may be utility increasing for purchasers.
- We ended up with 5 pixel groups: the new group 1 combined groups 1, 2 and 3; old group 4 became the new group 2, old groups 5 and 6 were combined to give us the new group 3, old groups 7, 8 and 9 were combined to be the new group 4 and the old group 10 became the **new group 5**.
- $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \Sigma_{j=2}^{7} b_{Sj} D_{Sj} + \boxed{\Sigma_{j=2}^{5} b_{Pj} D_{Pj}} + e.$

A Hedonic Regression that Added *HDMI* as an Additional Characteristic.

- The dummy variable that indicates the presence of HDMI in the laptop has already been generated and is listed in the Data Appendix as the column vector HDMI. Denote this column vector as D_{H2} in the following hedonic regression which adds D_{H2} to the other regressor columns:

- $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \Sigma_{j=2}^{7} b_{Sj} D_{Sj} + \Sigma_{j=2}^{5} b_{Pj} D_{Pj} + \boxed{D_{H2}} + e.$

A Hedonic Regression that Added *Brand* as an Additional Characteristic.

- BRAND takes on values from 1 to 12 but there are no brands that correspond to the number 2 in our sample for the 24 months in the years 2020 and 2021.
- Here are the numbers of observations in each of the 12 BRAND categories: 4, 0, 3,101, 6, 235, 107, 389, 489, 439, 327, 479.
- $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \Sigma_{j=2}^{7} b_{Sj} D_{Sj} + \Sigma_{j=2}^{5} b_{Pj} D_{Pj} + b_{H2} D_{H2} + \boxed{\Sigma_{j=2}^{11} b_{Bj} D_{Bj}} + e$

Time Dummy Characteristics Hedonic Regression Model: + CPU and Weight

- $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \Sigma_{j=2}^{7} b_{Sj} D_{Sj} + \Sigma_{j=2}^{5} b_{Pj} D_{Pj} + b_{H2} D_{H2} + \Sigma_{j=2}^{11} b_{Bj} D_{Bj} + \Sigma_{j=2}^{10} \boxed{b_{Uj} D_{Uj} + \Sigma_{j=2}^{7} b_{Wj} D_{Wj}} + e$
- The R^2 between the observed price vector and the predicted price vector was **0.8926**.

Weighted Time Dummy Characteristics Hedonic Regression Model.

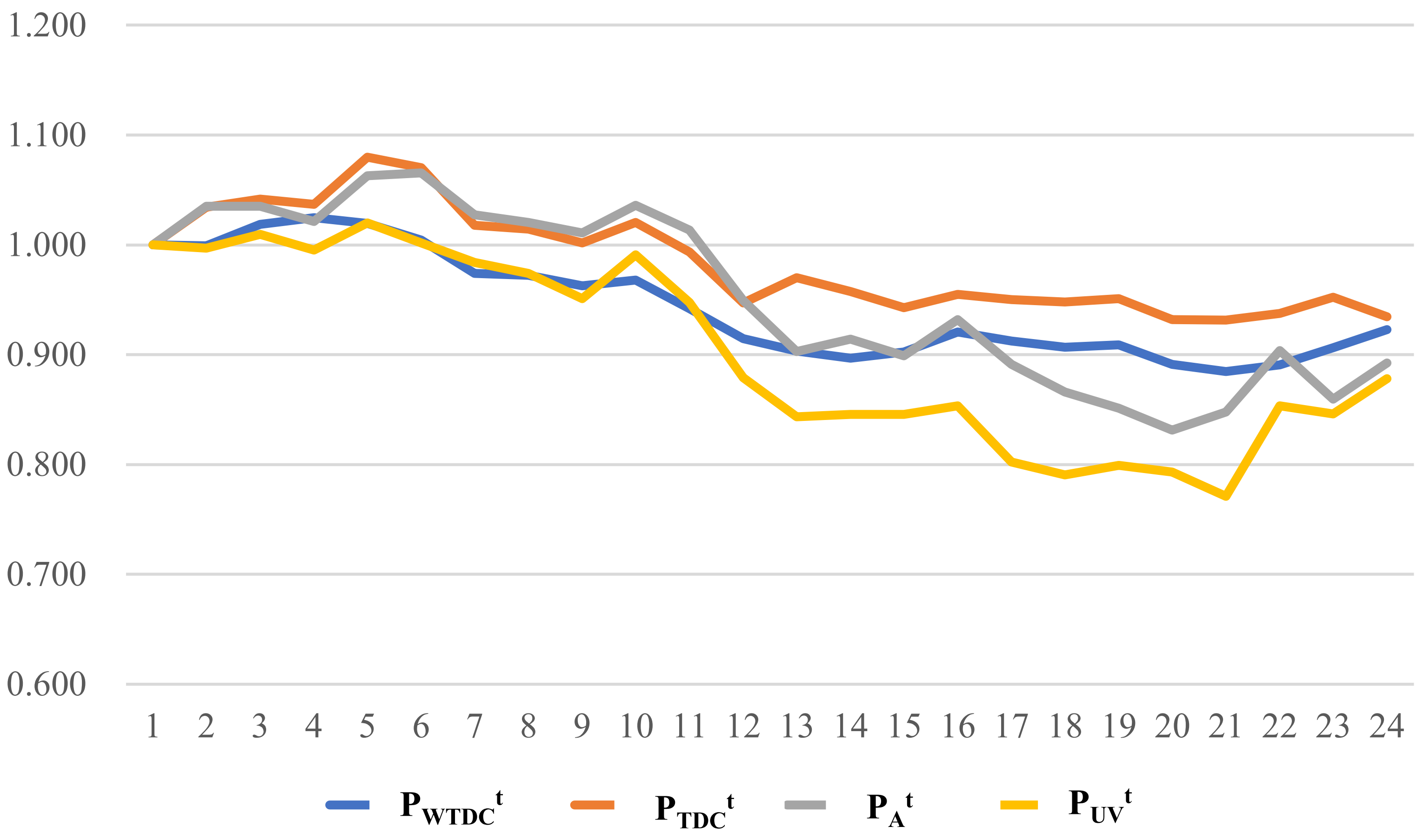
- Recall that the **expenditure share** that corresponds to purchased product n in month t is defined as $\mathbf{s_{tn} = p_{tn}q_{tn}/\Sigma_{j\in S(t)} p_{tj}q_{tj}}$ for t = 1,…,24 and n$\in$S(t).

- To obtain the weighted counterpart to the hedonic regression, we just form a share vector of dimension 2639 that corresponds to the $\ln p_{tn}$ and then form a new vector of dimension 2639 that consists of the positive square roots of each s_{tn}.

Price Indexes: Chart 1.

- **(1)P_{WTDC}^{t}** : ***Weighted Time Dummy*** *Characteristics Price Index.*
- **(2) P_{TDC}^{t}** : ***Unweighted (or equally weighted) Time Dummy*** *Characteristics Price Index.*
- **(3) P_{A}^{t}**: The *equally weighted* ***average price*** of a laptop.
- **(4) P_{UV}^{t}**: The period t ***unit value price***.

Chart 1. Time Product Dummy and Average Price Indexes.



The *Adjacent Period* Time Dummy Characteristics Hedonic Regression Model.

- There are ***two problems*** **with our "*best*" directly defined "Weighted Hedonic Price Index using characteristics, P_{WTDC}^{t}:**
  - **It is not *a real time index*; i.e., it is a *retrospective* index that is calculated using the data covering two years;**
  - **It does not allow for *gradual taste change* on the part of purchasers.**

- These difficulties can be avoided if we restrict the number of months **T to be equal to 2.** → *Adjacent period* Hedonic Regression.
  - P_{ATDC}^{t} : *Adjacent period equally weighted characteristics index*.
  - P_{WATDC}^{t} : *Weighted Adjacent period characteristics index*.

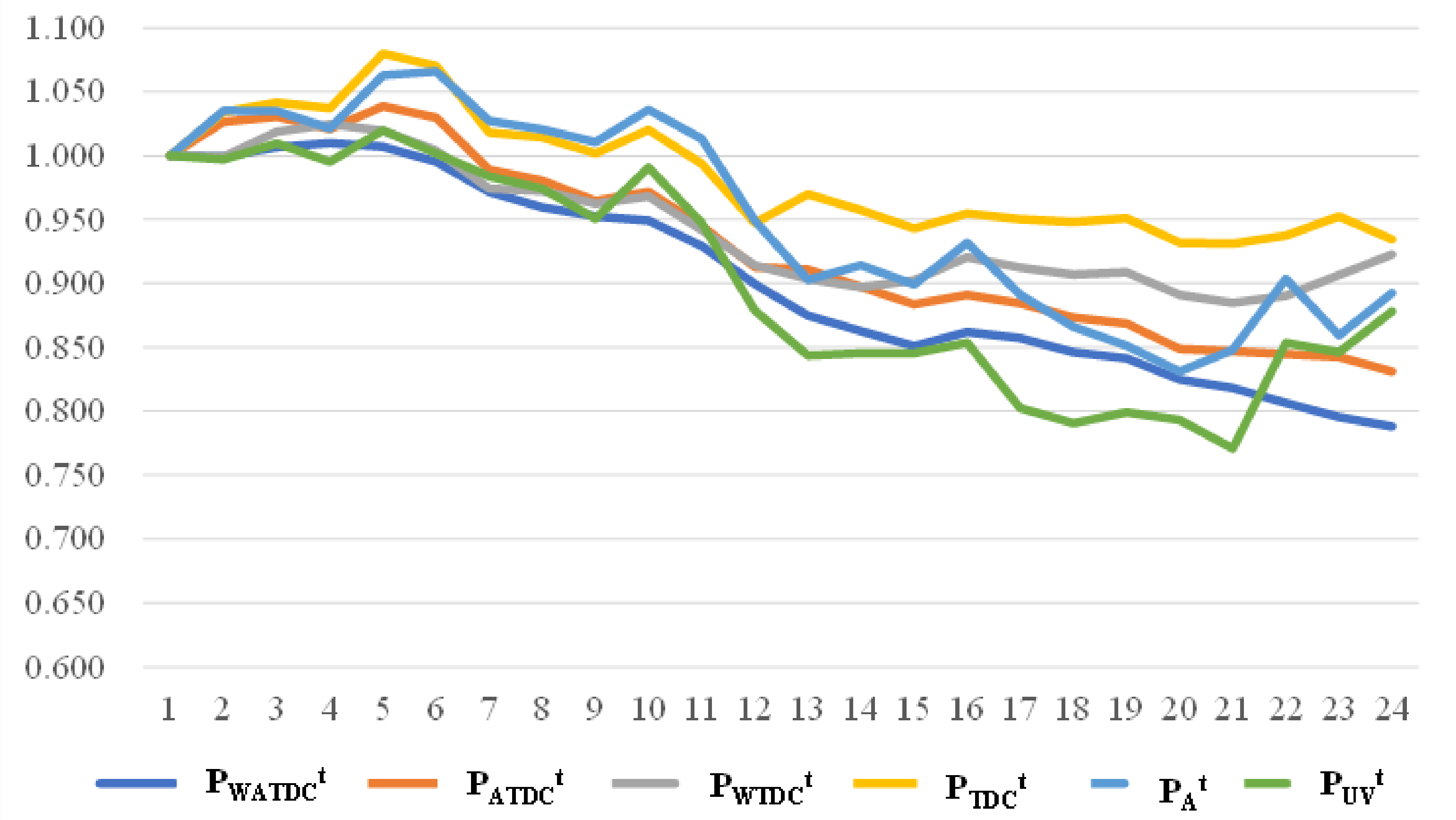
- $\mathbf{lnP = \rho_2 D_2 + b_0 ONE + \Sigma_{j=2}^{7}\, b_{Cj}\, D_{Cj} + \Sigma_{j=2}^{3}\, b_{Mj} D_{Mj} + \Sigma_{j=2}^{7}\, b_{Sj} D_{Sj} + \Sigma_{j=2}^{5}\, b_{Pj} D_{Pj} + b_{H2} D_{H2} + \Sigma_{j=2}^{11}\, b_{Bj} D_{Bj} + \Sigma_{j=2}^{10}\, b_{Uj} D_{Uj} + \Sigma_{j=2}^{7}\, b_{Wj} D_{Wj} + e}$
  - where lnP is now the vector of log prices for the products which were **sold only in months 1 and 2**.

Price Indexes: Chart2.

- **(1)P_{WATDC}^{t}** : ***Weighted Adjacent period*** *characteristics index.*
- **(2)P_{ATDC}^{t}** : ***Adjacent period equally weighted*** *characteristics index.*

- **Chart1:**
- **(3)P_{WTDC}^{t}** : *Weighted Time Dummy Characteristics Price Index.*
- **(4)P_{TDC}^{t}** : Unweighted (or equally weighted) Time Dummy Characteristics Price Index.
- **(5)P_{A}^{t} : Average Price**.
- **(6)P_{UV}^{t} : Unit Value Price**.

Chart 2. Sample Wide and Adjacent Period Time Dummy Characteristics Price Indexes.



Advantages of the *Weighted Adjacent Period* Time Dummy Characteristics indexes P_{WATDC}^{t} over the (sample wide) Weighted Time Dummy Characteristics indexes P_{WTDC}^{t}:

- The adjacent period indexes fit **the data much better since each bilateral regression estimates a new set of quality adjustment parameters** whereas the panel regression approach fixes the quality adjustment parameters over the entire window of observations.
- The adjacent period methodology that **allows the quality adjustment parameters to change every month** means that purchasers may not have stable consistent preferences over time and some economists may object to the resulting inconsistency of these indexes.

Disadvantages of the ***Weighted Adjacent Period*** Time Dummy Characteristics indexes $P_{WATDC}{}^{t}$ over the (sample wide) Weighted Time Dummy Characteristics indexes $P_{WTDC}{}^{t}$:

- The adjacent period indexes are **chained indexes**. If there are large fluctuations in the monthly product prices and quantities, then there is a danger that these indexes may be subject to the **chain drift problem**.
- Since there are large fluctuations in monthly prices and quantities in our data, there is some danger that our adjacent period indexes may be subject to some **downward chain drift**.

5. Time Product Dummy Variable Regression Models.

- We also defined the **366 product dummy variable vectors** of dimension 2639, D_{J1}, …, D_{J366}. Define the vector of the logarithms of observed laptop prices as lnP as was done in previous sections.
- Then ***the unweighted Time Product Dummy regression model*** can be expressed as the following estimating equation for the log price levels $\rho_2, \rho_3, \ldots, \rho_{24}$ and the 366 product log quality adjustment factors $\beta_1, \beta_2, \ldots, \beta_{366}$:
- $\mathbf{lnP = \Sigma_{t=2}^{24}\, \rho_t D_t \; + \Sigma_{k=1}^{366}\, \beta_k D_{Jk} + e^t.}$

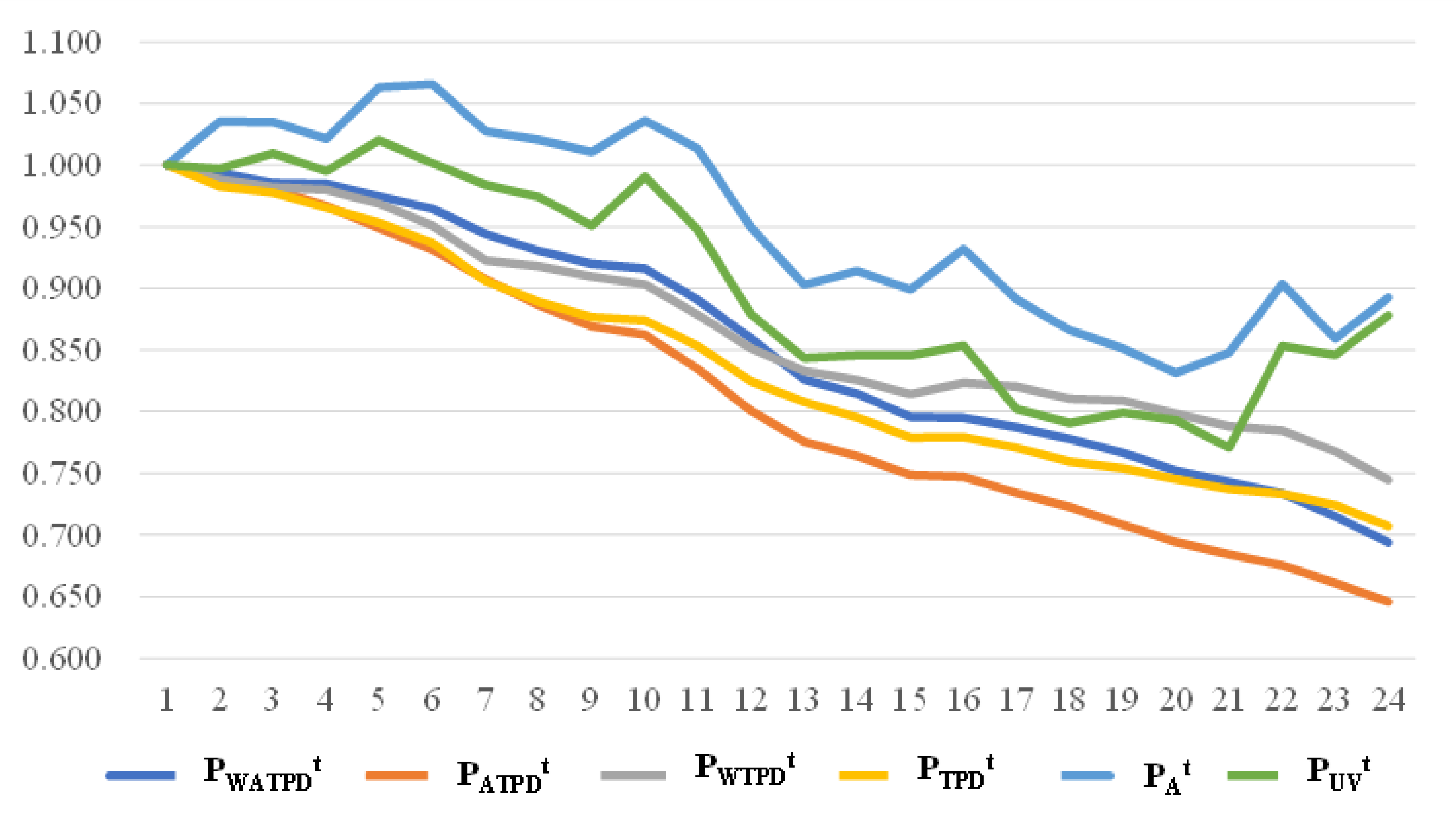
- ***Adjacent Period Time Product Dummy Price Indexes P_{ATPD}^{t}*** for t = 2, 3, …, 24.

- ***Weighted Adjacent Period Time Product Dummy Price Indexes, P_{WATPD}^{t}***.
  for t = 2, 3, …, 24.

Price Indexes: Chart 3.

- (1)P_{WATPD}^{t} : *Weighted Adjacent period Time Product Dummy Price Index.*
- (2)P_{ATPD}^{t} : *Adjacent period Time Product Dummy Price index.*
- (3)P_{WTPD}^{t} : *Weighted Time Product Dummy Price Index.*
- (4)P_{TPD}^{t} : *Unweighted (or equally weighted) Time Product Dummy Price Index.*
- (5)P_{A}^{t} : **Average Price.**
- (6)P_{UV}^{t} : **Unit Value Price.**

Chart 3. Sample Wide and Adjacent Period Weighted and Unweighted Time Product Dummy Indexes.



- We *prefer the Adjacent Period Weighted Time Product Dummy Indexes P_{WATPD}^{t}* **over** their single regression counterpart indexes, the **Weighted Time Product Dummy Indexes P_{WTPF}^{t}** for two reasons:
  - (i) the regressions which generate the P_{WATPD}^{t} **fit the data much better than the single regression** which generated the P_{WTPD}^{t} and
  - (ii) the P_{WATPD}^{t} appear to be ***smoother*** than the P_{WTPD}^{t}.
  - **Thus P_{WATPD}^{t} is our preferred index thus far.**

- Our preferred index, the ***Adjacent Period Weighted Time Product Dummy Index $P_{WATPD}{}^{t}$, is a chained index*** and thus, it is subject to possible ***chain drift***.

- In order to reduce or eliminate possible ***chain drift***, we will calculate ***Predicted Share Price Similarity linked indexes***.

6. Similarity Linked Price Indexes for Laptops.

- **The *Predicted Share* method of linking months with the most similar structure of relative prices will be explained under the assumption that it is necessary to construct a price Index P^t in real time.**

- **The *matched model Laspeyres and Paasche indexes***, $P_L(r,t)$ and $P_P(r,t)$, that relate the prices of month t to month r are defined as follows:

- $P_L(r,t) \equiv \Sigma_{k\in S(r,t)}\, p_k^t q_k^r / \Sigma_{k\in S(r,t)}\, p_k^r q_k^r$; $1 \le r, t \le 24$;
- $P_P(r,t) \equiv \Sigma_{k\in S(r,t)}\, p_k^t q_k^t / \Sigma_{k\in S(r,t)}\, p_k^r q_k^t$; $1 \le r, t \le 24$.

Price Indexes: Chart 4.

- **(1) $P_S{}^t$** :***Predicted Share Similarity Linked indexes*** .
- **(2) $P_{FCH}{}^t$** : Chained maximum overlap ***Fisher*** indexes.
- **(3) $P_{LCH}{}^t$** : Chained maximum overlap ***Laspeyres*** indexes.
- **(4) $P_{PCH}{}^t$** : Chained maximum overlap ***Paasche*** indexes.

- **(5) $P_{WATPD}{}^t$:Weighted Adjacent Period Time Product Dummy Index.**
- **(6) $P_A{}^t$: Average Price.**
- **(7) $P_{UV}{}^t$: Unit Value Price.**

Chart 4: The Predicted Share Similarity Linked Index and Other Comparison Price Indexes.

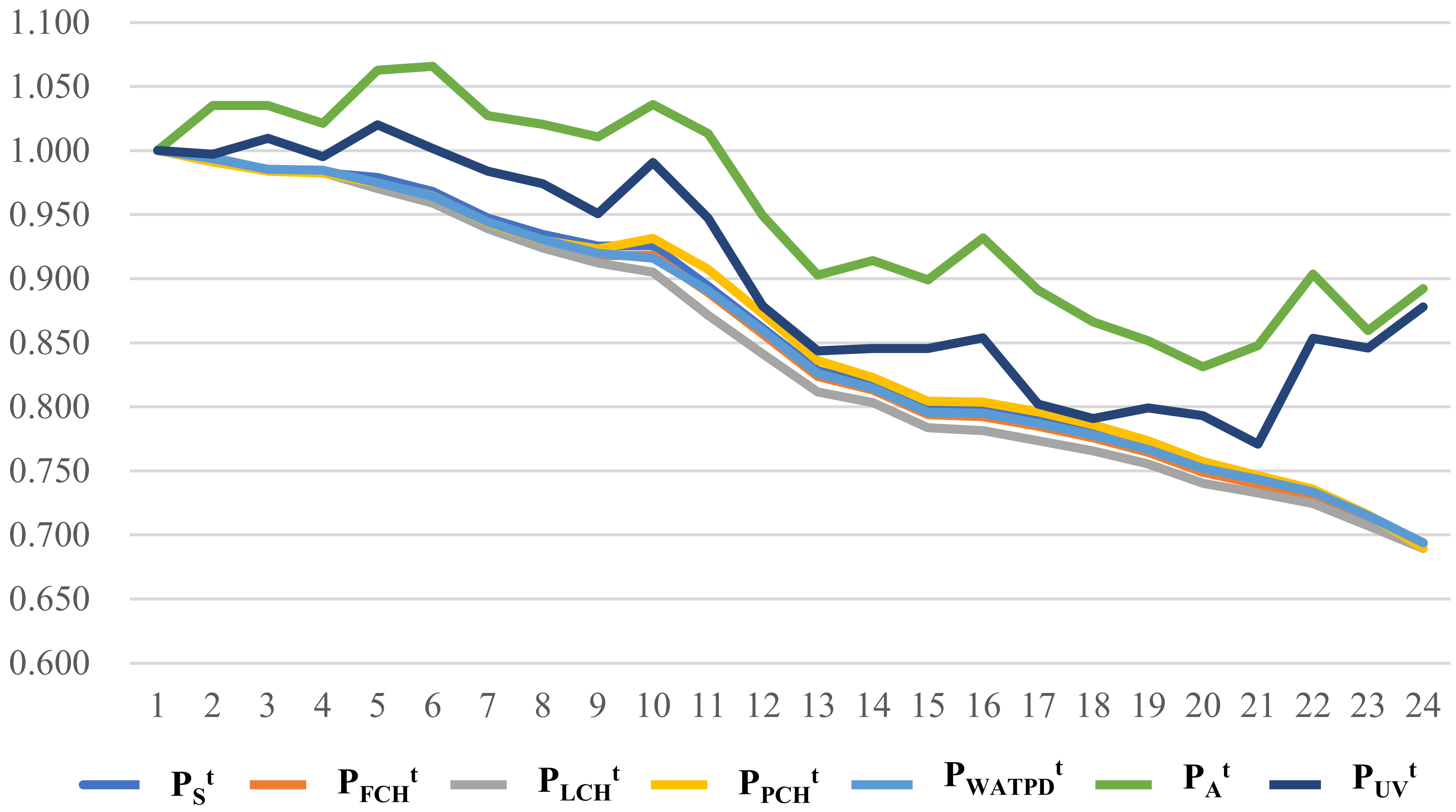
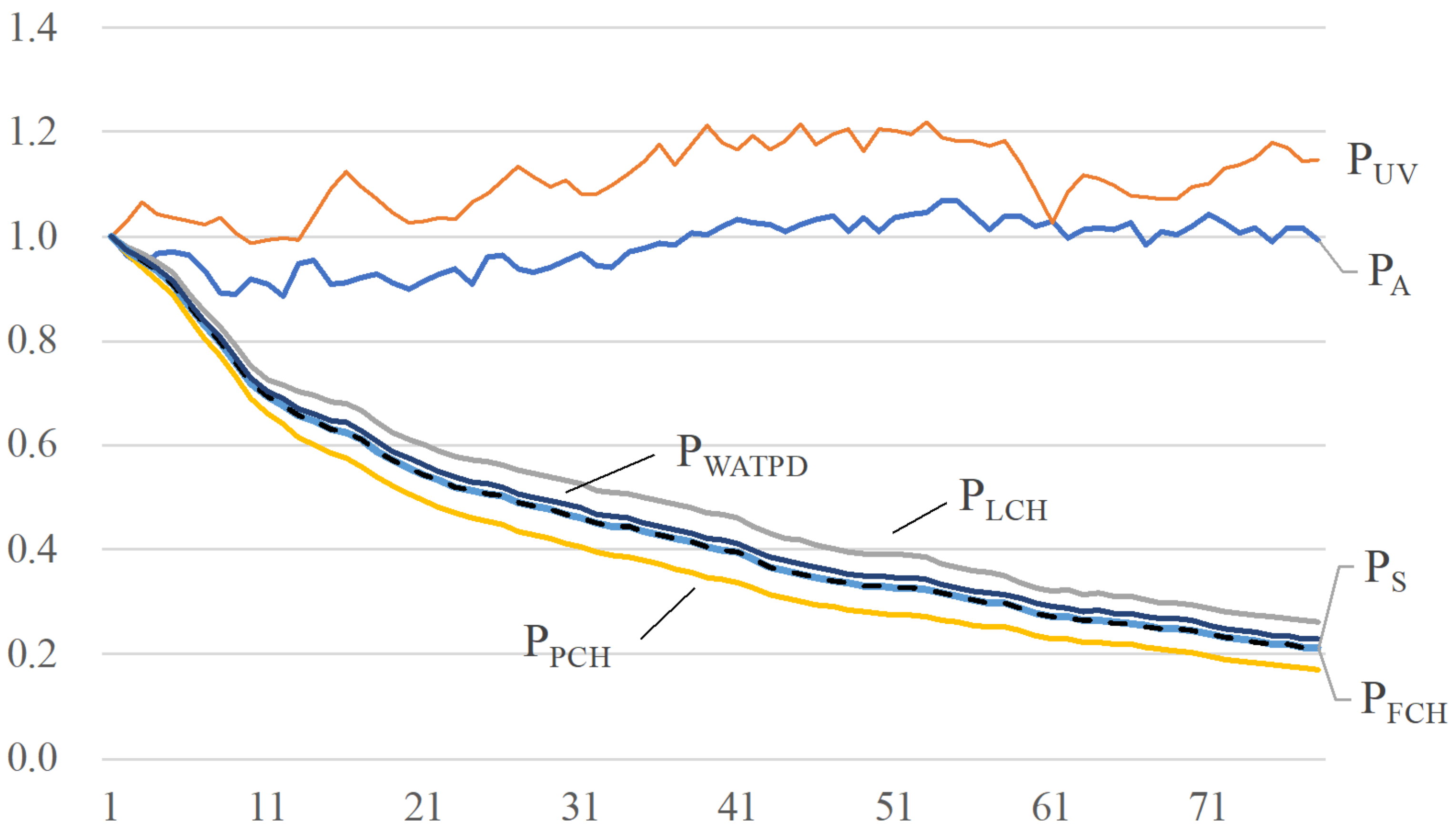


Chart 4b: The Predicted Share Similarity Linked Index and Other Comparison Price Indexes: 5 years.



- It can be seen that ***the similarity linked indexes P_S^t, the Chained Fisher maximum overlap indexes P_{FCH}^t*** **and the *Adjacent Period Weighted Time Product Dummy price indexes* P_{WATPD}^t are all extremely close to each other.**

- **These three indexes seem to be "*best*" for our particular application.**

- It can also be seen that the chained **Laspeyres and Paasche** indexes, P_{LCH}^t and P_{PCH}^t, are very close to our "***best***" indexes.

7. Expanding Window Weighted Time Product Dummy Indexes.

Expanding Window Weighted Time Product Dummy Indexes.

- We can determine whether a given price index suffers from a ***chain drift problem*** by comparing it to a "***reasonable***" index that does not suffer from ***chain drift***.
- But how exactly can we find a "***reasonable***" target index that is not subject to ***chain drift***?
  - The month t aggregate quantity levels Q^{t**} **and the Implicit Weighted Time Product Dummy price levels P^{t**}**; i.e., define:
  - $Q^{t**} \equiv \alpha^{*}\cdot q^{t}$; $P^{t**} \equiv p^{t}\cdot q^{t}/\alpha^{*}\cdot q^{t}$; t = 1,…,24.
  - It can be seen that the Q^{t**} are transitive; i.e., $Q^{3**}/Q^{1**} = (Q^{2**}/Q^{1**})(Q^{3**}/Q^{2**})$. Expenditures $e^{t} \equiv p^{t}\cdot q^{t}$ are also transitive so that $e^{3**}/e^{1**} = (e^{2**}/e^{1**})(e^{3**}/e^{2**})$.
  - **$P_{IWTPD}{}^{t} \equiv P^{t**}/P^{1**}$; t = 1,…,24.**

- ***Expanding Window Weighted Time Product Dummy price indexes*, $P_{EW}{}^{t}$, for t = 1, 2, …, 24.**
- **Step 1**: define $P_{EW}{}^{1} \equiv 1$.
- **Step 2**: Run the **weighted Time Product Dummy regression** that links months **1 and 2** .
- **Step 3**: Run a weighted Time Product Dummy regression using the data for months **1, 2 and 3**.
- **Step 4**: Run a weighted Time Product Dummy regression using the data for months 1, 2, 3 and 4 to get estimates for the β_k that correspond to products that were purchased in months **1, 2, 3 and 4**.
- …
- **Step 24**: The final step simply sets **$P_{EW}{}^{24}$ equal to the month 24 *Implicit Weighted Time Product Dummy price index***, $P_{IWTPD}{}^{24}$.

Modified Expanding Window Weighted Time Product Dummy Indexes.

- If the statistical agency is able to collect price and quantity data on the products in scope on a retrospective basis, then **Modified Expanding Window price indexes P_{MEW}^{t}** could be used.
- To construct these indexes, start off with a window of 12 months of data and construct **Implicit Weighted Time Product Dummy indexes** for this 12 month window.
- Then simply switch over to the Expanding Window price indexes for months 13 to 24. Thus the resulting indexes will be real time indexes over months 13-24.

Price Indexes: Chart 5.

- **(1)** P_{EW}^{t} : ***Expanding Window*** price indexes.
- **(2)** P_{MEW}^{t} : ***Modified Expanding Window*** price indexes.
- **(3)** P_{IWTPD}^{t} : ***Implicit Weighted*** Time Product Dummy indexes.
- **(4)** P_{WTPD}^{t} : **Weighted Time Product Dummy indexes**.

Chart 5: Expanding Window and Weighted Time Product Dummy Price Indexes.

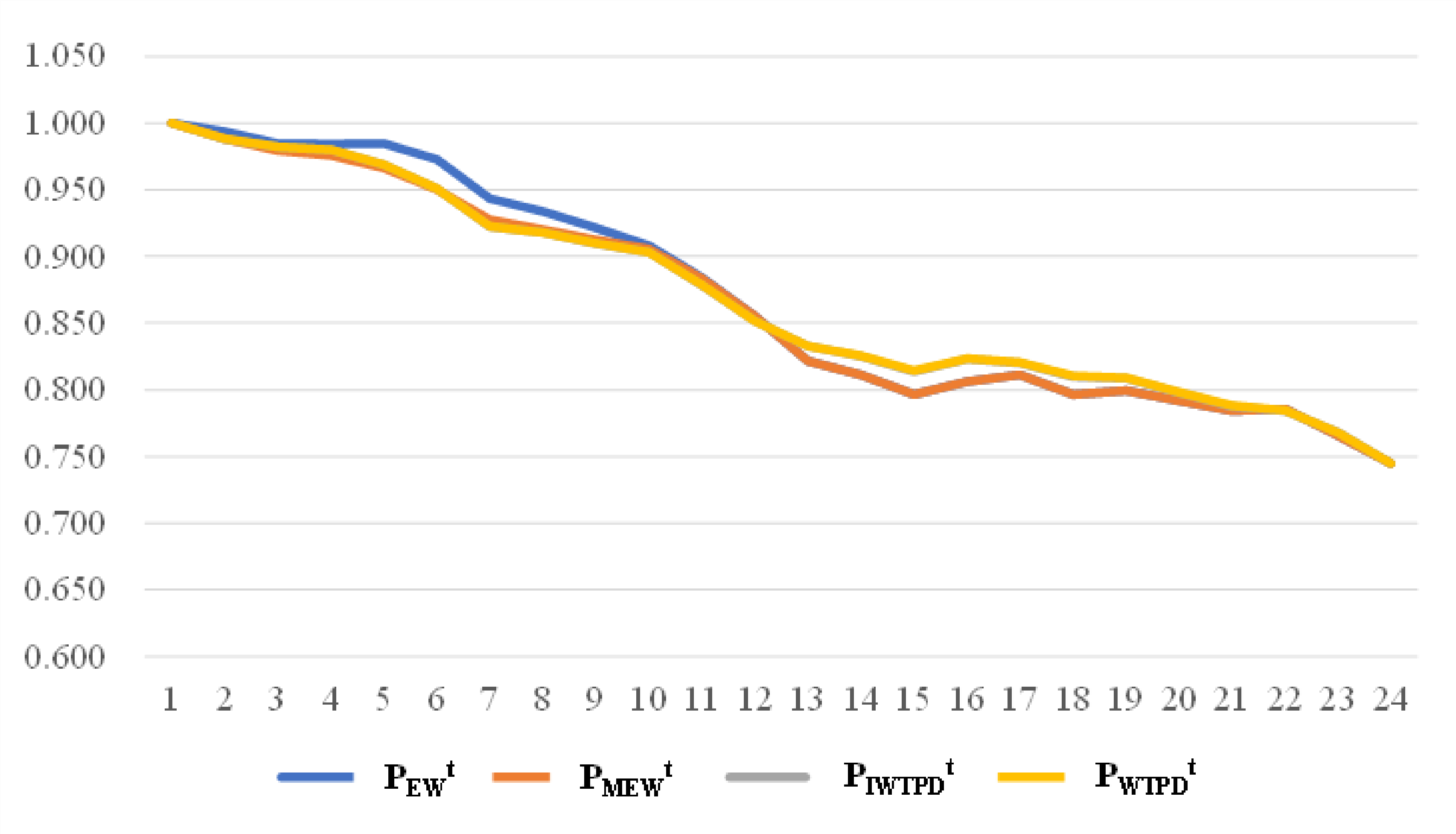
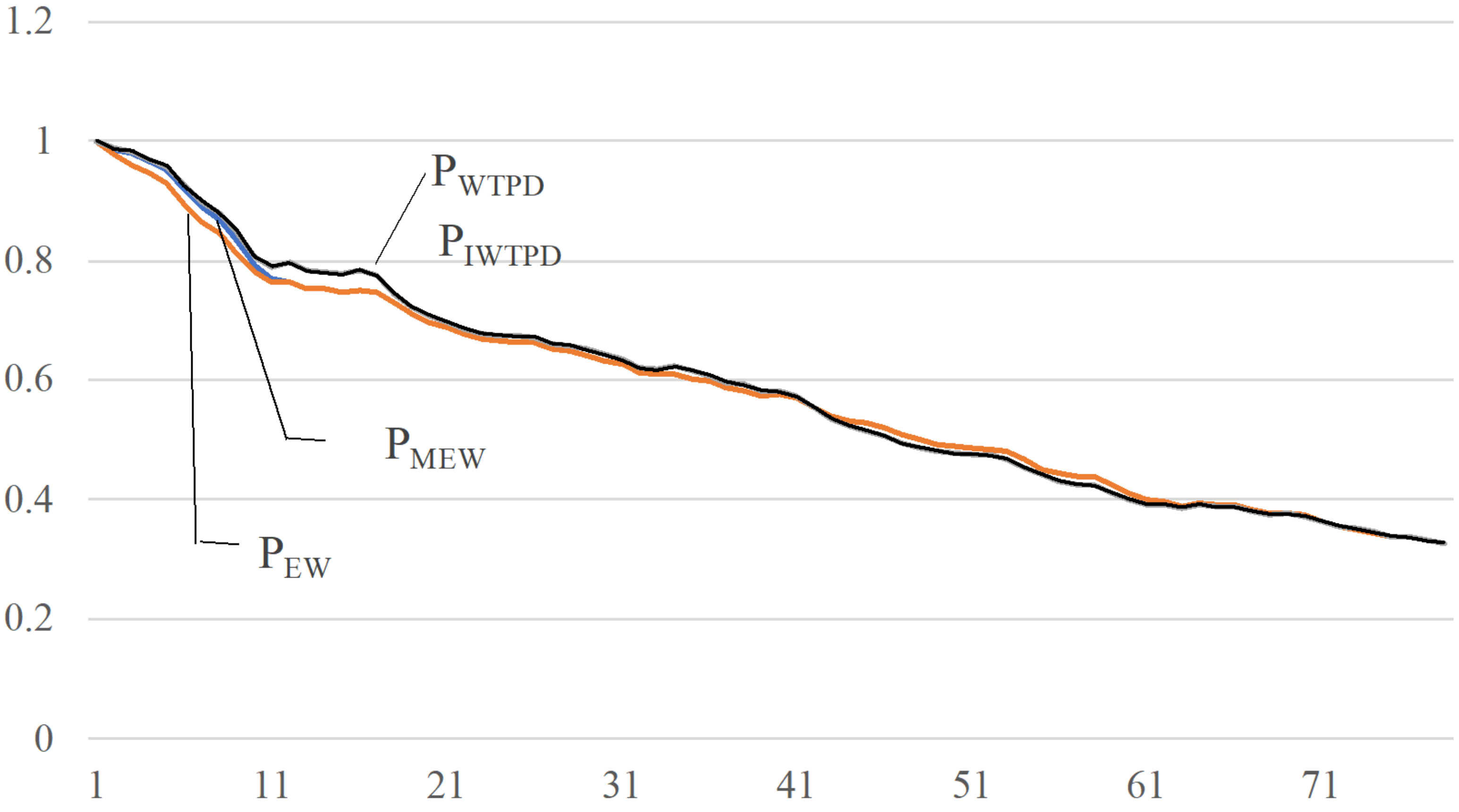


Chart 5b: Expanding Window and Weighted Time Product Dummy Price Indexes: 5 years



The *limitations* of the Expanding Window Weighted TPD model are or more generally, of a TPD hedonic regression.

- New and existing products must compete in the marketplace for more than one period.
- The Time Product Dummy model relies on the assumption that **purchasers have linear preferences over the products in scope**, at least to some degree of approximation.
- The Expanding Window Weighted TPD model does not allow for preference changes. **The Rolling Window TPD model does allow for gradual preference changes at the cost of introducing some possible chain drift.**

The ***limitations*** of the Expanding Window Weighted TPD model are or more generally, of a TPD hedonic regression.

- This limitation of the TPD methodology was recognized by Krsinich (2016; 400-401) and de Haan, Hendricks and Scholz (2021; 395).
- Again consider the extreme example **where a new product enters the marketplace every period but exits after only one month**.
- There is an extreme ***lack of matching bias***.

8. Conclusions.

- When price and quantity data for the products in scope are available, **it is best to use weighted hedonic regressions** that take into account the economic importance of the products.
  - **We found substantial differences between our weighted and unweighted (or more accurately, equally weighted) hedonic regressions.**
- The Time Dummy Characteristics approach to hedonic regressions ***did not work well*** for our particular example.
  - This approach requires data on characteristics (which is expensive) and it is subject to the missing characteristics problem. We found that the indexes changed substantially as we added additional characteristics to the regressions. **However, there are situations where the TPD approach does not work well and other approaches should be used.**

- There was a **chain drift problem** with all of our models that used chaining to link ***adjacent periods***.
  - The chain drift problem was not cured by the use of the **multilateral predicted share method** because most of the bilateral links chosen by the method were chain links.
- A satisfactory solution to the chain drift problem for our example was provided by the use of the ***Expanding Window methodology*** explained.
  - This method should work well for many product classes where substitution between the competing products is high and each product is available on the marketplace for a number of consecutive periods.

- **An interesting question for further research** is a comparison of the Expanding Window Weighted TPD method explained above with the Expanding Window Geary (1958) Khamis (1970) index that was introduced by Chessa (2016) (2021).
  - Both indexes are exact for linear preferences but the GK index can be implemented without econometric estimation.
- There are many other problems for further research that could be explored such as determining what is the “best” approach to aggregation of microeconomic data at the individual product level.
  - Simple unit value aggregation, a Rolling Window method based on a bilateral superlative index like the Rolling Window GEKS and CCDI methods or is it a regression based approach like the Expanding Window Weighted Time Product Dummy method?

W. Erwin Diewert,

University of British Columbia;

Email: erwin.diewert@ubc.ca

Chihiro Shimizu,

Hitotsubashi University;

Email: c.shimizu@r.hit-u.ac.jp