Estimating Components of Variance of Price Change from a Scanner-Based Sample

Sylvia G. Leaver and William E. Larson
U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 3655, Washington, D.C. 20212
Leaver_S@bls.gov

Key Words: Scanner data, Weighted REML, Price Index

Opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

In this paper we present estimates of components of variance of price change for cereal for several publication areas. Components of variance for 1-, 6-, and 12-month lags were computed using a weighted restricted maximum likelihood estimation method. Estimates are contrasted among publication areas using two different random effects models, and findings are discussed with respect to approaches to sample design.

In section one the official CPI and scanner-based geometric price index estimators are described. Section two presents the random effects model fitted to our data and the construction of components of variance estimators for the scanner index series. Section three presents computational results and compares estimates of the components over time and across cities. Sources of price change variability are identified and discussed. Conclusions are given in section four.

1. Publication and Scanner-based Indexes

The CPI is calculated monthly for the total US metropolitan and urban non-metropolitan population for all consumer items, and it is also estimated at other levels defined by geographic area and item groups such as cereal, women's suits, and tobacco products.

The CPI is estimated for items grouped into 211 strata for each index area, although not all such indexes are published every month. It is constructed in two stages. In the first or elementary level stage, the price index for an item-area is updated every 1 or 2 months via a function of sample price changes called a price relative. Let X_{ia}^{t} denote the index at time t, in item stratum i, area a, relative to time period *0*. Then

$X_{ia}^{t} = \quad R_{ia}^{t,t-1} X_{ia}^{t-1}$ where $R_{ia}^{t,t-1}$ denotes the price relative between times t and t-1 . Since 1999, elementary indexes for most commodities and services, including cereal, have been computed using a weighted geometric average (BLS, 1997):

$$R_{ia}^{t,t-1} = \prod_{j \in S_{ia}} \left(\frac{P_{iaj,t}}{P_{iaj,t-1}} \right)^{w'_{iaj}} = e^{\sum_{j \in S_{ia}} w'_{iaj} \ln\left(P_{iaj,t} \middle/ P_{iaj,t-1} \right)} ;$$

Here S_{ia} represents the sample for item i in area a, P represents the price and w' represents the quote-level sampling weight of sample item j, normalized to the same sample rotation base for all quotes in the item-index area.

Indexes for higher level item I and area A groupings are computed as a Laspeyres-type weighted sum of elementary indexes:

$$X_{IA}^{t} = \sum_{i \in \mathrm{I}} \; \sum_{a \in \mathrm{A}} r_{ia}^{b} X_{ia}^{t} \text{, where}$$

r_{ia}^{b} is the item-area relative importance or relative expenditure share, computed from the Consumer Expenditure Survey for reference period *b*.

Earlier studies of components of variance of the CPI have utilized CPI price and housing survey data. The principal objective of these studies was to estimate components of variance whose sum in turn would be minimized by means of an optimal allocation of sample resources in the survey design. Baskin (1992, 1993) estimated components of price change for the CPI shelter index using hierarchical Bayes and Gibbs sampling techniques. Baskin and Johnson (1995) and Shoemaker (2001 and 2002) estimated components of variance of rent and commodities and services price change, where, for commodities and services, data were grouped into larger design group classes comprising several item strata each. This paper builds on the REML approach of Shoemaker, but uses instead a large sample of scanner data for just one item stratum, cereal.

BLS has purchased from A.C. Nielsen Corporation scanner data for cereal from their sample of retail establishments markets corresponding to 36 of 37 index areas for which BLS publishes an index. The data examined in this study cover all cereal sales, coded at the Universal Product Code (UPC) level, in the Nielsen sample recorded for January 1999 through December 2001.

The Nielsen sample is stratified by major chain, and sample stores within chains were selected using a Peano key equal probability selection scheme (Garrett and Harter, 1995). The average sampling rate within a chain is approximately one in ten, but this rate varies by chain. Weekly per unit prices and quantity data are recorded for each UPC in each sample store in which sales occur. Quantity values for each sample store were inflated by a projection factor, equal to the ratio of the sampled stores' total sales to chain-level total sales for the same week. Total cereal sales estimates were computed by multiplying reported per unit prices by projected quantities.

Using these data, one can construct a series of scanner-based price indexes for cereal for each CPI index area. This study examines the variance structure for individual store-UPC-level price change, using chain and cereal type classification variables.

The estimator considered in this study is a geometrically averaged scanner price relative for k-month change:

$$R_a^{t,t-1} = \prod_{c=1}^{n_a} \prod_{r=1}^{n_{ac}} \prod_{\substack{q \in cr \\ c \in a}} I_q^{t,t-1} \left(\frac{p_{acrq}^{t}}{p_{acrq}^{t-k}} \right)^{S_{acrq}}$$

$$= e^{\sum_{c=1}^{n_a} \sum_{r=1}^{n_{ac}} \sum_{q \in cr} S_{acrq} \ln \left(p_{acrq}^{t} \big/ p_{acrq}^{t-k} \right)}$$

, where product q is an item corresponding to a unique UPC or two or more UPCs judged to be sufficiently similar to combine, and S_q is the expenditure share of q, i.e., the ratio of the previous year's expenditure for q to the sum of the previous year's expenditures for all items available at both times t and t - 1. Counts $n_{a,}$ n_{ac} and n_{acr} refer to the number of chains in index area a, the number of stores in chain c and the number of products sold in store r in chain c in index area a, respectively. The price p_{acrq}^{t} is computed as the unit value price per ounce of product q in store r in chain c at time t,

$$p_{acrq}^{t} = \frac{\sum_{i=1}^{n_{acrq}} P_{acrqi}^{t} Q_{acrqi}^{t}}{\sum_{i=1}^{n_{acrq}} Q_{acrqi}^{t} \; Z_{acrqi}^{t}} \text{,where}$$

$P_{acrqi}^{t}, Q_{acrqi}^{t}$ and Z_{acrqi}^{t} are the price, quantity in units, and size in ounces per unit, respectively, of the weekly summary i for product group q in store r in chain c in area a at time t. Sales data for the first three weeks in each month were averaged to produce unit valued prices.

2. Components of Variance Estimation

We fit a random effects log-linear model to our data, which were store-product level measures of unit-valued price change. Namely, we considered:

Y = Z $\boldsymbol{\gamma}$ + $\boldsymbol{e}$, where

$\mathbf{Y} = \ln \left(\frac{p_{arcq}^{t}}{p_{arcq}^{t-k}} \right)$, = store-UPC-level k-month log price change.

γ_1 = Chain effect

γ_2 = Cereal type (hot, sugary, fruity, plain) effect

γ_3 = store within chain effect and

e is the residual effect (product variation within chain, store and type). Here γ and e are assumed to be uncorrelated Gaussian variables with mean 0 and variances G and σ^2, the elements of G being the components of variance of interest. A reduced model, omitting the store effect, was also fitted.

Estimates of G were computed via the SAS PROC MIXED procedure, using a restricted maximum likelihood (REML) method, for 1-, 6- and 12-month price change for each month and index area. Store-product level expenditure share estimates were inserted as weights for the estimation of the error variance σ^2 (Pffefferman, et al., 1998).

3. Findings

Most estimates presented herein are based on the scanner data set with remainder chains, representing independent stores, excepted from each area sample. Figures 1-4 present components of variance for each index area, averaged over the 36-month study period. Figures 5-7 depict the behavior of components of variance over the study period, averaged at the national level for 1-, 6-, and 12-month price change. From figures 5-7 it is clear to see that the values of the two principal components, chain and cereal type, vary from month to month. This effect obtained in each index area also.

Table 1. Price change variance component estimates averaged across index areas

Region	Total Variance	Between Chain	Between Cereal Type	Between Store within Chain	Residual	% Share, Between Chain	% Share, Between Cereal Type	% Share, Between Store within Chain	% Share, Residual
1-month Price Change Variance Component Estimates									
National	0.004975	0.002712	0.002102	0.000152	9.76E-06	51.29	42.19	5.85	0.67
Northeast	0.006368	0.003295	0.003019	4.28E-05	1.15E-05	55.82	42.54	1.16	0.48
North Central	0.004279	0.001882	0.002258	0.000131	8.68E-06	43.73	48.08	7.01	1.18
South	0.004614	0.003138	0.001387	8.36E-05	5.47E-06	59.91	35.48	4.21	0.40
West	0.005017	0.002747	0.001961	0.000296	1.31E-05	48.24	42.09	9.13	0.54
6-month Price Change Variance Component Estimates									
National	0.006462	0.003297	0.002883	0.000273	9.8E-06	52.02	44.45	3.10	0.43
Northeast	0.007507	0.003628	0.003732	0.000135	1.09E-05	40.95	50.11	8.47	0.47
North Central	0.005908	0.002673	0.002963	0.000262	8.82E-06	55.39	37.80	6.58	0.23
South	0.005764	0.003438	0.002117	0.000203	5.54E-06	48.04	43.43	8.04	0.48
West	0.006873	0.003537	0.002897	0.000426	1.35E-05	48.67	44.06	6.87	0.41
12-month Price Change Variance Component Estimates									
National	0.005513	0.004257	0.000914	0.000332	1.01E-05	62.88	26.49	10.15	0.49
Northeast	0.005456	0.00451	0.000765	0.000172	1.07E-05	67.41	24.06	7.73	0.79
North Central	0.005005	0.003759	0.000917	0.00032	9.79E-06	56.51	30.67	12.23	0.58
South	0.005811	0.00477	0.000714	0.000321	5.62E-06	68.92	21.79	9.09	0.20
West	0.005766	0.004129	0.001169	0.000454	1.38E-05	60.83	28.08	10.65	0.45

Table 1 gives estimates of average 1-, 6-,and 12-month price change components, averaged across all 37 index areas and 36 months of the study . From Table 1 we can see that the two most important components of variance were between chain and between cereal type, and that this was consistent across all regions and lags. The chain effect appeared to gain importance with length of lag and again this effect was consistent across regions. Components of variance were similar in magnitude, across regions, with total variance and between chain variance components being slightly larger on the average in index areas in the Northeast and South.

The between store effect was generally but not uniformly small. It was significant in many months for several index areas, particularly the non-self-representing ones, and even in the analyses in which independent stores were excluded. Figure 8 gives the percentage of months of the study where the between store effect for 1-month price change was significant at the .05 level. The larger values for this component in the NSR areas are attributable to the lack of market identification in the scanner data. That is, between store variation included between city or market variation within chains for these areas.

The residual component was very small, both in absolute terms and as a percentage of total variance. This result was consistent across index areas and regions. This was primarily due to the use of expenditure weights in components computation. Figure 9 gives the percentage of total reported UPC groups, by decile of total expenditure share, averaged across all index areas. From these we see that, on the average, fewer than 30% of all UPC groups accounted for 90% of total cereal expenditures.

Cereal type piqued our curiosity. Figure 10 depicts the behavior of 1-month price change at the national level for each of the 4 cereal types. Price change for hot cereal exhibited cyclic behavior, taking steep dips in September and January, with extreme rebounds in October and February, leveling off in other months to behavior similar to the other cereal types. We investigated this, and discovered that the extreme behavior was attributable to several hot cereals, all from one national brand. Hot cereals represent about 14% of total volume sales for 2000; the extreme dips in September and January appear to have greater influence on the aggregate cereal index than their rebounding peaks, largely due to the dampening effects of price change behavior among the other cereal types in those months.

4. Conclusions

The findings of this research point to chains and cereal types as important stratification variables in sample selection for cereal. Within many areas, stores within chains exhibit little variation in pricing policy. In other areas, stores do vary within chain; however the magnitude of between store variation is dwarfed by their between chain and cereal type counterparts. Hot cereals as a group exhibit remarkably cyclic and extreme price change behavior, which is unlike other cereal types and that appears to be attributable to the pricing policy for one national brand.

Stratification by chain of the outlet frames in the CPI sample would represent a remarkable shift from current sample selection procedures for cereal, as well as for other food commodities. Currently, the CPI uses probability proportional to reported expenditure sampling for outlet selection, where individual outlets are included in sample frames if they are reported in the CPI program's Telephone Point of Purchase Survey. Outlets are not stratified by chain. It would be instructive to explore scanner data for other food commodity groups to determine whether the chain effect is as strong, and thus would be useful in stratifying outlets that sell multiple items, such as grocery stores.

5. Acknowledgments

The authors would like to thank Owen Shoemaker, William Johnson, William Cook, and Alan Dorfman for their careful reading of this paper. The authors also wish to acknowledge the contributions of Sonja Mapes and Scott Pinkerton in the classification of cereals into types. The authors thank Marvin Kasper and Janet Williams for their support on this project.

6. References

Baskin, Robert M. (1992), ``Hierarchical Bayes estimation of variance components for the U.S. Consumer Price Index", *ASA Proceedings of the Section on Survey Research Methods*, 716-719

Baskin, Robert M. (1993), ``Estimation of variance components for the U.S. Consumer Price Index via Gibbs sampling", *ASA Proceedings of the Section on Survey Research Methods*, 808-813

Baskin, Robert M. , and Johnson, William H. (1995), ``Estimation of variance components for the U.S. Consumer Price Index", *ASA Proceedings of the Section on Survey Research Methods*, 126-131.

Bureau of Labor Statistics (1992), *BLS Handbook of Methods*, Washington. DC: U.S. Government Printing Office, pp. 176-235.

Bureau of Labor Statistics, (1997), "The Experimental CPI using Geometric Means (CPI-U-XG)," April 10, 1997 (Washington: Bureau of Labor Statistics).

Dorfman, A.H., Lent, J., Leaver, S.G.., and Wegman, E., (2001) "On Sample Survey Designs for Consumer Price Indexes," *Proceedings of the 53rd Session of the ISI.*

Garrett, J. K. and Harter, R. M. (1995), "Chapter 10: Sample Design using Peano Key Sequencing in Market Research" *Business Survey Methods*, . Wiley & Sons, Inc., pp.205-217.

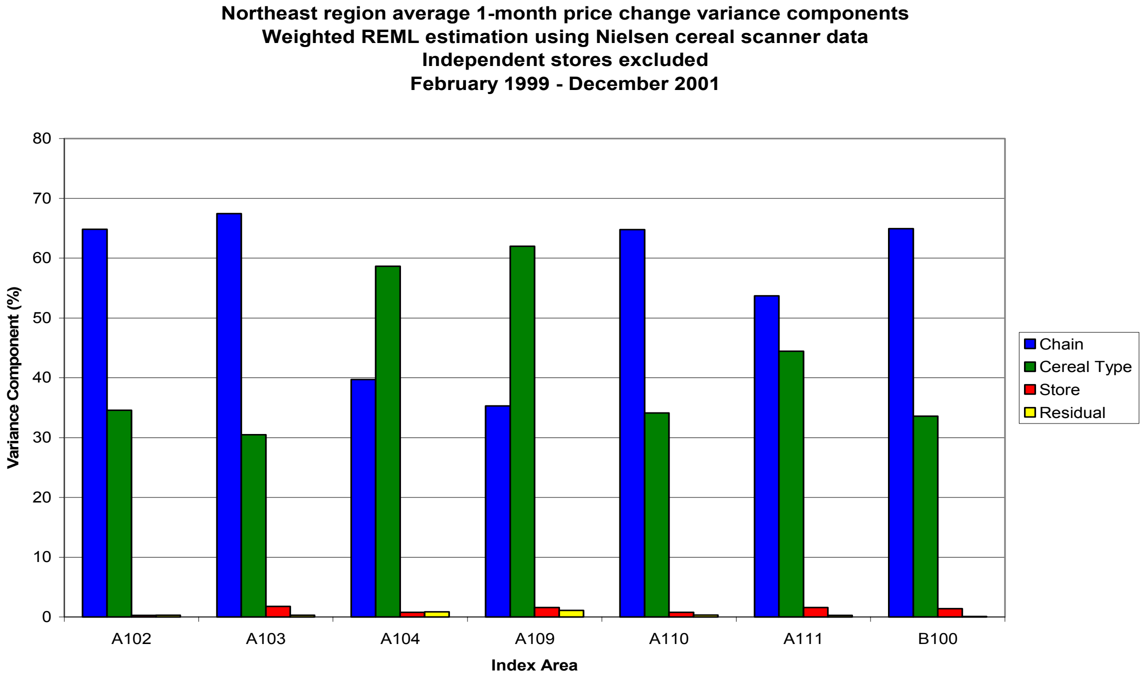
Leaver, S. G. and Valliant, R. L. (1995), "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods.* Wiley & Sons, Inc., pp. 543-566.

Pfeffermann, D., Skinner, C.J., Goldstein, H., Holmes, D.J., Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models (with discussion)", Journal of the Royal Statistical Society, Series 60, pp. 23-40.

Shoemaker, Owen J. (2001), "Estimation of Variance Components for the U.S. Consumer Price Index: A Comparative Study," *Proceedings of the Government Statistics Section, American Statistical Association.*

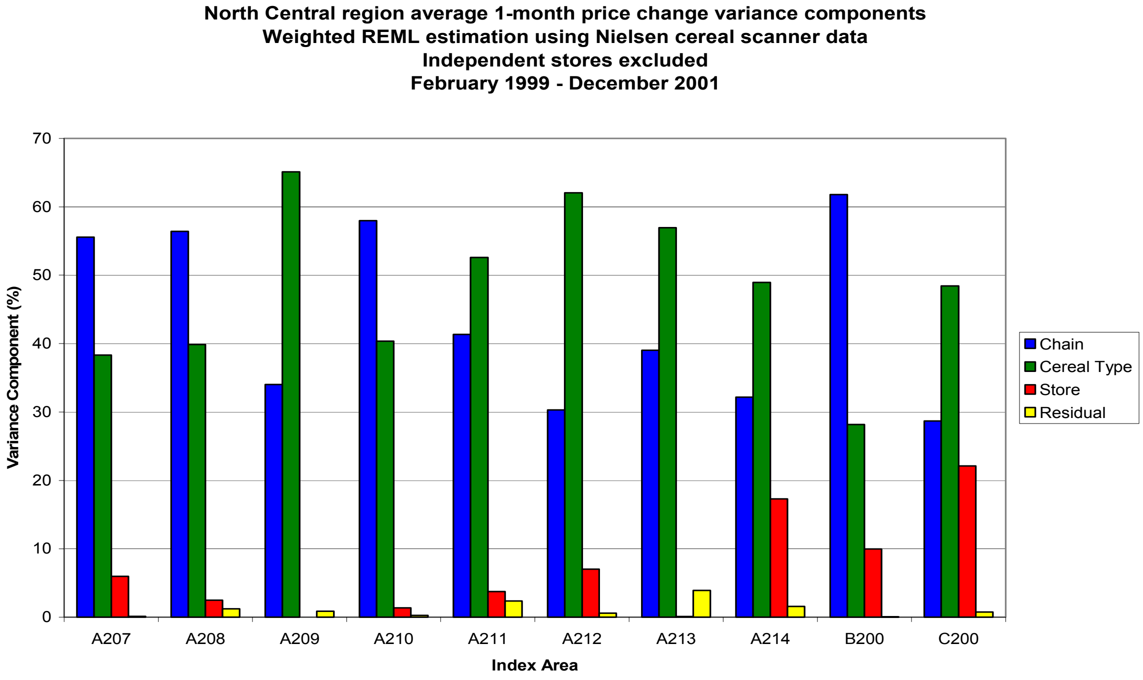
Shoemaker, Owen J. (2002), "Estimation and Analysis of Variance Components for the Revised CPI Housing Sample." *Proceedings of the Government Statistics Section, American Statistical Association.*

Figure 1



Northeast region average 1-month price change variance components
Weighted REML estimation using Nielsen cereal scanner data
Independent stores excluded
February 1999 - December 2001

Figure 2



North Central region average 1-month price change variance components
Weighted REML estimation using Nielsen cereal scanner data
Independent stores excluded
February 1999 - December 2001

Figure 3

South region average 1-month price change variance components
Weighted REML estimation using Nielsen cereal scanner data
Independent stores excluded
February 1999 - December 2001

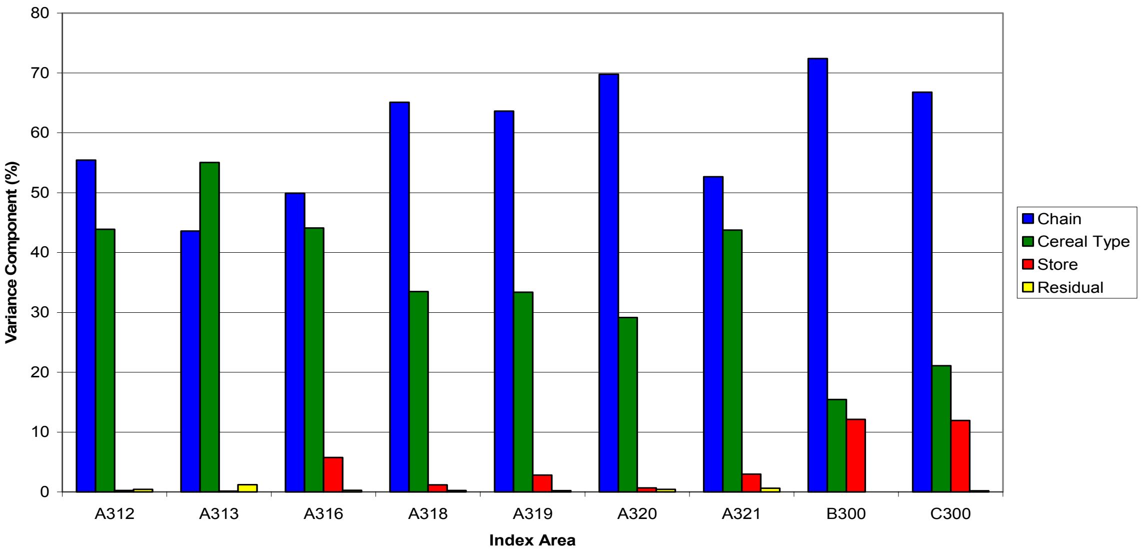


Figure 4

West region average 1-month price change variance components
Weighted REML estimation using Nielsen cereal scanner data
Independent stores excluded
February 1999 - December 2001

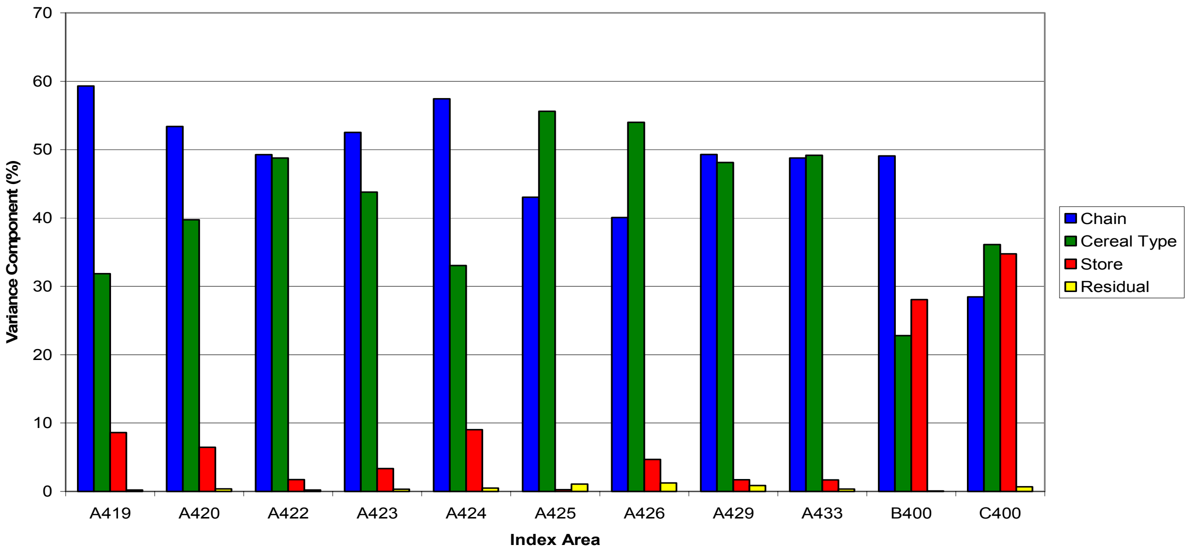


Figure 5

National Components of Variance Percentage for 1-Month Price Change
Nielsen Cereal Scanner Data
Independent Stores Excluded
February 1999-December 2001

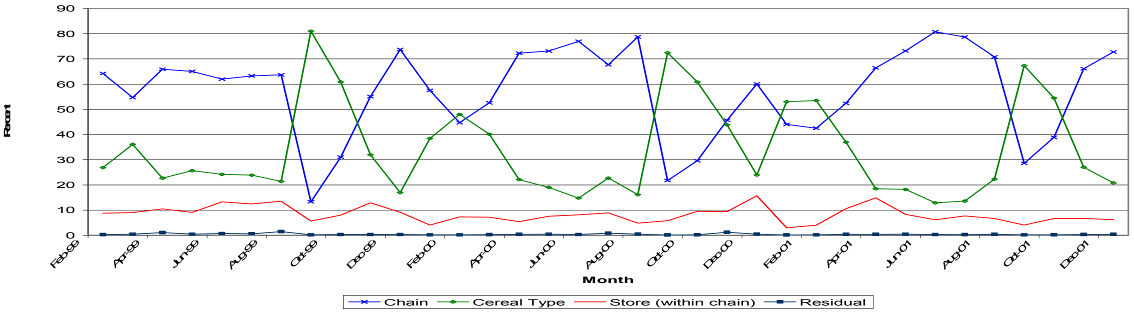


Figure 6

National Components of Variance Percentage for 6-Month Price Change
Nielsen Cereal Scanner Data
Independent Stores Excluded
July 1999-December 2001

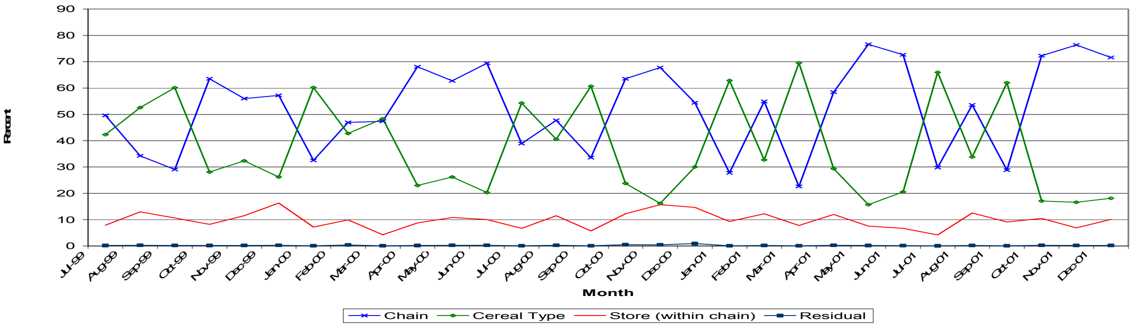


Figure 7

National Components of Variance Percentage for 12-Month Price Change
Nielsen Cereal Scanner Data
Independent Stores Excluded
January 2000-December 2001

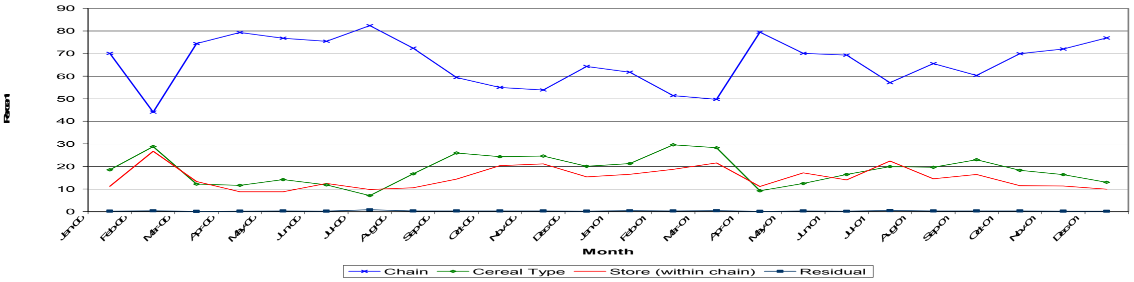


Figure 8

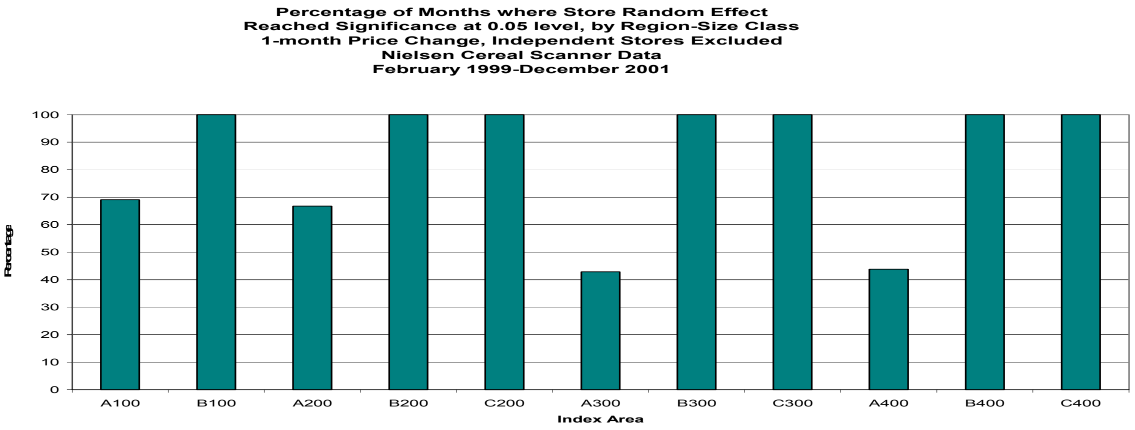

Percentage of Months where Store Random Effect
Reached Significance at 0.05 level, by Region-Size Class
1-month Price Change, Independent Stores Excluded
Nielsen Cereal Scanner Data
February 1999-December 2001

Figure 9

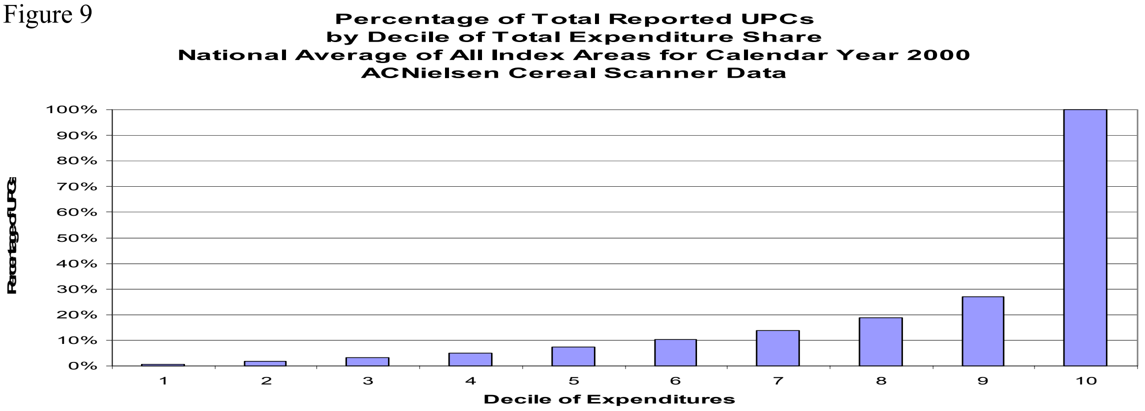
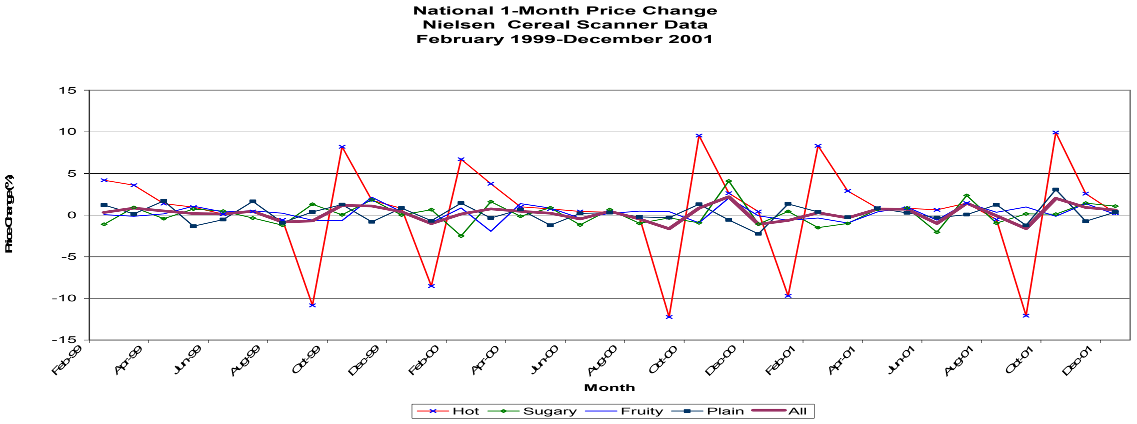

Percentage of Total Reported UPCs
by Decile of Total Expenditure Share
National Average of All Index Areas for Calendar Year 2000
ACNielsen Cereal Scanner Data

Figure 10


National 1-Month Price Change
Nielsen Cereal Scanner Data
February 1999-December 2001