Statistics Finland

The Making of Hedonic Index Numbers

Ville Auno, Henri Luomaranta-Helmivuo, Hannele Markkanen, Satu Montonen, Kristiina Nieminen, Antti Suoperä

Presenter: Satu Montonen

Meeting of the Group of Experts on Consumer Price Indices
07 - 09 June 2023, Geneva

Content

1. Background
2. Data and data pre-processing
3. Steps of the process for producing the hedonic price index
4. Results
5. Conclusions

1. Background

- Previously, the price index for second-hand cars was calculated by Autovista Group for the purpose of CPI
- From the beginning of 2023, Statistics Finland has done the calculation itself

- The same second-hand car is not sold every month, so it is impossible to follow the price of the same car over time

- In this study, we combine hedonic quality adjusting and traditional index calculation
- In Finland, the same method is used for the prices of houses as well as for the rents of offices and shops

2. Data and data pre-processing

- Data is received on a daily basis from one major selling portal for second-hand cars in Finland
- Only the latest sales announcement of the month is considered
- The sales announcement data is supplemented with additional characteristics information from the vehicle register data from Finnish Transport and Communications Agency
- The monthly data contains approximately 75 000 individual sales announcements of second-hand cars
- For index calculation purposes, only the following are taken into account:
  - Second-hand cars with ”sold”-status purchased from car dealers
  - Passenger cars
  - Cars aged between one and twenty years
  - Cars with price greater than 2000 euros
  - Mileage needs to be less than one million kilometers

3. Steps of the process for producing the hedonic price index

Definition and estimation of price model
incl. statistical tests

Aggregation and Oaxaca-decomposition

Index calculation

3.1 Definition and estimation of price model 1/5

- The price model is semilogarithmic:

$$log(p_{it}) = \alpha_{01t} + \cdots + \alpha_{0k_1t} + x'_{it}\beta_t + \varepsilon_{it},$$

where p is the unit price of a second-hand car, parameters α represent stratum effects and term ε is random error term

- The unknown parameters β and α are estimated using the ordinary least squares method (OLS)

The explanatory variables used in the price model

Variable	Description
x_1	Gearbox type: If automatic $x_1 = 1$, else $x_1 = 0$.
x_2	Towing hook: If towing hook $x_2 = 1$, else $x_2 = 0$.
x_3	Service history: If service history is available $x_3 = 1$, else $x_3 = 0$.
x_4	Cruise control: If cruise control $x_4 = 1$, else $x_4 = 0$.
x_5	Selling time of a car, months.
$x_6 = sqrt(x_5)$	Square root of the selling time of a car.
x_7	Age of a car, years.
$x_8 = sqrt(x_7)$	Square root of the age of a car.
x_9	Mileage (ten thousand).
$x_{10} = sqrt(x_9)$	Square root of mileage.
x_{11}	Power/Weight ratio of a car.
$x_{12} = sqrt(x_{11})$	Square root of Power/Weight of a car.

3.1 Definition and estimation of price model 2/5

- We define several hierarchical partitions of second-hand cars (homogenous stratums)
- Using the F-test, we select the suitable partition: model 6

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
	No categori-zation	Size of a car	Size of a car × Make	Size of a car × Make × Model	Size of a car × Make × Model × Driving Power	Size of a car × Make × Model × Driving Power × Type of a car
		Model 1 vs 2	**Model 2 vs 3**	**Model 3 vs 4**	**Model 5 vs 4**	**Model 6 vs 5**
Test statistic		11896	1872	711	36.8	10.7

3.1 Definition and estimation of price model 3/5

- We define several classifications of price models
- Using the F-test, we select the suitable classification of price model: model 8

	Model 6	**Model 7**	**Model 8**
	No heterogeneity	Size of a car	Size of a car × Make
		Model 7 vs 6	**Model 8 vs 7**
Test statistic		206.5	45

3.1 Definition and estimation of price model 4/5

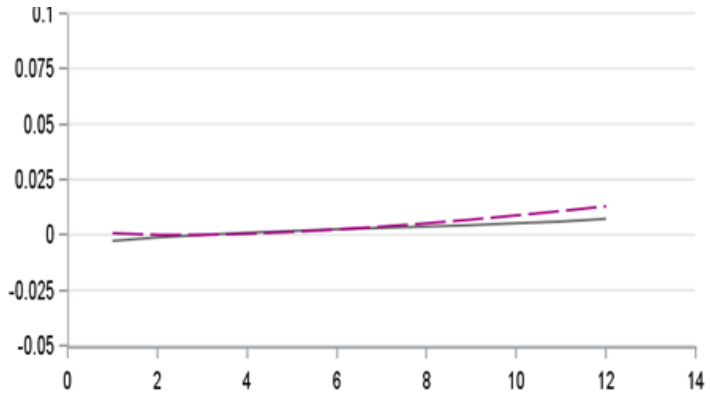
Year	2020	2021
Number of observations	287936	269663
Number of equations	72	74
Number of stratums/categories	1594	1691
Degrees of freedom	285478	267084
SSE	5401.6405077	4908.43633
R2	0.9645034599	0.9675392005
RMSE	0.1375550427	0.1355650208

	2020	2021
Constant	9.9126394001	9.8211262087
If automatic gearbox $x_1 = 1$, else $x_1 = 0$	0.0902673948	0.0923941505
If towing hook $x_2 = 1$, else $x_2 = 0$	0.0118209506	0.0113174535
If service history is available $x_3 = 1$, else $x_3 = 0$	-0.010492392	-0.008856039
If cruise control $x_4 = 1$, else $x_4 = 0$	0.017682513	0.0190084745
Selling time of a car, x_5	-0.000386744	0.0036841099
$x_6 = x_5^{1/2}$	0.0054383443	-0.012634214
Age of a car, x_7	-0.138809764	-0.135251635
$x_8 = x_7^{1/2}$	0.2915511757	0.2950576677
Mileage, x_9	-0.033047764	-0.033221364
$x_{10} = x_9^{1/2}$	0.0180405738	0.026330353
Power/Weight ratio of a car, x_{11}	12.089654612	9.8976375615
$x_{12} = x_{11}^{1/2}$	-2.549090343	-1.520907481

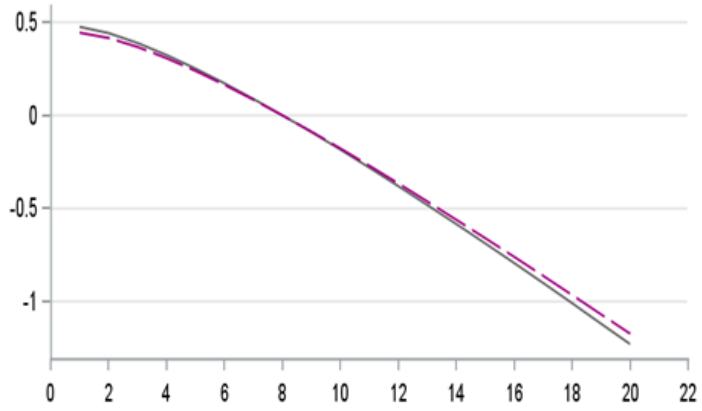
- The price model is estimated for each year
- Estimation results for model 8
  - Selling time of a car has little effect on price
  - Age of a car and mileage have a negative effect on price
  - Power/Weight ratio of a car has a positive effect on price

3.1 Definition and estimation of price model 5/5

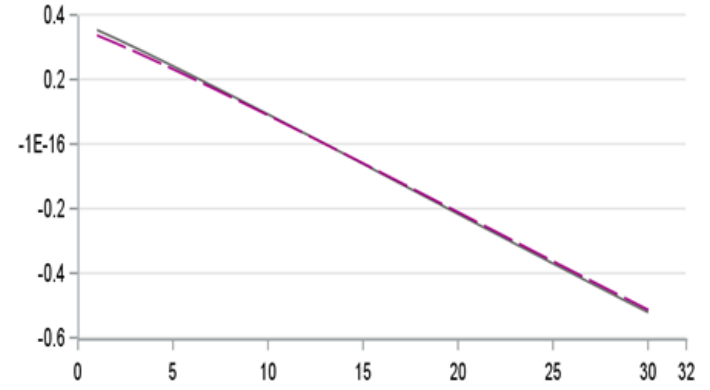
The price effect of selling time (months) on the average log-prices in year 2020 and 2021



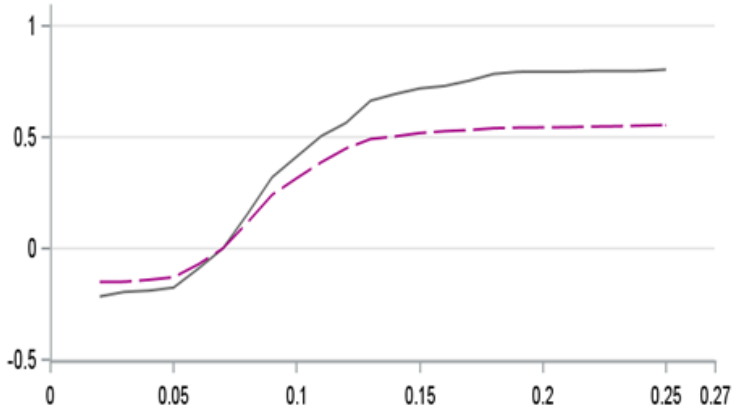
The price effect of age (years) on the average log-prices in year 2020 and 2021



The price effect of mileage (ten thousand) on the average log-prices in year 2020 and 2021



The price effect of power/weight ratio (kW/kg) on the average log-prices in year 2020 and 2021



3.2 Aggregation and Oaxaca-decomposition

- We aggregate price models from observations into stratums of the partition
- We test unweighted geometric and arithmetic averages in aggregation
- The quality adjusting is performed using decomposition introduced by Oaxaca (1973)
  - The decomposition splits the actual average price change into quality corrections and quality adjusted price changes for any stratum

(1) Price-ratio = {Quality corrections } + {Quality adjusted price change conditional on $\overline{\boldsymbol{x}}'_{kt}$}

A = QC + QA

- The equation (1) can be represented as

$$log(\bar{p}_{kt}/\bar{p}_{k0}) = log(\tilde{p}_{kt}/\bar{p}_{k0}) + log(\bar{p}_{kt}/\tilde{p}_{kt}),$$

where $log(\bar{p}_{kt})$ is the average price for the current month, $log(\bar{p}_{k0})$ is the average price for the base period and

$log(\tilde{p}_{kt}) = \hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{j0}$ is the current month's estimated price using the base period valuation of characteristics $\widehat{\boldsymbol{\beta}}_{j0}$

- The price model estimates used are always from the base period

3.3 Index calculation

- The averaged stratum-level price decompositions are summed up to COICOP7-level using weights $w_{k,f}$ of index number formula f

  $exp\{\sum_k w_{k,f}\ log(\bar{p}_{kt}/\bar{p}_{k0})\} = P_{f,A}^{t/0}$ is the price index for average prices (A)

  $exp\{\sum_k w_{k,f}\ log(\tilde{p}_{kt}/\bar{p}_{k0})\} = P_{f,QC}^{t/0}$ is the price index for quality corrections (QC)

  $exp\{\sum_k w_{k,f}\ log(\bar{p}_{kt}/\tilde{p}_{kt})\} = P_{f,QA}^{t/0}$ is price index for quality adjusted price changes (QA)

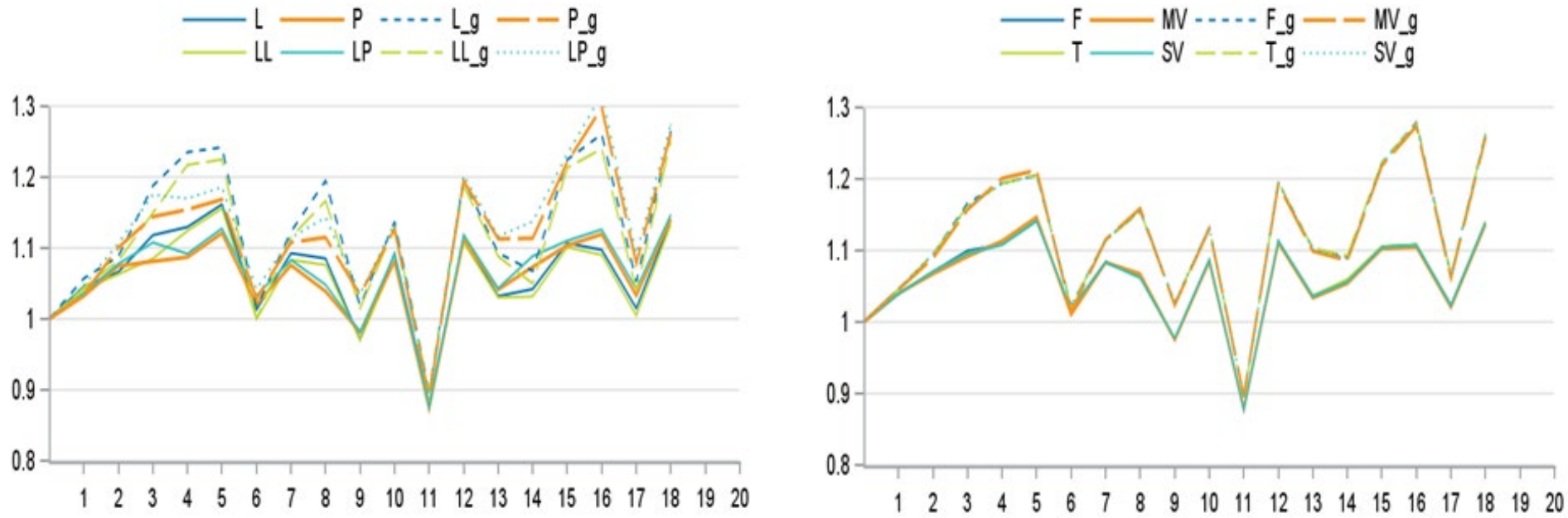
  that satisfy the following equation

$$P_{f,A}^{t/0} = P_{f,QC}^{t/0} \cdot P_{f,QA}^{t/0}$$

- In our case the base period is a previous year normalized as an average month
  - We use the flexible basket approach
- We test different index number formulas

4. Results 1/3

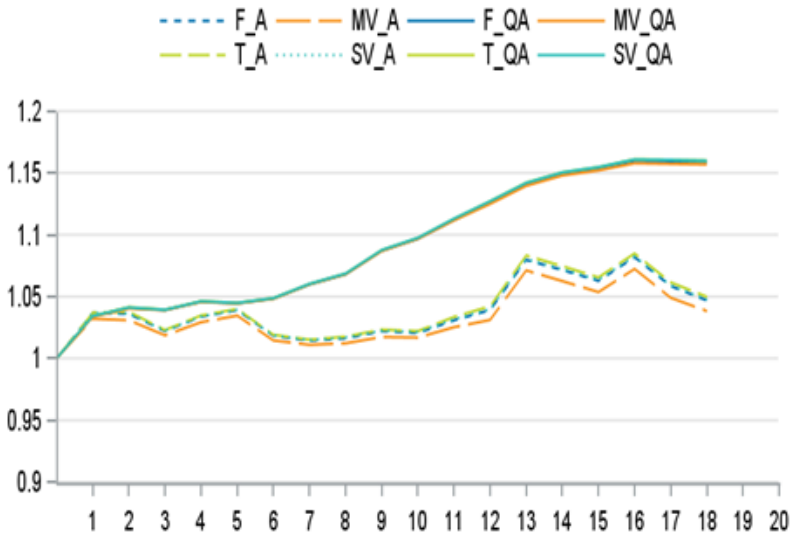
- Index series for actual average prices for ‘Small cars’ make ‘Honda’. Indices based on geometric are dotted lines and arithmetic are solid lines

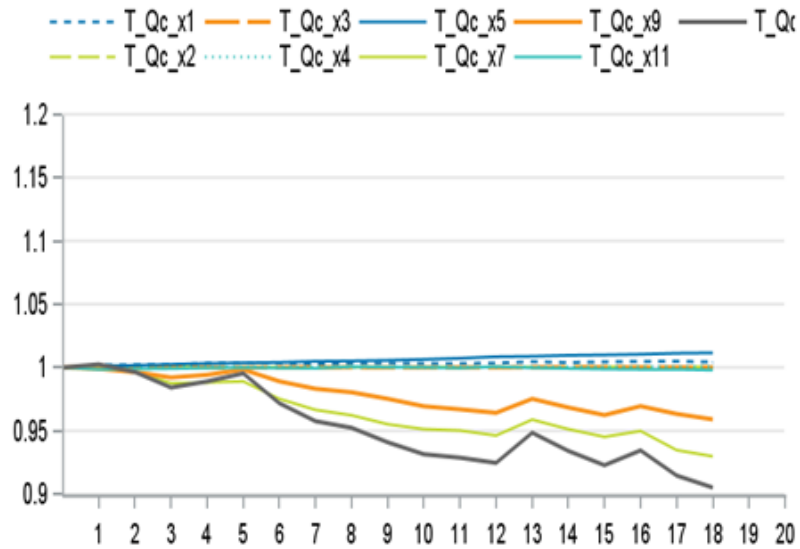


- Basic formulas are contingently biased, deviating from each other
- Price ratios using unweighted arithmetic or geometric average prices are closely related

4. Results 2/3

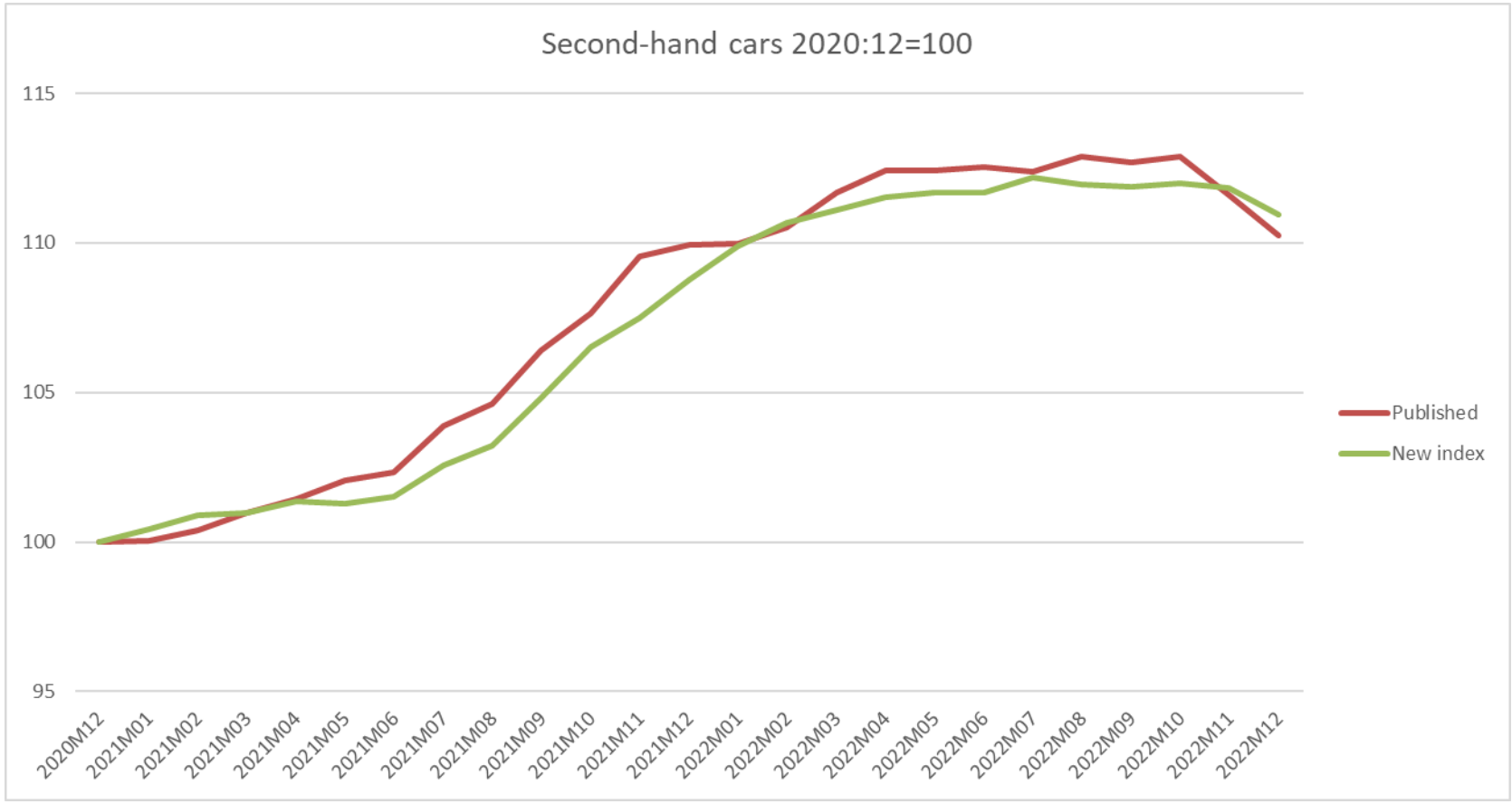
- Hedonic index series for actual arithmetic average prices (A), quality adjusted prices (QA) and quality corrections (Qc_x)





- Age of a car (x7) and mileage (x9) have a negative effect on actual average prices
  - Sold cars are older and more driven in the current period
- Index series for actual prices must be corrected upwards, which is index series for quality adjusted prices

4. Results 3/3



- The differences between the series are due to the data source, regression model variables, index formula and strategy

Things to consider when designing a hedonic application (HICP Manual)

- How many and which quality-related variables to include in the regression equation: Our model has 12 variables (slide 6)
- Whether to use another (finer or coarser) stratification when estimating the regression coefficients than when computing the index: We use a coarser stratification for estimation (slide 8)
- How frequently to re-estimate the regression coefficients: We re-estimate every year
- Whether to weight the prices when estimating the regression coefficients: We use equal weights
- Which function form to use; semi-logarithmic, double-logarithmic or other: Our model is semi-logarithmic (slide 6)
- Whether valid or spurious results are obtained: Statistical inference leads to selection of the best price models. Estimators of the price models are the best linear unbiased estimates (BLUE)
- Whether the method improves the accuracy of the index so much that it outweighs the often relatively high cost for design work and for collection of quality-related data: Yes, see slide 14

5. Conclusions

- Our proposal for producing a hedonic price index is as follows:
    1. Use suitable partition in estimation of price models
    2. Aggregate price models into stratum-level by using arithmetic average
        - Arithmetic average is more interpretable than geometric average
    1. Form price decompositions for stratums (Oaxaca)
    2. Aggregate stratum-level price decompositions into COICOP-level using Törnqvist formula and base strategy with a flexible basket, that is free of chain drift
- This method is widely used in Statistics Finland

Statistics Finland

Thank You!

Satu Montonen
satu.montonen@stat.fi