Estimating the Benefits and Costs of New and Disappearing Products

by Erwin Diewert
University of British Columbia and University of New South Wales

and Robert Feenstra

University of California at Davis

Meeting of the Group of Experts on Consumer Price Indices, UNECE, Geneva

Session 3: Methodological Issues in Compilation

May 8, 2018

Introduction

- **One of the more pressing problems facing statistical agencies and economic analysts is the new goods (and services) problem; i.e., how should the introduction of new products and the disappearance of (possibly) obsolete products be treated in the context of forming a consumer price index?**
- **Hicks (1940) suggested a general approach to this measurement problem in the context of the economic approach to index number theory.**
- **His approach was to apply normal index number theory but estimate hypothetical prices that would induce utility maximizing purchasers of a related group of products to demand 0 units of unavailable products.**
- **With these virtual (or reservation or imputed) prices in hand, one can just apply normal index number theory using the augmented price data and the observed quantity data.**
- **The practical problem facing statistical agencies is: *how exactly are these virtual prices to be estimated*? We address this problem in this paper.**

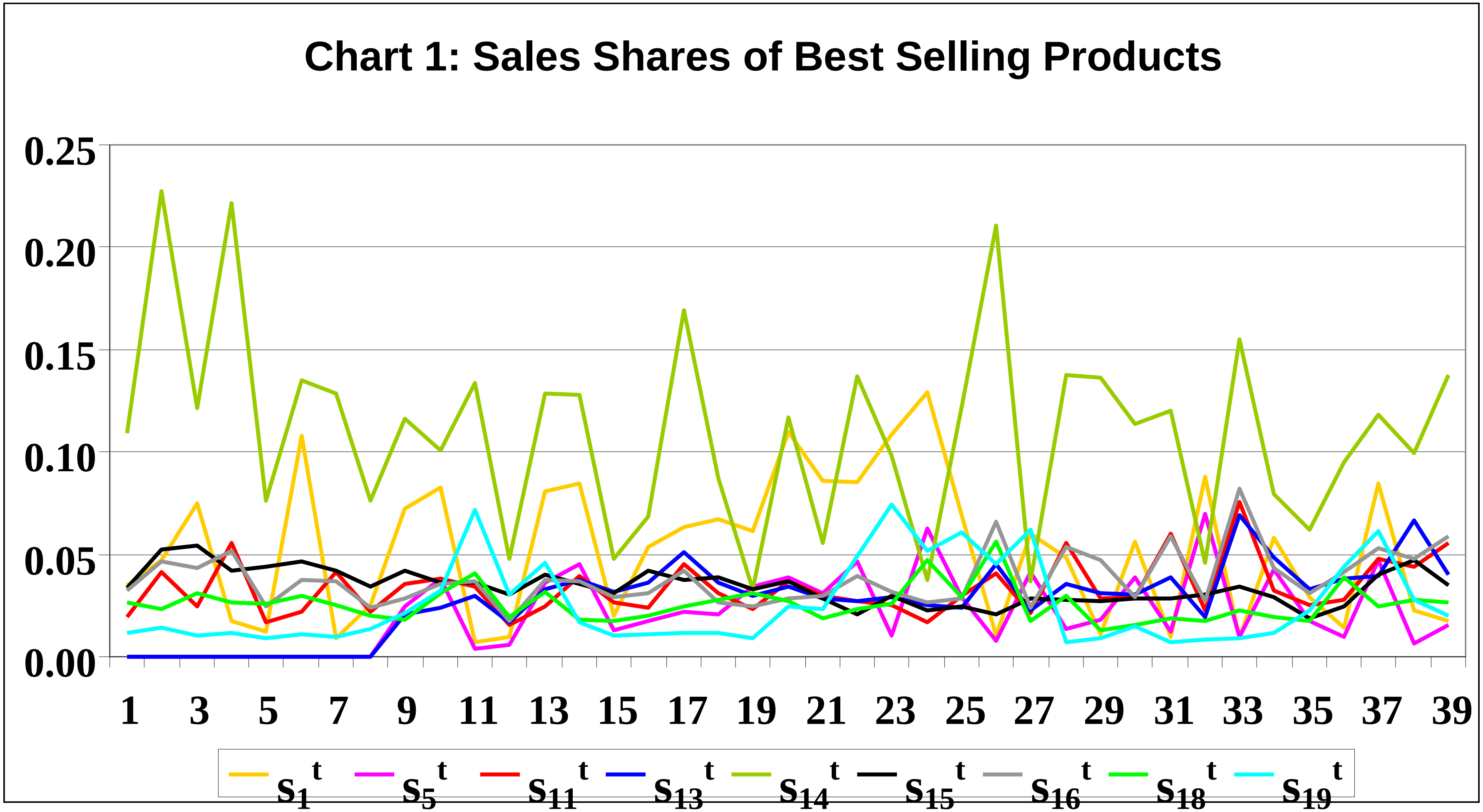
Introduction (cont)

- The two main contributors in this area are Feenstra (1994) and Hausman (1996).
- Feenstra's method largely avoids econometric estimation but is likely to overstate the benefits of new products.
- His method relies on the properties of CES utility and cost functions. It turns out that CES reservation prices for new and disappearing products are equal to plus infinity and hence, it is not necessary to estimate reservation prices using the Feenstra methodology. His method does require an estimate for the elasticity of substitution.
- Hausman's method requires the estimation of consumer expenditure functions. His method is theoretically sound but there are econometric difficulties in implementing his methodology.
- We suggest a new methodology which avoids the problems with the Feenstra and Hausman methodologies.

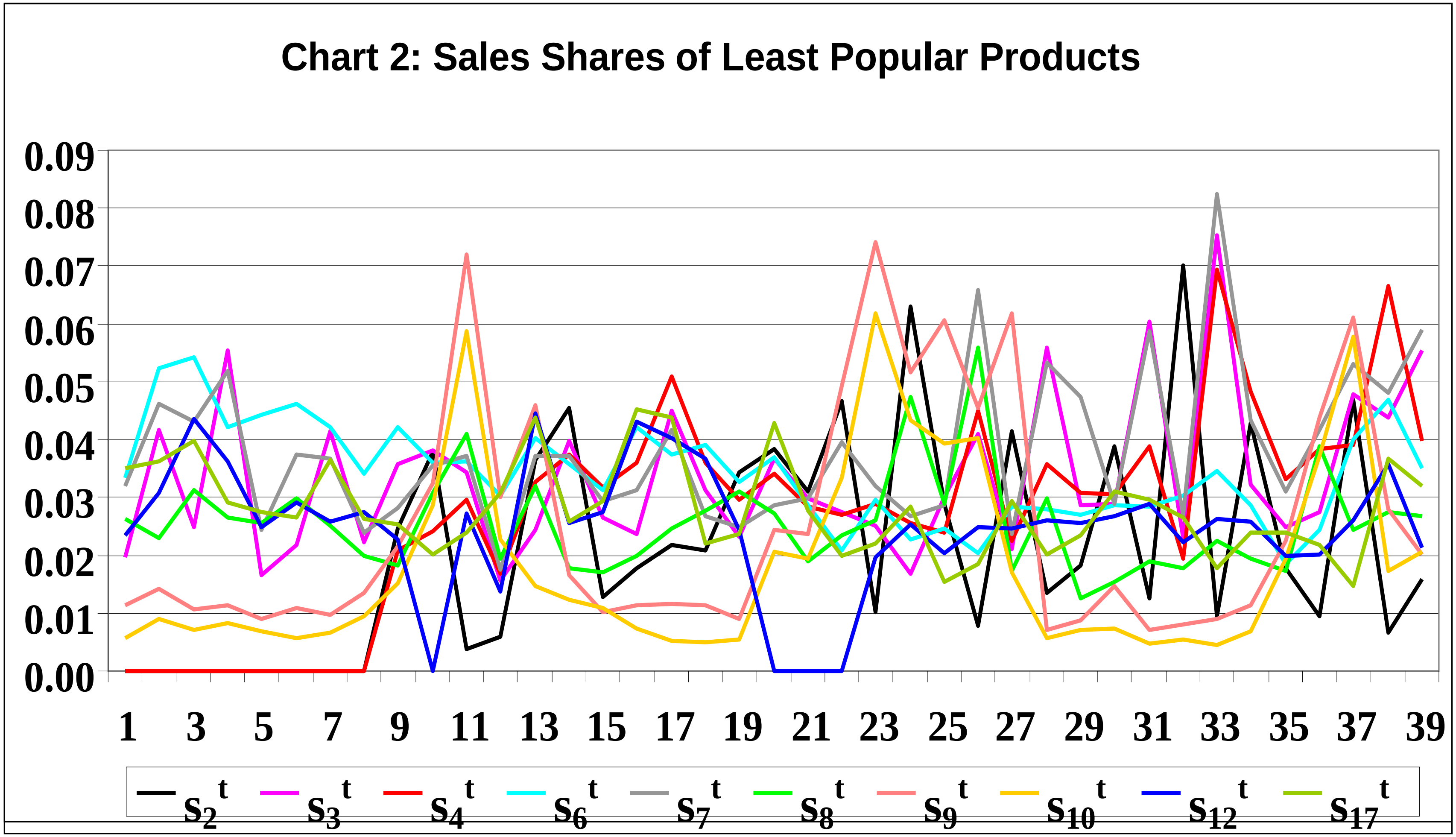
The Scanner Data Set for Sales of Frozen Juice

- In order to compare Feenstra's CES methodology with our new methodology, we used the data from Store Number 5 in the Dominick's Finer Foods Chain of 100 stores in the Greater Chicago area on 19 varieties of frozen juice for 3 years in the period 1989-1994.
- We aggregated weekly price and sales data into 39 "months", where each "month" consisted of 4 consecutive weeks of sales data so we have 39x19 = 741 price and quantity observations.
- However, products 2 and 4 had no sales for months 1-8 and product 12 was missing for month 10 and 20-22 so that there were 20 missing prices out of the 741 price-quantity observations.
- We estimated CES preferences as well as KBF preferences for this data set and compared Feenstra's methodology for new and disappearing products with our new methodology.

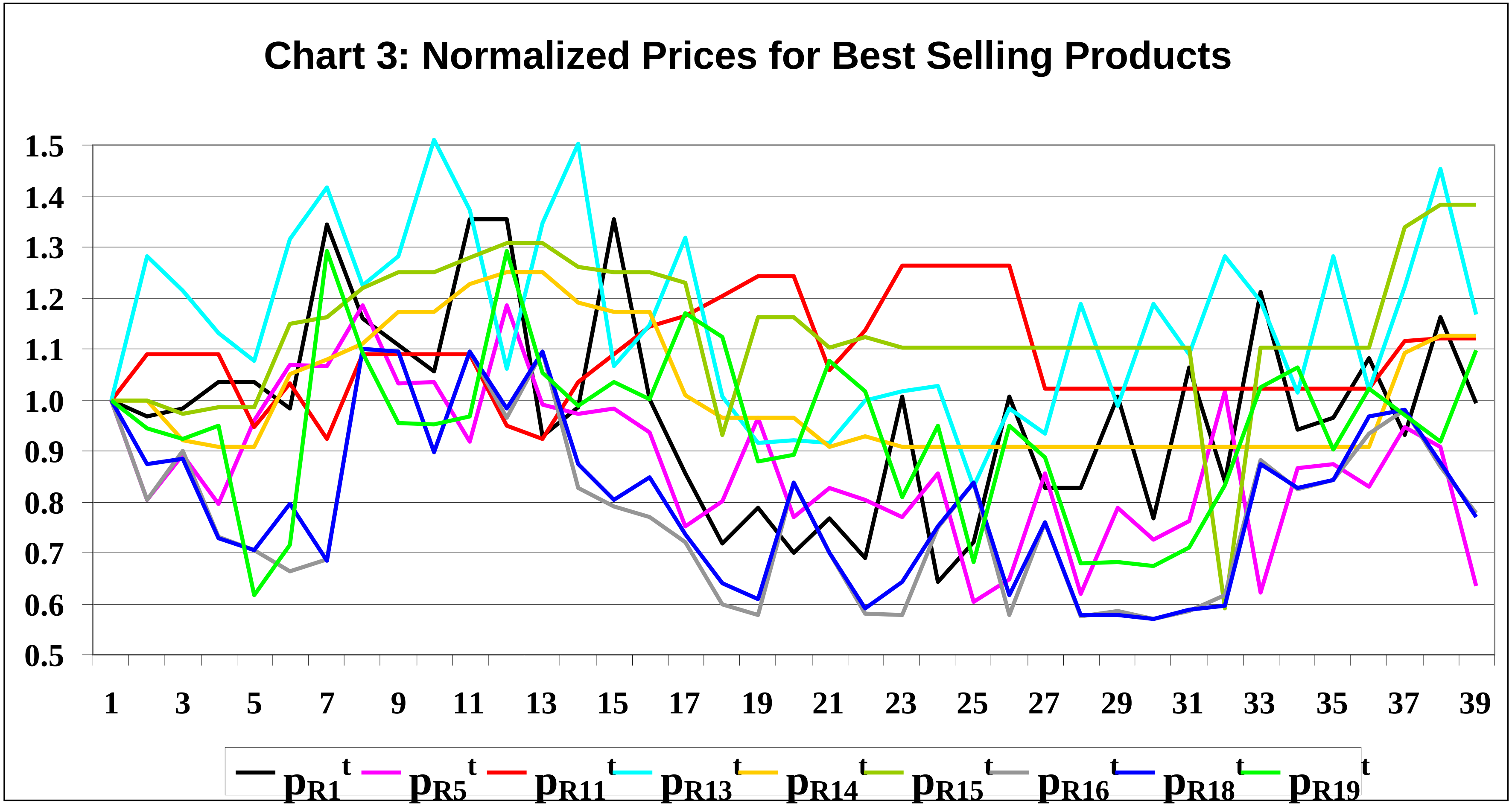
The Scanner Data Set for Sales of Frozen Juice (cont)



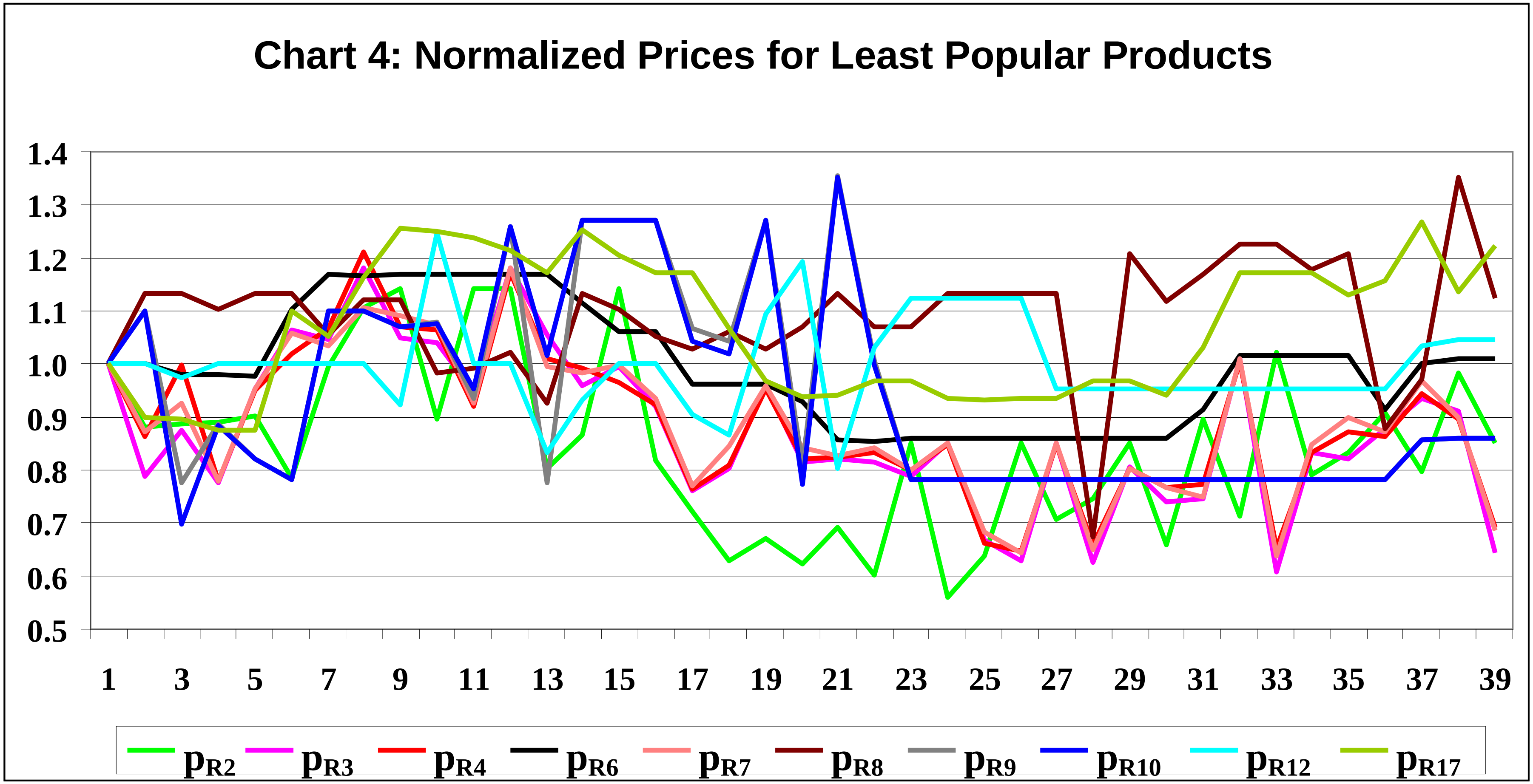
The Scanner Data Set for Sales of Frozen Juice (cont)



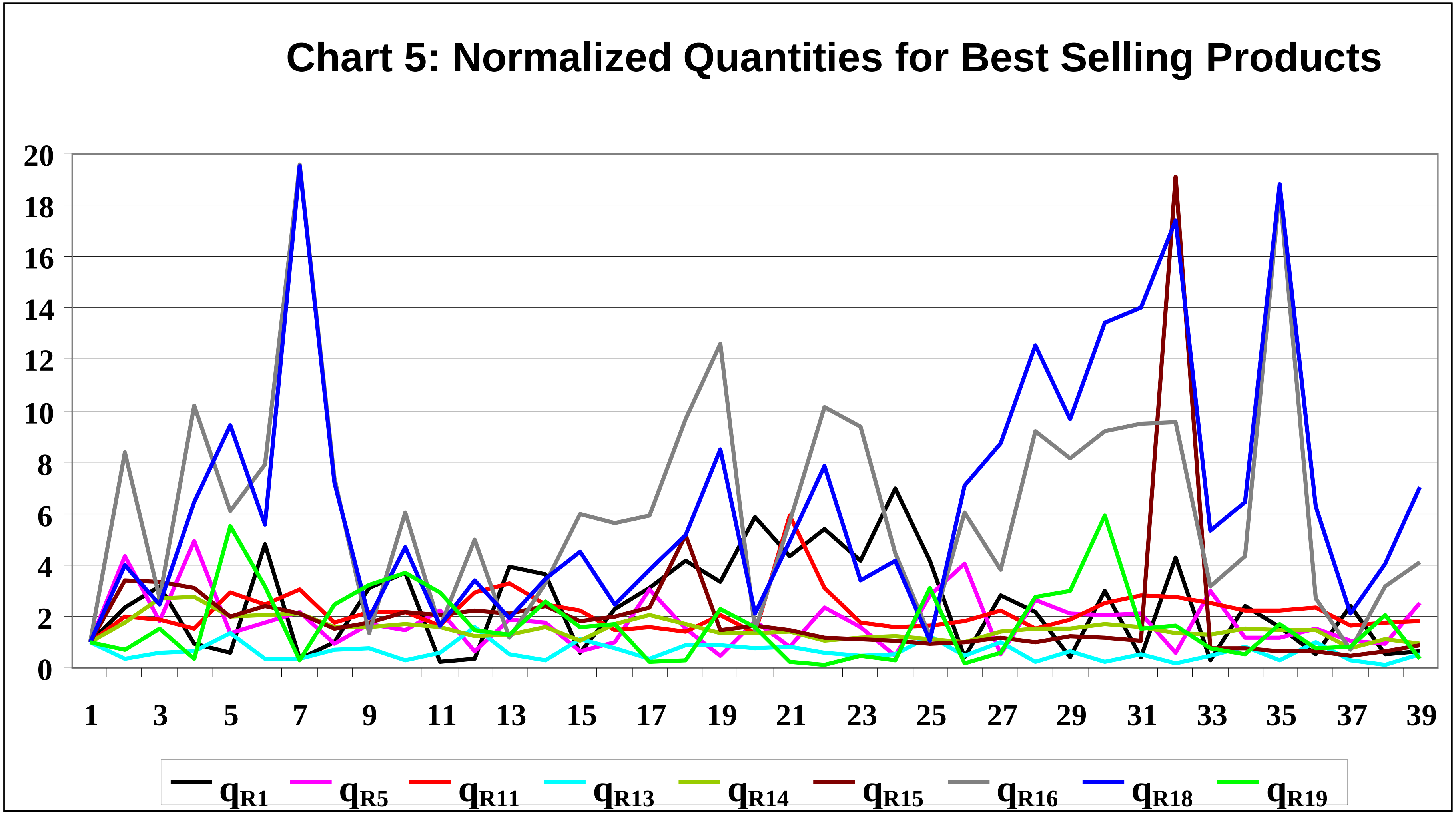
The Scanner Data Set for Sales of Frozen Juice (cont)



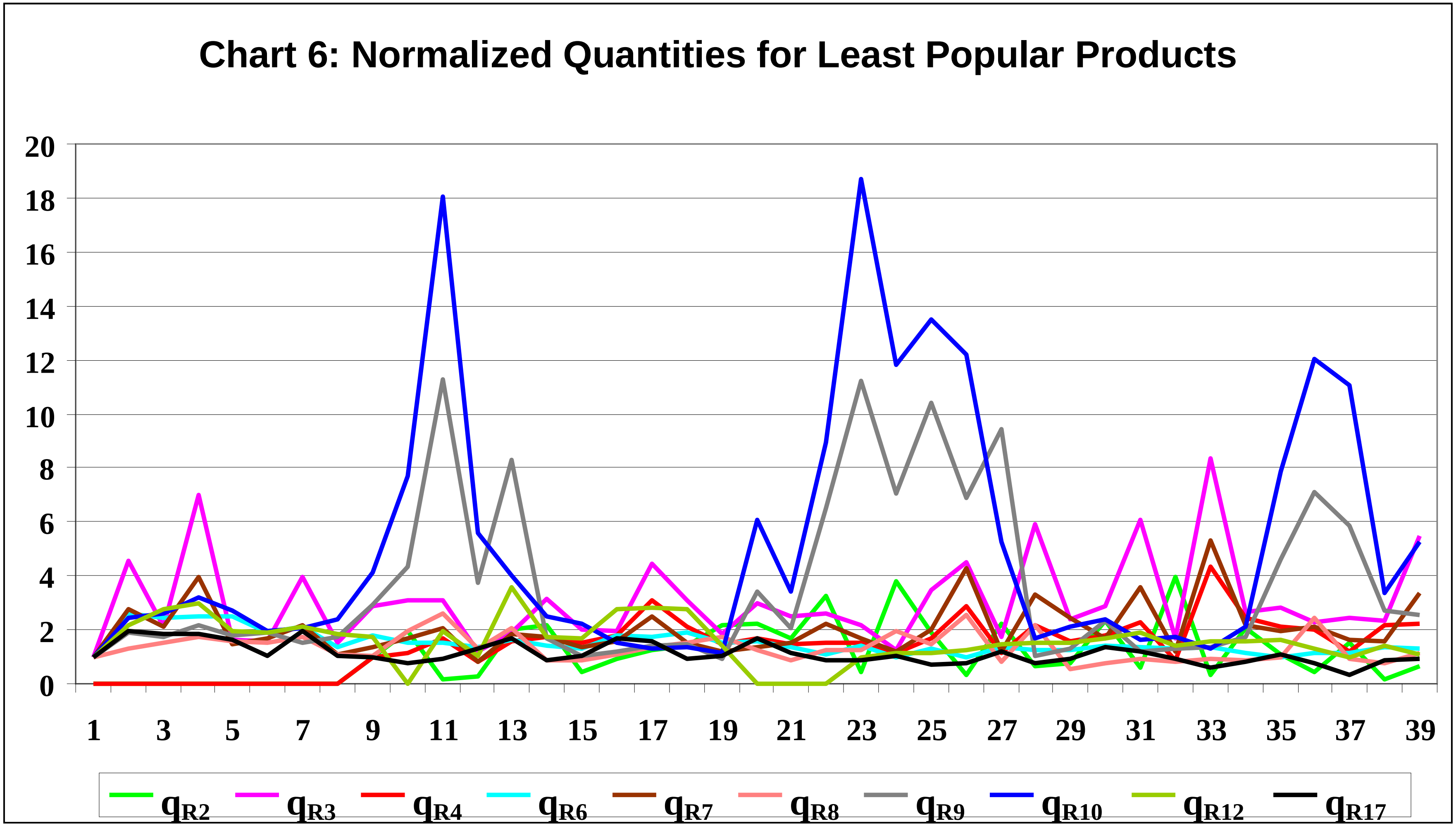
The Scanner Data Set for Sales of Frozen Juice (cont)



The Scanner Data Set for Sales of Frozen Juice (cont)



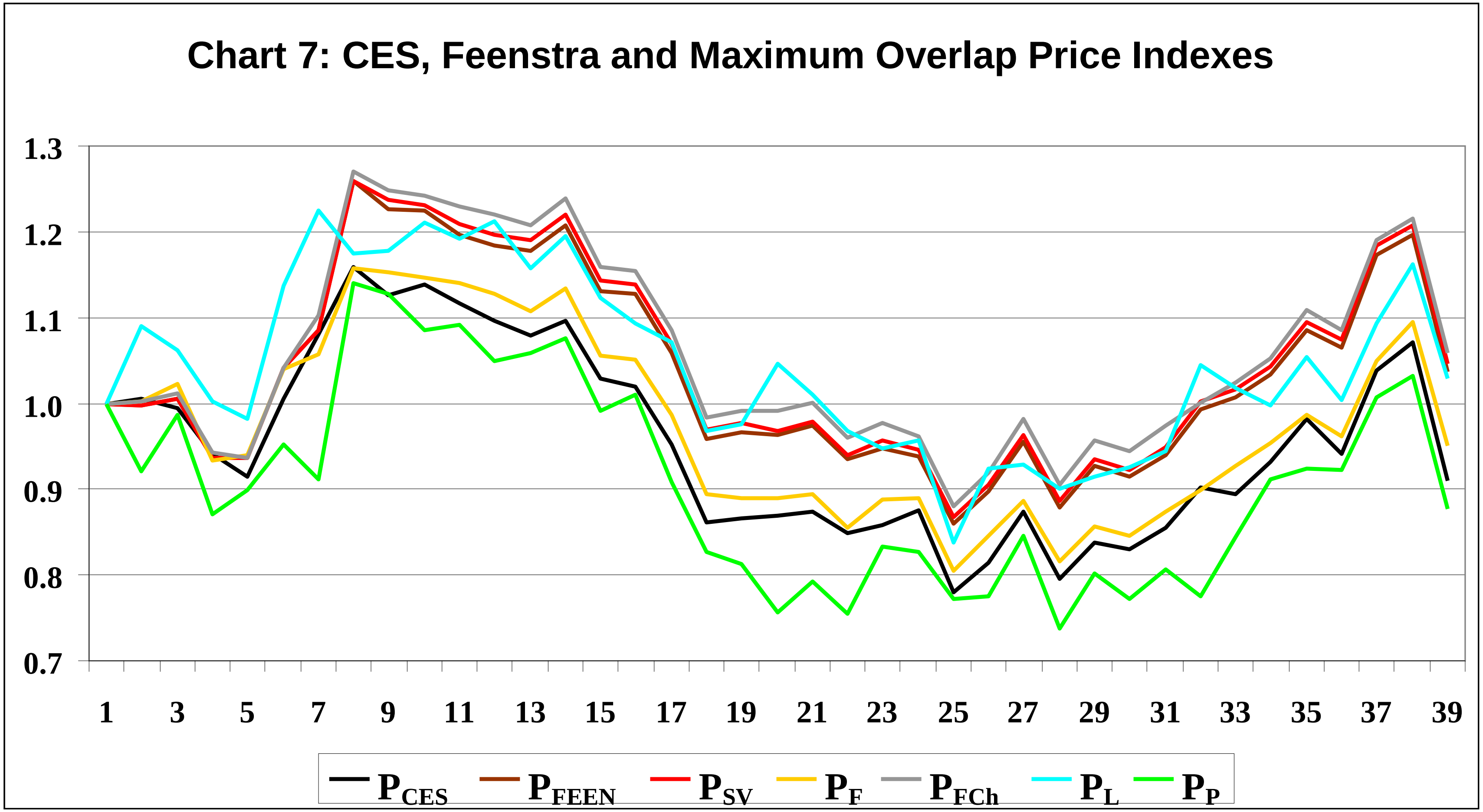
The Scanner Data Set for Sales of Frozen Juice (cont)



CES and Feenstra Price Indexes and Sato-Vartia, Fisher, Laspeyres, Paasche Fixed Base and Chained Maximum Overlap Price Indexes

- **The fixed base Paasche indexes are on the lowest curve on Chart 7. $P_{CES}{}^{t}$ is slightly below the fixed base maximum overlap (fixed base) Fisher index $P_{F}{}^{t}$ and they are the second and third curve from the bottom.**
- **The highest curves are $P_{FCh}{}^{t}$, the *chained* maximum overlap Fisher index, $P_{SV}{}^{t}$, the Sato-Vartia maximum overlap *chained* index followed by the *chained* Feenstra indexes, $P_{FEEN}{}^{t}$. These indexes suffer from some chain drift. The wide gap between the fixed base maximum overlap Laspeyres and Paasche indexes indicates that these two indexes suffer from substantial substitution bias.**
- **The most reasonable indexes are the econometrically determined CES price index, $P_{CES}{}^{t}$, and the fixed base maximum overlap Fisher index, $P_{F}{}^{t}$. (The black and gold lines on the chart which follows).**

CES and Feenstra Price Indexes and Sato-Vartia, Fisher, Laspeyres, Paasche Fixed Base and Chained Maximum Overlap Price Indexes



The Konüs-Byushgens-Fisher Utility Function

- The functional form for a purchaser's utility function $f(q)$ that we will introduce in this section is the following one:

(75) $f(q) = (q^T A q)^{1/2}$

- where the N by N matrix $A \equiv [a_{nk}]$ is symmetric (so that $A^T = A$) and thus has $N(N+1)/2$ unknown a_{nk} elements.
- We also assume that A has one positive eigenvalue with a corresponding strictly positive eigenvector and the remaining $N-1$ eigenvalues are negative or zero.
- Konüs and Byushgens (1926) showed that the Fisher (1922) quantity index $Q_F(p^0,p^1,q^0,q^1) \equiv [p^0 \cdot q^1 p^1 \cdot q^1 / p^0 \cdot q^0 p^1 \cdot q^0]^{1/2}$ is exactly equal to the aggregate utility ratio $f(q^1)/f(q^0)$ provided that all purchasers maximized the utility function defined by (75) in periods 0 and 1 where p^0 and p^1 are the price vectors prevailing during periods 0 and 1 and aggregate purchases in periods 0 and 1 are equal to q^0 and q^1.
- The KBF functional form has 2 advantages over the translog functional form: (i) the translog utility function is not well defined when a quantity is 0 and (ii) it is possible to impose the correct curvature conditions on the KBF functional form without destroying its flexibility.

The KBF Functional Form (cont)

- **The KBF functional form also has an advantage over the normalized quadratic functional form, $f(q) \equiv q^TAq/q^T\alpha$, which requires the econometrician to specify (or estimate) the α vector of parameters.**
- **Thus we will concentrate on alternative specifications for the estimation of the KBF functional form.**
- **The unit cost function which is dual to the KBF utility function defined by (75) is:**

(1) $c(p) \equiv \min_q\{f(q) \geq 1; q \geq 0_N\}$.

- **The first order necessary (and sufficient) conditions that can be used to solve the unit cost minimization problem defined by (1) when the utility function f is defined by (75) are the following conditions:**

(76) $p = \lambda Aq/(q^TAq)^{1/2}$;

(77) $1 = (q^TAq)^{1/2}$.

The KBF Functional Form (cont)

- **Multiply both sides of equation n in (76) by q_n and sum the resulting N equations.**
- **This leads to the equation $p \cdot q = \lambda(q^TAq)^{1/2}$. Solve this equation for λ and use this solution to eliminate the λ in equations (76).**
- **The resulting equations (where equation n is multiplied by q_n) are the following *system of inverse demand share equations*:**

(78) $s_n \equiv p_nq_n/p \cdot q = q_n \Sigma_{k=1}^{N} a_{nk}q_j/q^TAq$; $\quad n = 1,\ldots,N$

- **where a_{nk} is the element of A that is in row n and column j for n, k = 1,...,N.**
- **Equations (78) can be used as a system of nonlinear estimating equations. In the following 3 slides, we show how the parameters in the A matrix can be written in a different form that allows us to impose the concavity property on f(q).**

The KBF Functional Form (cont)

- **We set A equal to the following expression:**

(79) $A = bb^T + B$; $b >> 0_N$; $B = B^T$; B is negative semidefinite;

$Bq^* = 0_N$.

- **The vector $b^T \equiv [b_1,\ldots,b_N]$ is a row vector of positive constants and so bb^T is a rank one positive semidefinite N by N matrix.**
- **The symmetric matrix B has N(N+1)/2 independent elements b_{nk} but the N constraints Bq^* reduce this number of independent parameters by N.**
- **Thus there are N independent parameters in the b vector and N(N–1)/2 independent parameters in the B matrix so that $bb^T + B$ has the same number of independent parameters as the A matrix.**
- **Diewert and Hill (2010) showed that replacing A by $bb^T + B$ still leads to a flexible functional form.**

The KBF Functional Form (cont)

- **The reparameterization of A by bb^T + B is useful in our present context because we can use this reparameterization to estimate the unknown parameters in stages.**
- **Thus we will initially set B = $O_{N\times N}$, a matrix of 0's. The resulting utility function becomes $f(q) = (q^Tbb^Tq)^{1/2} = (b^Tqb^Tq)^{1/2} = b^Tq$, a linear utility function.**
- **The matrix B is required to be negative semidefinite.**
- **We can follow the procedure used by Wiley, Schmidt and Bramble (1973) and Diewert and Wales (1987) and impose negative semidefiniteness on B by setting B equal to $-CC^T$ where C is a lower triangular matrix.**
- **Write C as $[c^1,c^2,\ldots,c^N]$ where c^k is a column vector for k = 1,...,K.**
- **If C is lower triangular, then the first k−1 elements of c^k are equal to 0 for k = 2,3,...,N.**

The KBF Functional Form (cont)

- **Thus we have the following representation for B:**

(80) $B = -CC^T$

$$= -\Sigma_{n=1}^{N} c^n c^{nT}$$

- **where we impose the following restrictions on the vectors c^n in order to impose the restrictions $Bq^* = 0_N$ on B:**

(81) $c^n \cdot q^* = c^{nT} q^* = 0$; $\quad n = 1,\ldots.,N.$

- **In the first stage, we estimate the linear utility function $f(q) = b^T q$. In the second stage, we estimate $f(q) = (q^T[bb^T - c^1c^{1T}]q)^{1/2}$ where $c^{1T} \equiv [c_1{}^1, c_2{}^1, \ldots, c_N{}^1]$ and $c^{1T}q^* = 0$.**
- **For starting coefficient values in the second nonlinear regression, we use the final estimates for b from the first nonlinear regression and set the starting $c^1 \equiv 0_N$.**
- **We continue adding c^i columns as long as the log likelihood for the system of estimating equations increases significantly.**

The Systems Approach to the Estimation of KBF Preferences

- **Our system of nonlinear estimating equations for *Model 5* is the following stochastic version of equations (75) above where $A = bb^T - c^1c^{1T}$:**

(82) $s_i^t = q_i^t \sum_{k=1}^{19} a_{ik}q_k^t/[\sum_{n=1}^{19}\sum_{m=1}^{19} a_{nm}q_n^tq_m^t] + \varepsilon_i^t$

$t = 1,\ldots,39;\ i = 1,\ldots,19$

- **where $b^T = [b_1,\ldots,b_{19}]$, $c^{1T} = [c_1^1,\ldots,c_{19}^1]$ and the error term vectors, $\varepsilon^{tT} = [\varepsilon_1^t,\ldots,\varepsilon_{19}^t]$ are assumed to be distributed as a multivariate normal random variable with mean vector 0_{19} and variance-covariance matrix Σ for $t = 1,\ldots,39$.**
- **There were no problems in estimating this rank 2 A matrix model using the econometric estimation package Shazam.**
- **The equation by equation R^2 values were as follows: 0.9661, 0.9787, 0.9623, 0.9889, 0.9608, 0.9521, 0.9628, 0.8002, 0.9657, 0.9752, 0.8337, 0.9224, 0.9867, 0.8936, 0.9673, 0.9555, 0.9064 and 0.9599.**
- **Note that this framework can handle 0 quantity observations perfectly well; if a quantity is 0, the corresponding expenditure share is also 0.**

The Systems Approach to the Estimation of KBF Preferences (cont)

- **Our system of nonlinear estimating equations for *Model 6* are equations (82) where $A = bb^T - c^1c^{1T} - c^2c^{2T}$ with $c^{2T} = [0,c_2^2,\ldots,c_{19}^2]$ and the normalizations $b_{19} = 1$, $c_{19}^1 = -\sum_{n=1}^{18} c_n^1$ and $c_{19}^2 = -\sum_{n=2}^{18} c_n^2$.**
- **Thus there are 18 + 18 + 17 unknown parameters to estimate in the A matrix.**
- **However, the nonlinear maximum likelihood estimation package in Shazam did not converge for this model. The problem is that the error specification that is used in the system command for the Nonlinear estimation option in Shazam also estimates the elements of the variance covariance matrix Σ.**
- **Thus for our Model 6, it is necessary to estimate the 53 unknown parameters in the A matrix plus 19×18/2 = 171 unknown variances and covariances. Too hard to do!**

The Single Equation Approach to the Estimation of KBF Preferences using Share Equations

- **In order to deal with the nonconvergence problem, we decided to stack up our 18 estimating equations into one big nonlinear regression model which involves estimating a single variance parameter instead of the 171 parameters required for the systems approach.**
- **We stopped adding columns to the lower triangular C matrix at 4 columns, which was our *Model 11*.**
- **For this model, we have** $A = bb^T - c^1c^{1T} - c^2c^{2T} - c^3c^{3T} - c^4c^{4T}$ **with** $c^{4T} = [0,0,0,c_4^4,\ldots,c_{19}^4]$ **and the additional normalization** $c_{19}^4 = -\sum_{n=4}^{18} c_n^4$.
- **The final log likelihood for this model was 2629.182, an increase of 14.656 for adding 15 new parameters to the Model 10 parameters.**
- **Thus the increase in log likelihood is now less than one per additional parameter so we decided to stop adding columns at this point.**
- **The single equation** R^2 **increased to 0.9922.**

KBF Utility Function: the One Big Equation Approach (cont)

- **However, this single equation R^2 is not comparable to the equation by equation R^2 that we obtained using the systems approach in the previous section.**
- **The comparable R^2 for each separate product share equation are as follows: 0.9859, 0.9930, 0.9773, 0.9853, 0.9814, 0.9543, 0.9755, 0.8581, 0.9760, 0.9694, 0.8923, 0.9278, 0.9908, 0.9202, 0.9874, 0.9566, 0.9111 and 0.9653.**
- **The average R^2 is 0.9560 which is a relatively high average when estimating share equations.**
- **Once we have an estimated A matrix, it is straightforward to form the reservation prices for the 20 observations where we have missing products.**
- **We explain the relevant algebra on the next slide.**
- **Using the one big equation approach, we dropped the observations where the prices were missing.**

KBF Utility Function: the One Big Equation Approach (cont)

- With the estimated b and c vectors in hand (denote them as b^* and c^{k*} for k = 1,2,3,4), form the estimated A matrix as follows:

(76) $A^* \equiv b^* b^{*T} - c^{1*}c^{1*T} - c^{2*}c^{2*T} - c^{3*}c^{3*T} - c^{4*}c^{4*T}$

- and denote the ij element of A^* as $a_{ij}{}^*$ for i,j = 1,...,19. The *predicted expenditure share* for product i in month t is $s_i{}^{t*}$ defined as follows:

(77) $s_i{}^{t*} \equiv q_i{}^t \Sigma_{k=1}{}^{19} a_{ik}{}^* q_k{}^t / [\Sigma_{n=1}{}^{19} \Sigma_{m=1}{}^{19} a_{nm}{}^* q_n{}^t q_m{}^t]$.

- The *predicted price* for product i in month t is defined as follows:

(78) $p_i{}^{t*} \equiv e^t \Sigma_{k=1}{}^{19} a_{ik}{}^* q_k{}^t / [\Sigma_{n=1}{}^{19} \Sigma_{m=1}{}^{19} a_{nm}{}^* q_n{}^t q_m{}^t]$

- The predicted prices for products 2 and 4 for the first 8 months in our sample period were 1.62, 1.56, 1.60, 1.52, 1.61, 1.52, 1.70. 1.97 and 1.85, 1.46, 1.80, 1.37, 1.77, 1.83, 1.88, 2.27 respectively.
- The predicted prices for product 12 for months 10 and 20-22 were 1.37, 1.20, 1.22 and 1.28. These prices are not infinite!

KBF Utility Function: the One Big Equation Approach (cont)

- **Problem: the predicted prices were not particularly close to the actual prices!**
- **Thus the equation by equation R^2 for the 19 products were as follows: 0.7571, 0.8209, 0.8657, 0.8969, 0.9025, 0.7578, 0.8660, 0.0019, 0.2517, 0.1222, 0.0000, 0.0013, 0.9125, 0.6724, 0.4609, 0.7235, 0.5427, 0.8148 and 0.4226.**
- **The average R^2 is only 0.5681 which is not very satisfactory.**
- **How can the R^2 for the share equations be so high while the corresponding R^2 for the fitted prices are so low?**
- **The answer appears to be the following one: when a price is unusually low, the corresponding quantity is unusually high and vice versa.**
- **Thus the errors in the fitted price equations and the corresponding fitted quantity equations tend to offset each other and so the fitted share equations are fairly close to the actual shares.**

KBF Utility Function: the One Big Equation Approach II

- **Our interest is not in predicting shares; our interest is in finding predicted prices for the observations when quantities are equal to 0.**
- **Thus we turned to another econometric specification of the KBF utility function model where prices replaced shares as the dependent variables in the One Big Regression Approach.**
- **Thus the new estimating equations become:**

$$(79)\ p_i^t \equiv e^t\, \Sigma_{k=1}^{19}\, a_{ik}^* q_k^t / [\Sigma_{n=1}^{19} \Sigma_{m=1}^{19}\, a_{nm}^* q_n^t q_m^t] + \varepsilon_i^t\ ; \qquad t = 1,\ldots,39;\ i = 1,\ldots,18.$$

- **Again, the observations that correspond to missing products are dropped from the stacked estimating equations defined by (79). This is an advantage of the one Big Regression Approach.**
- **As before, $A = bb^T - CC^T$ where C is a lower triangular matrix.**
- **With the new dependent variables, we were able to estimate a rank 6 substitution matrix (which is the matrix $-CC^T$).**

KBF Utility Function: the One Big Equation Approach II (cont)

- The final log likelihood for the rank 6 substitution matrix *Model 14* was 568.877, an increase of 18.531 over the rank 5 substitution Model 13. The single equation R^2 was 0.9527.
- Model 14 had 111 unknown parameters that were estimated (plus a variance parameter). We had only 680 observations and so we decided to call a halt to our estimation procedure.
- The equation by equation R^2 that compares the predicted prices for the 19 products with the actual prices were as follows: 0.8274, 0.8678, 0.9001, 0.9174, 0.8955, 0.8536, 0.9047, 0.0344, 0.3281, 0.4242, 0.0516, 0.2842, 0.8650, 0.7280, 0.4872, 0.8135, 0.8542, 0.8479 and 0.3210.
- The average R^2 for Model 14 was 0.6424. For Model 11, it was 0.5681 so by switching from shares as the dependent variables to prices as the dependent variables, we have improved the accuracy of our estimated predicted prices.

KBF Utility Function: the One Big Equation Approach II (cont)

- **The *month t utility level* or aggregate quantity level implied by the KBF model, $Q_{KBF}{}^{t}$, is defined as follows:**

(81) $Q_{KBF}{}^{t} \equiv (q^{tT}A^{*}q^{t})^{1/2}$; t = 1,...,39.

- **The corresponding *KBF (unnormalized) implicit price level*, $P_{KBF}{}^{t*}$, is defined as period t sales of the 19 products, e^{t}, divided by the period t aggregate KBF quantity level, $Q_{KBF}{}^{t}$:**

(82) $P_{KBF}{}^{t*} \equiv e^{t}/Q_{KBF}{}^{t}$; t = 1,...,39.

- **The month t *KBF price index*, $P_{KBF}{}^{t}$, is defined as the month t KBF price level divided by the month 1 KBF price level; i.e.,**

  $P_{KBF}{}^{t} \equiv P_{KBF}{}^{t*}/P_{KBF}{}^{1*}$ for t = 1,...,39.

- **These econometrically based KBF price indexes can be compared to our econometrically based CES price indexes $P_{UCES}{}^{t}$ that are defined in a similar manner using the results of Model 4, which estimated a direct CES utility function.**
- **However, before we make this comparison, we estimate one more model.**

CES Utility Function; One Big Equation; Prices as Dependent Variables (instead of shares)

- This leads to the following system of estimating equations:

(83) $p_i^t = [e^t/q_i^t][\beta_i (q_i^t)^s/\sum_{n=1}^{19} \beta_n (q_n^t)^s] + \varepsilon_i^t$; $t = 1,\ldots,39; i = 1,\ldots,18.$

- Now stack the above 702 equations into a single estimating equation and drop the 20 observations where $q_i^t = 0$. Call this ***Model 15***.
- The final log likelihood was equal to **483.834**, which is below the final LL from the KBF Model 14 which was **568.877**.
- The single equation R^2 was 0.9393, which is below the single equation R^2 from Model 14, which was 0.9527.
- The estimated parameter s was $s^* = 0.85365$. This is virtually identical to our estimate for s from **Model 4** (which used the systems approach to CES utility function estimation with shares as dependent variables) which was **0.85374**. ($\sigma = 1/(1-s^*) = 6.8$)
- Since the estimated s for Model 15 is the same as it was for Model 4, **the Feenstra gains and losses from changes in product availability will not change**.

CES Utility Function; One Big Equation; Prices as Dependent Variables (cont)

- **The *month t utility level* or aggregate quantity level implied by the New Single equation CES Model 15, $Q_{CESN}{}^{t}$, is defined as follows:**

(84) $Q_{CESN}{}^{t} \equiv [\sum_{n=1}^{19} \beta_{n}{}^{*} (q_{n}{}^{t})^{s^*}]^{1/s^*}$; $\qquad t = 1,\ldots,39.$

- **The corresponding *New CES (unnormalized) implicit price level*, $P_{CESN}{}^{t*}$, is defined as period t sales of the 19 products, e^{t}, divided by the period t aggregate quantity level, $Q_{CESN}{}^{t}$:**

(85) $P_{CESN}{}^{t*} \equiv e^{t}/Q_{CESN}{}^{t}$; $\qquad t = 1,\ldots,39.$

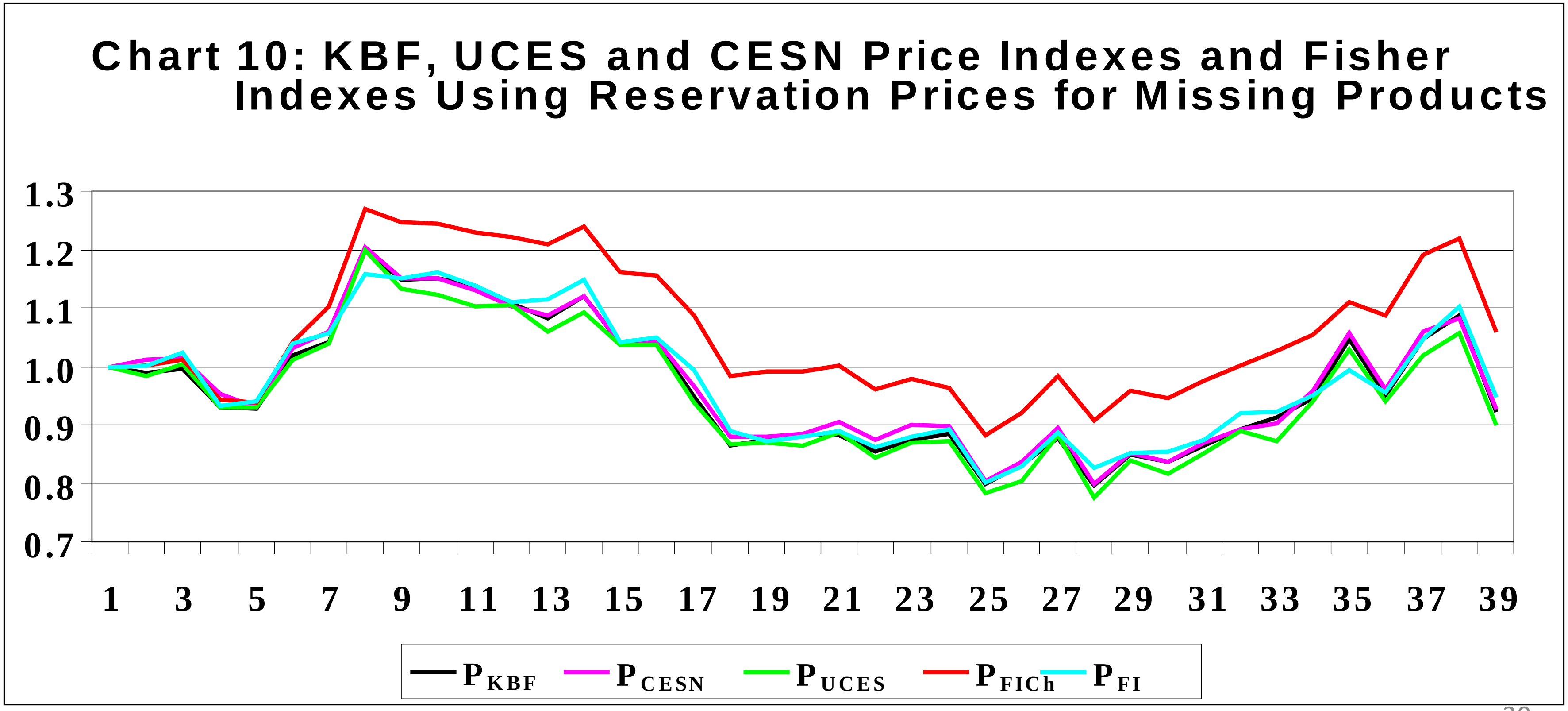
- **The month t *New CES price index*, $P_{CESN}{}^{t}$, is defined as the month t CESN price level divided by the month 1 CESN price level; i.e.,**

$$P_{CESN}{}^{t} \equiv P_{CESN}{}^{t*}/P_{CESN}{}^{1*} \text{ for } t = 1,\ldots,39.$$

- **The CESN price index will be compared to its econometric counterpart indexes $P_{KBF}{}^{t}$ (Model 14) and $P_{UESN}{}^{t}$ (Model 4).**

Comparison of 3 Econometric Based Price Indexes with Chained and Fixed Base Fisher Indexes that use Reservation Prices

- $P_{KBF}{}^{t}$ and $P_{CESN}{}^{t}$ are very close to each other. The Model 4 CES price index, $P_{UESN}{}^{t}$, is pretty close to the Model 14 and 15 indexes. The fixed base Fisher index that uses reservation prices for the missing products is fairly close as well. The chained Fisher index that uses reservation prices has upward chain drift.



KBF and CES Gains from Changes in Product Availability:

Table 10: Gains and Losses of Utility that can be Attributed to Changes In Product Availability Holding Expenditure Constant

	KBF	CES
$G_{A2,4}^{9}$	1.00127	1.00746
L_{A12}^{10}	0.99748	0.99512
G_{A12}^{11}	1.00304	1.00529
L_{A12}^{20}	0.99881	0.99644
G_{A12}^{23}	1.00078	1.00296
Product	1.00138	1.00724

- **Since there is a net gain in product availability over the sample period, both estimated utility functions register a net gain.**
- **But the net gain from the KBF utility function is only about 1/5 of the gain that accrued to the CES utility function using Approach 3. The CES approach consistently overestimates the gains from increased product availability!**

Conclusion: The Important Points to Take Away!

- **When dealing with scanner data where there are periodic sales of products, chain drift is a huge problem.**
- **Multilateral index number theory can be used to deal with the chain drift problem; see the ABS (2016) and Diewert and Fox (2017)**
- **It is not a trivial matter to estimate the elasticity of substitution in the CES context. Estimation of the CES unit cost function may give very different results from estimation of the CES direct utility function.**
- **The CES methodology developed by Feenstra for measuring the gains from increased product availability appears to overestimate the gains by a substantial amount.**
- **The KBF utility function can be estimated and it can be used to calculate "reasonable" reservation prices but it is too labour intensive (and subject to many econometric uncertainties) to be adopted by statistical agencies as a practical approach to the estimation of reservation prices.**