

Visualizing Cereal World

Edward J. Wegman
Center for Computational Statistics
George Mason University

and

Alan Dorfman
Office of Survey Methods Research
Bureau of Labor Statistics

Abstract

We discuss the basic theory of price indices. These form the basis for the Consumer Price Index in the United States. We note situations which can lead to non-robustness and volatility in several different price indices. An experiment to replace survey-based data with point of sales (POS) scanner data is being carried out. We visually explore the scanner data and conclude that certain aspects of scanner data can lead to non-robustness of commonly used price indices.

1. Introduction and Background

The consumer price index (CPI) is arguably one of the most important government statistics. As the principal measure of inflation, it is a major factor in driving social and economic policy with significant impact on the daily lives of virtually every resident of the United States. As such it is constantly under review for technical improvements in methodology with the idea of increasing accuracy. With the continuing improvement in the electronic acquisition of data and storage of data in data warehouses, new possibilities have arisen in ways of calculating the CPI.

Any price index is intended to measure the general increase or decrease in prices from one time period to another when aggregated over a large number of items. This is typically done by averaging price ratio weighted by the level of expenditures. Traditionally data for these computations are obtained by surveys. Household Expenditure Surveys are used to accumulate information on amount spent and on the variety of items purchased. This is normally done by selecting a representative sample of households and interviewing them or asking them to keep a diary of purchases over a specified time period. Retail Establishment Surveys are used to accumulate data on prices. These are normally carried out by having field agents collect price information on selected items in a retail establishment. Data from these two surveys are combined to establish expenditures and price ratios. Indices are thus computed by combining data from two surveys.

Many retail establishments have computerized their inventory control and pricing control by introducing electronic scanning of universal product codes (UPC). This computerized acquisition of price and expenditure data suggests that a substantial improvement in accuracy of a price index might be realized by using these data in place of or as supplement to the traditional survey collected data. On the downside, the scanner data are not collected for the purpose of constructing the CPI and, therefore, must be regarded as opportunistically collected data not necessarily collected with appropriate statistical designs. Not all retail establishments are equipped with scanner systems so complete reliance on scanner data would not be feasible. Moreover, it is not clear that traditional methods for calculating price indices can be easily adapted to the scanner data because the noisiness of scanner data has the potential to distort the interpretation of the price index. In this article, we examine some of the traditional methods for computing price indices, discuss some of the criteria for choosing a price index, and investigate, using exploratory data analysis (EDA) techniques, the potential impact of scanner data on price index computation.

The Bureau of Labor Statistics has launched a limited, multi-year experiment studying the impact of scanner data on the consumer price index. The data, which are involved in this experiment and which we examine in this paper, are scanner data on breakfast cereals sold in a large metropolitan area during the period 1996 to 2001. This paper describes the use of EDA methods to visually explore this vast "cereal world."

2. Price Index Theory

Perhaps the most surprising aspect of the uninitiated is the subtlety of the issues surrounding the price index. As described above, the price index would seem to be a straightforward weighted average. Most statisticians, indeed, most scientists deal with ordinary arithmetic averages and give little thought beyond. Traditionally, however, there are two other averages worth considering: the geometric and the harmonic averages. We review each of these and show that a wide variety of price indices can be invented based on these simple averages.

Arithmetic Mean - Consider a set of observations x_1, \dots, x_n . The traditional statistical mean is the arithmetic average given by

$$\bar{x}_a = \frac{1}{n} \sum_{j=1}^n x_j.$$

The arithmetic average equally weights each observation which is useful for randomly sampled observations. However, for many purposes it is more useful to consider a weighted arithmetic average

$$\bar{x}_{aw} = \sum_{j=1}^n w_j x_j \text{ where } \sum_{j=1}^n w_j = 1.$$

Geometric Mean - The arithmetic mean is tied to linear inferences in the sense that if points fall on a straight line the arithmetic average will also fall on the same straight line. However, if the observations fall on an exponential (or convex) curve, by Jensen's inequality, the arithmetic average will fall above the curve. However, the geometric mean will fall on the exponential curve. Here

$$\bar{x}_g = \prod_{j=1}^n x_j^{\frac{1}{n}}$$

and, as with the arithmetic mean, we can have a weighted geometric mean as well given by

$$\bar{x}_{g_w} = \prod_{j=1}^n x_j^{w_j}, \text{ where } \sum_{j=1}^n w_j = 1.$$

Harmonic Mean - The third type of mean is the harmonic mean. Here the equally weighted harmonic mean is given by

$$\bar{x}_h = \frac{1}{\frac{1}{n} \sum_{j=1}^n x_j^{-1}}$$

and the weighted harmonic mean is given by

$$\bar{x}_{h_w} = \frac{1}{\sum_{j=1}^n w_j x_j^{-1}} \text{ where } \sum_{j=1}^n w_j = 1.$$

For a given set of weights, these means satisfy the inequality $\bar{x}_{h_w} \leq \bar{x}_{g_w} \leq \bar{x}_{a_w}$ with equality when the x_i are equal.

Now let us consider a population of n items which are available for sale. Let p_{jt} and q_{jt} be the price and quantity sold respectively of item $j \in \{1, \dots, n\}$ in time period $t \in \{B, 1, 2\}$. Here B is a base time period. We wish to measure inflation between periods 1 and 2. The ratio $\frac{p_{j2}}{p_{j1}}$ is called the *price relative*. A simple arithmetic average of the price relative is called the Carli Index, introduced in 1764, and given by

$$A = \frac{1}{n} \sum_{j=1}^n \frac{p_{j2}}{p_{j1}}.$$

Its weighted counterpart, $S = \sum_{j=1}^n w_j \frac{p_{j2}}{p_{j1}}$, is referred to as the Sauerbeck index. These indices possess a fatal flaw. In practice, long range indices are often produced by "chaining," i.e. multiplying intermediate indices. It is thus desirable that the product of the index between periods 1 and 2 with the index between periods 2 and 3 should

approximate what one gets between 1 and 3 directly. Suppose period 3 prices return exactly to their first period level, i.e. $p_{j3} = p_{j1}$ for all j . Then by the Cauchy-Schwartz inequality, $S_{12} \cdot S_{23} \geq S_{13}$, i.e., chaining this index must yield something too large.

The standard formula used by governments around the world is the Laspeyres index, which dates from 1871 given by

$$L = \sum_{j=1}^n w_{j1} \left(\frac{p_{j2}}{p_{j1}} \right), \text{ where } w_{j1} = \frac{p_{j1} q_{j1}}{\sum_{k=1}^n p_{k1} q_{k1}}.$$

This is a ratio of total expenditures $L = \frac{\sum_{j=1}^n q_{j1} p_{j2}}{\sum_{j=1}^n q_{j1} p_{j1}}$. The numerator is what the second period

expenditure would have been had the quantities of items equaled the quantities in the first period divided by the actual first period expenditures. A relatively simple algebraic computation shows that the Laspeyres Index is also a weighted arithmetic average of the price relatives where the weight w_{j1} is just the proportion of expenditure on item j in time period 1. The difference between this and the Sauerbeck index is that the weights are tied to the very price relatives that they are weighting.

A natural question is why should time period 1 determine the quantity. Time period 2 has an equal claim to being legitimate. The Paasche Index dating from 1874 provides an alternative. Here

$$P = \frac{\sum_{j=1}^n q_{j2} p_{j2}}{\sum_{j=1}^n q_{j2} p_{j1}} = \frac{1}{\sum_{j=1}^n w_{j2} \left(\frac{p_{j2}}{p_{j1}} \right)^{-1}} \text{ where } w_{j2} = \frac{p_{j2} q_{j2}}{\sum_{k=1}^n p_{k2} q_{k2}}.$$

The Paasche is essentially the same as the Laspeyres Index except that the quantities, q_{j2} , from time period 2 are used. Note that this subtle change has a rather dramatic effect: the Paasche Index is a harmonic mean of the price relatives.

The observation that prices tend to go up by percentages rather than fixed monetary increments suggests that an exponential model may be more realistic, hence, that geometric averages may be more appropriate. The straightforward equally weighted geometric mean of the price relatives is the so-called Jevons Index dating from 1863.

$$G = \prod_{j=1}^n \left(\frac{p_{j2}}{p_{j1}} \right)^{\frac{1}{n}}.$$

It has the drawback that if an item has $p_{j2} = 0$, then the Jevons index = 0 also. There is a similar problem with items for which $p_{j1} = 0$ makes the Jevons index infinite. The weighted version is called the Expenditure Weighted Jevons with base period B

$$G_w = \prod_{j=1}^n \left(\frac{p_{j2}}{p_{j1}} \right)^{w_{jB}} \quad \text{where} \quad w_{jt} = \frac{p_{jt}q_{jt}}{\sum_{k=1}^n p_{kt}q_{kt}}.$$

Another class of indices are the so-called *superlative indices*, which attempt to overcome the preference for expenditures in time period 1 or time period 2 by using both. The Tornqvist Index dating from 1936 is a commonly used geometric index which uses an average of expenditure weights from both time periods. It is calculated by

$$T = \prod_{j=1}^n \left(\frac{p_{j2}}{p_{j1}} \right)^{\frac{(w_{j1} + w_{j2})}{2}}.$$

Fisher's Ideal Index dating from 1922 combines the Laspeyres and the Paasche indices by the following formula

$$F = \sqrt{LP} = \left\{ \sum_{j=1}^n w_{j1} \left(\frac{p_{j2}}{p_{j1}} \right) \frac{1}{\sum_{j=1}^n w_{j2} \left(\frac{p_{j2}}{p_{j1}} \right)^{-1}} \right\}^{1/2}.$$

Tests for Indices - Given the plethora of indices available, one naturally asks the question of which index is best. One approach often preferred by economists is to create sensible tests of indices. In the following discussion, let $\mathbf{p}^t = (p_{1t}, \dots, p_{nt})$ and $\mathbf{q}^t = (q_{1t}, \dots, q_{nt})$. An example of a test is the so-called *identity test*. That is, if $I(\mathbf{p}^1, \mathbf{p}^2, \alpha \mathbf{q}^1, \beta \mathbf{q}^2) = 1$ for all $\alpha > 0, \beta > 0, \mathbf{p}^1 = \mathbf{p}^2, \mathbf{q}^1 = \mathbf{q}^2$. This test says that if the price doesn't change then the price index should be 1 irrespective of quantity sold. Another test is the *proportionality test*, which is formulated as $I(\mathbf{p}^1, \alpha \mathbf{p}^2, \mathbf{q}^1, \mathbf{q}^2) = \alpha I(\mathbf{p}^1, \mathbf{p}^2, \mathbf{q}^1, \mathbf{q}^2)$ for all $\alpha > 0$. This is to say that if all prices change by a fixed proportion, the index must change by the same proportion. There are many such tests. In effect we used the *symmetric treatment of time* (time reversal) test to reject the Carli and Sauerbeck indices earlier.

3. Impact of Outliers

It is well known that the arithmetic mean is not robust against large positive outliers. The impact of outliers on the geometric and harmonic means, however, is not well studied. The impact on the geometric mean can be considered as follows. Let

$$\bar{x}_{g_w} = \prod_{j=1}^n x_j^{w_j} = \prod_{j=1}^n e^{w_j \ln(x_j)} = e^{\sum_{j=1}^n w_j \ln(x_j)}.$$

The geometric mean is related to the arithmetic mean of the logarithm of the observations. Thus if x_j is a large positive outlier, the logarithm scales back the magnitude of x_j and, thus, makes the geometric mean relatively robust to large outliers in the price ratio. Conversely, small outlier, i.e. x_j near 0, causes $\ln(x_j)$ to be very large in a negative sense. The geometric mean is not robust to values near 0. An extreme value of

w_j , the proportion of expenditure on items j would not be mitigated by the logarithmic effect. Because $\sum_{j=1}^n w_j = 1$, an extreme value w_k here would make the corresponding price relative dominate the index. The effect is compounded with geometric indices because as the price relative decreases, making the item a better bargain, the expenditure is likely to go up creating increased emphasis on a now inexpensive item in the computation of the price index.

In a similar way consider the harmonic mean

$$\bar{x}_{h_w} = \frac{1}{\sum_{j=1}^n w_j x_j^{-1}}.$$

Notice that a large positive value of x_j would lead to a small, even negligible, value of x_j^{-1} . Thus a large value of price relative would tend to have much less of an effect on the harmonic mean and we would expect the harmonic mean to be relatively robust as well when compared to the arithmetic mean. However, as with the geometric indices, a 0 value of x_j will cause the harmonic to be 0, so that indices based on harmonic means are not robust to small positive values of x_j . Because $\sum_{j=1}^n w_j = 1$, an extreme value of w_k would make the corresponding price relative dominate the index. Thus the good news is that we would expect indices based on geometric and harmonic means to be relatively robust to large values of price ratios, but not robust to small values of price ratios. The bad news is that all of these indices are vulnerable to extreme values of expenditure proportions causing price relatives corresponding to large expenditures to dominate the price index.

Lent (2001) analyzes in detail the effect of outliers on several price indices, in particular the superlative indices taking into account the economic conditions that affect the weights. For example, under conditions where the first and second expenditures are close, the Tornqvist index tends to be more robust to large outliers than the Fisher Index. This is not unreasonable because the Fisher is formed by an inner product of arithmetic mean terms and harmonic mean terms while the Tornqvist is essentially a geometric mean. The harmonic mean would tend to be robust to outliers, but the arithmetic mean would not. Thus the Fisher index can be viewed as a product of a non-robust and a reasonably robust component. In all cases, these indices would be sensitive to extremes in the expenditure.

4. Particulars of the Cereal Data

The data that we will examine are point of sales (POS) scanner data collected in supermarkets and assembled by a commercial vendor to provide feedback to manufacturers on the relative popularity and viability of their products. This is a common strategy that manufacturers use. A third party vendor assembles, cleans and formats the data into a common format. One lesser role of the third party vendors is to organize the

and no promotions with magenta. Magenta and green are additive complimentary colors; the white areas indicate arise from the color addition of magenta and green.

Basic Exploratory Data Analysis - The basic exploratory analysis was begun essentially as a visual data mining operation. Five variables were used: UPC code, Quantity, Price, Store Code and a Promotion Code. There were 15 distinct types of promotion, which were distinguished by level of aggressiveness. These were given as 15 binary variables, which we decided to consider as a 15 digit number. This choice was made for both simplicity and because the actual types of promotion were in fact rather limited to about 6 or 7 major types.

Figure 1 illustrates a significant finding of our exploratory analysis. There were three "aggressive" types of promotions and three less aggressive types of promotions which dominated the marketing efforts. Of course there were also cereals that were not promoted at all in this time period. See Figure 2.



Figure 2. Detail of scatter plot matrix showing the color coding of promotions. The same color coding as in Figure 1. The vertical axis is the promotion variable, the horizontal is the store code variable.

The most aggressive promotions were brushed with green as shown in Figure 2. These are respectively starting with the largest numerical code: 1) national advertising by the manufacturer, 2) Sunday supplement advertising by stores/change in local newspapers, and 3) distribution of coupons either directly or in newspapers. The less aggressive promotions are coded in yellow and represent marketing devices such as endcaps (displays at the end of an aisle in a store), coupon machines in the aisle, and similar devices. Items not promoted are coded as magenta. Figure 3 is another detail from this scatterplot matrix.

Figure 3 maintains the same color coding as in Figures 1 and 2. The price is the price per item, i.e. per box or package. What is dramatically clear from this scatterplot is that aggressive promotion dramatically increases sales as much as 200-fold. Also clear from

this plot are that there are fixed price points and that aggressive promotion can dramatically increase sales even at relatively large price points. Of course, large quantities sold imply that large expenditures are made and as indicated in Section 3, all of the price indices are not robust relative to large expenditures. The traditional sampling-based data collection would not tend to pick up these large expenditures that are apparent with the scanner-based data. Hence, one could expect a significant difference between scanner-based computation and sampling-based computation.

One additional comment is perhaps useful on Figures 1 to 3 as well as other Figures in this paper. We are using additive color schemes in our images. In additive color schemes, blue and red together result in magenta. Magenta and green are complimentary colors. Hence when blended together they yield white. Thus the white seen in these figures is the result of green and magenta being mixed as additive colors.

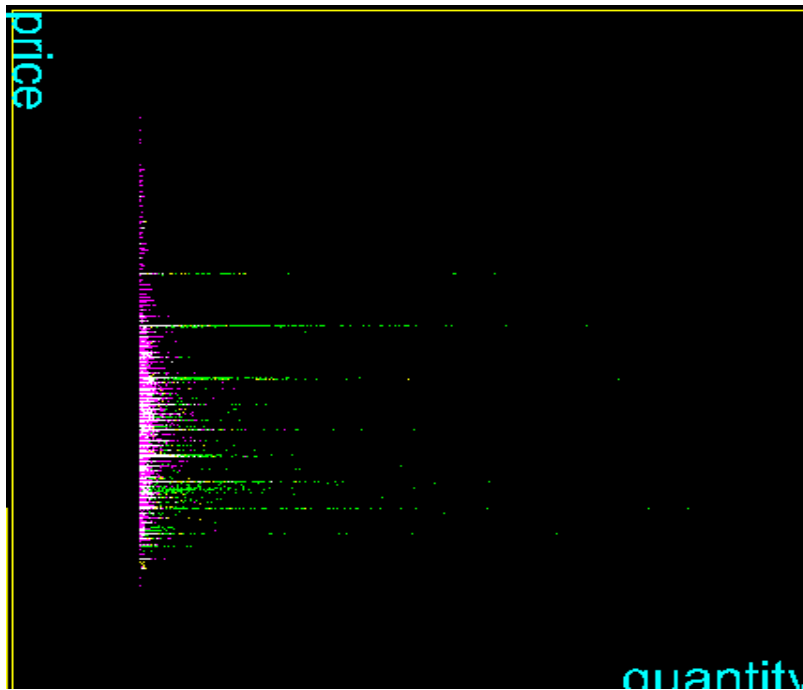


Figure 3. Detail of the scatterplot in Figure 1 showing the effect of promotion on quantity. The color coding is the same as in Figures 1 and 2.

Figure 4 is the result a small amount of (grand tour) rotation among the variables price, promotion and quantity. In this scatterplot all promoted items are brushed green and all non-promoted items are brushed red. This plot suggests that of the aggressive promotion schemes, the Sunday supplement advertising (second tier in the larger plot) is responsible for substantially increased sales, more so than national advertising by manufacturers (first tier) and coupons (third tier). This is an interesting result because advertising does not intrinsically discount the price, while coupons and store promotions do. In this figure, green and red in additive color result in yellow.

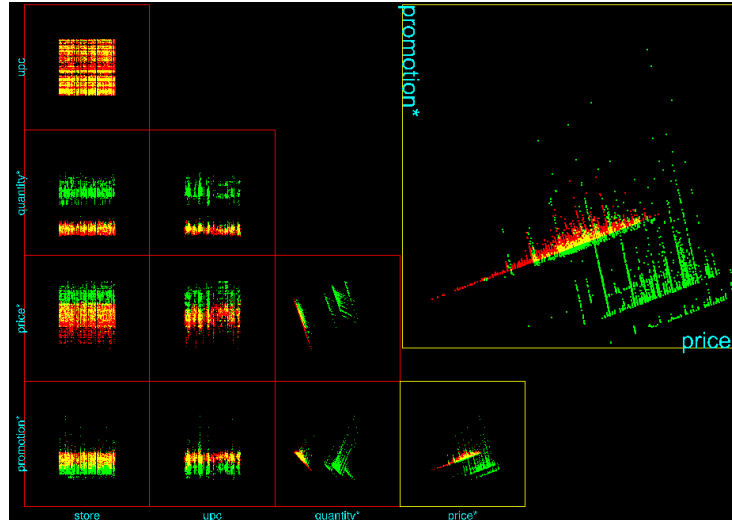


Figure 4. A scatterplot matrix with a grand tour rotation of the price, promotion and quantity variables. In this figure, all promotions are coded as green, no promotion is coded as red. In additive color, red plus green makes yellow. This perspective indicates that Sunday supplement advertising (the second tier in the lower right-hand corner of the larger plot) is a winning strategy in boosting sales volume. National advertising (the first tier) is less effective.

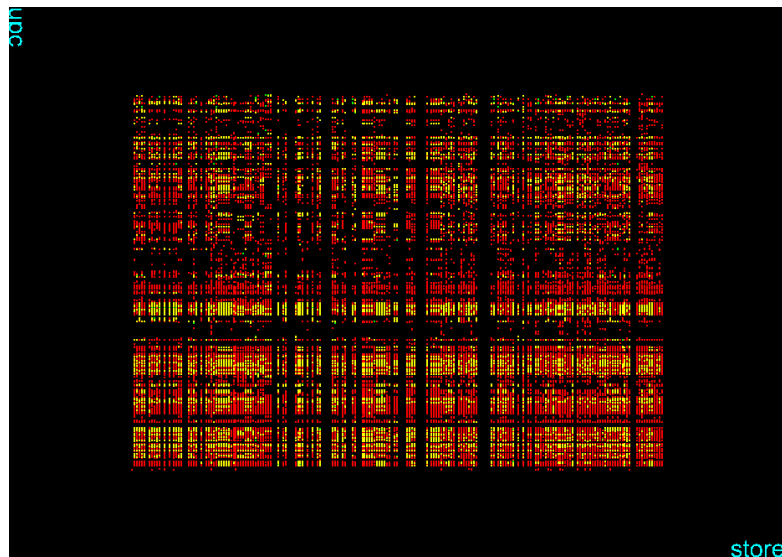


Figure 5. Scatterplot of UPC versus store codes. Red indicates the cereal was not promoted in April, 1999 while yellow (green plus red) indicates that at least in some markets the cereal was promoted. Horizontal banding is to be expected and represents marketing efforts by some manufacturers. Vertical banding should not occur because store codes are in theory randomized.

Figure 5 has same color coding as Figure 4 and represents a scatterplot of UPC codes versus store codes. This scatterplot records discrete variables against discrete variables and illustrates the utility of a scatterplot in situations with a very large number of discrete variables. An interesting aside is that this plot strongly resembles the microarrays found in microbiology studies of the genomics. UPCs are given to manufacturers in blocks, so that adjacent runs of UPCs belong to the same manufacturer. For this reason, in Figure 5

it is reasonable to expect horizontal banding, i.e. bands of yellow or red corresponding to promotions. What is perhaps unexpected is the vertical banding as the labeling of stores was supposed to have been randomized to prevent manufacturers from identifying individual stores. Other data exploration suggests strongly that adjacent store codes correspond to geographic adjacency so that in principle one could localize stores to specific geographic regions if not identify the stores individually.

Price Relative and Expenditure Effects over a One Year Period - The first part of our visual data analysis suggests that promotions have a dramatic effect on expenditures, which in turn change the character of price indices. A natural question is what is the effect of promotions on price relatives?

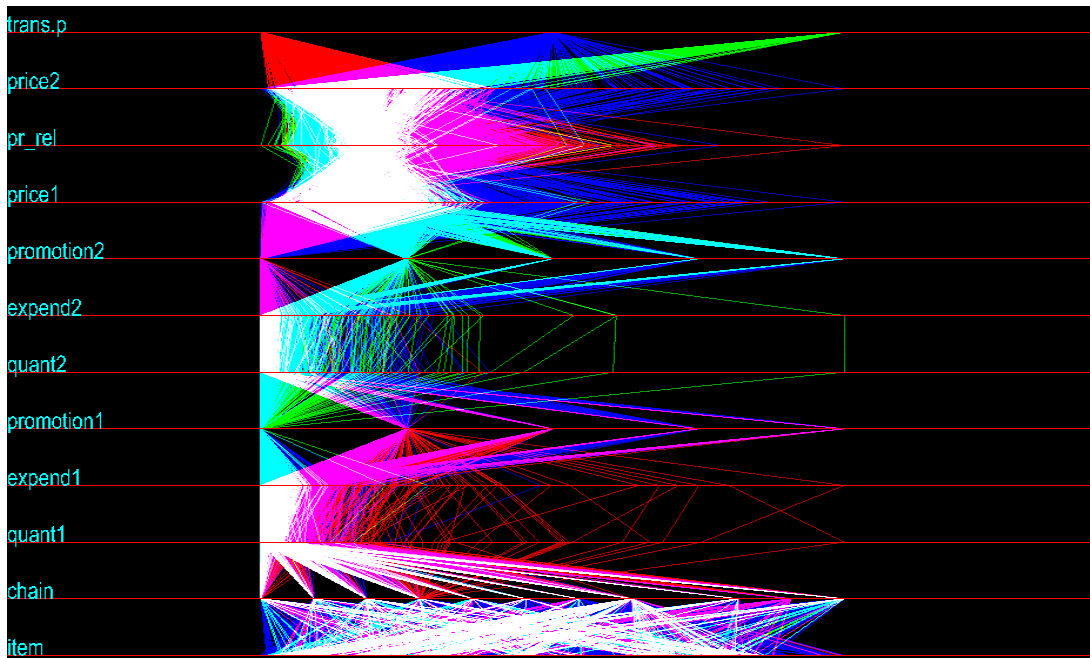


Figure 6. Parallel coordinate plot exploring impact of promotions on price relatives and on expenditures. The color coding is as follows: 'red' means promoted the first year, but not the second, 'green' means promoted the second year, but not the first, and 'blue' means same promotion strategy for both the first and second years.

Figure 6 represents data over a one year period (February, 1999 to February, 2000). The variables involved here are an item code, a chain code, quantity, expenditure, price and promotion code in year one, quantity, expenditure, price and promotion code in year two, price relative, and a code indicating promotion strategies in year one and in year 2. This latter variable is labeled trans.p and is coded to be 1 and brushed with red if the item was promoted in the first time period (February, 1999) but not promoted in the second time period (February, 2000). Trans.p is coded as 3 and brushed with green if the item was promoted in the second time period, but not the first. Finally, trans.p is coded with 2 and brushed with blue if there was no change in the promotion strategy. In this second phase,

we investigate unit pricing, i.e. the price variable is price per ounce, not price per package. The data are now 12 dimensional.

In Figure 6, we are again using an additive color scheme. Red + green = yellow, red + blue = magenta, green + blue = cyan and of course red + green + blue = white. Complimentary color pairs are red and cyan, green and magenta, and blue and yellow. Complimentary pairs also sum to white. Figure 6 again suggests that aggressive promotion yields outliers in quantity sold and total expenditure. This is consistent with our earlier findings. Examining the `pr_rel` (Price Relative) variable suggests that promotion the first year, but not the second tends lead to an increase in the price relative (red and magenta), while promotion the second year, but not the first tends to be associated with a decrease in the price relative (green and cyan). However there is a strong area of overlap with all three promotion strategies, i.e. where the data are coded as white.

The source of potential concern is that the expenditures and the price relatives are strongly linked by the promotion effect. Generally speaking, heavy promotion lowers prices and raises expenditures. Because the promotions tend to emphasize different products from year to year, promotions in the second year will drive down price relatives for promoted items and increase expenditures (and sales volumes). Ordinary survey-based methods for collecting data have a sample size fixed by design so that the impact of a huge sales volume is not captured. A family may adjust what product they buy because of a promotion, but will probably not adjust the amount of what they buy by very much. Scanner data, to the contrary, reflect the large increase in sales volume and consequently the total expenditure increase. Thus the price indices based on scanner data are much more likely to be sensitive to the promotion effect and will give a substantially different impression of inflation.

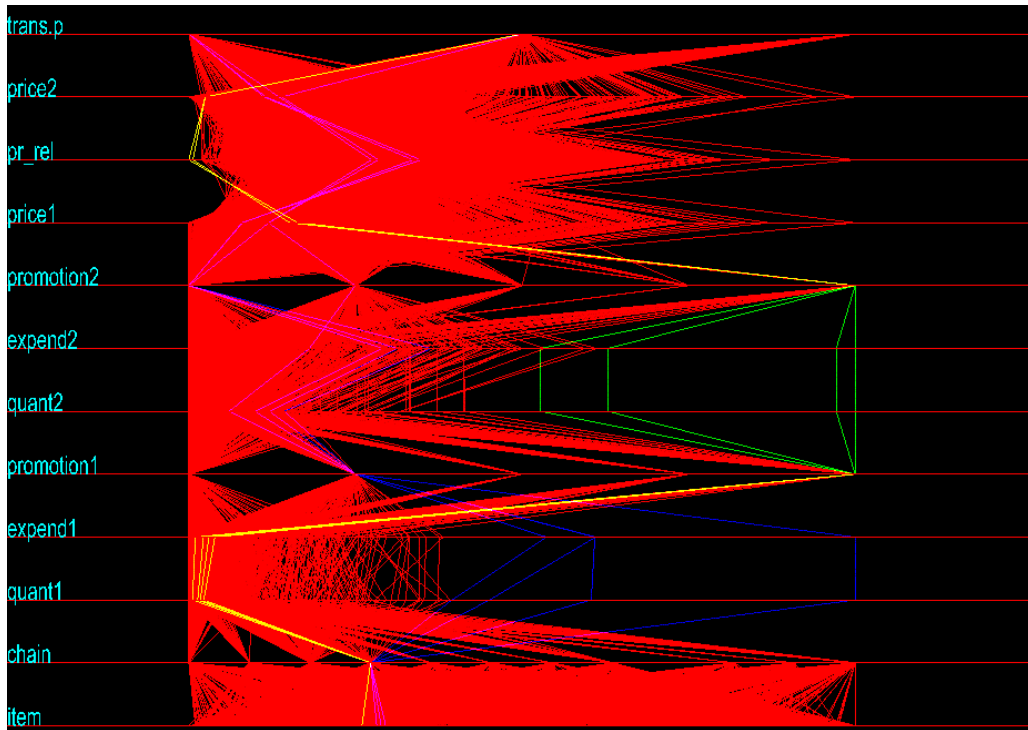


Figure 7. Outliers Belong to a Single Chain. The color coding here is 'green' means outliers in quantity year 2 variable, 'blue' means outliers in quantity year 1 variable, 'red' is all others.

Figure 7 illustrates an interesting, but not very important fact. We were curious to identify the properties of the outliers in expenditures and quantities for year 1 (1999) and for year 2 (2000). By brushing all the data with red and then brushing the year 1 outliers with blue and the year 2 outliers with green, we are able to trace these outliers through the graphic. The interesting effect is that these promotions are both traced to the same chain of stores. Recall that blue + red is magenta and green + red is yellow. Thus the magenta is the continuation of the blue traces and the yellow is the continuation of the green traces. Also of interest is that the blue outliers belong to items that were promoted the first year but not the second. The green outliers were promoted both year, but only in the second year did the sales volume take off. This suggests that these were new products in year 1 for which the supply chain had not been fully stocked.

Figure 8 illustrates one other interesting piece of detective work. In this image it is noted that one chain ceased promotions in the second year (2000). This chain is brushed with green. For $trans.p = 1$, the data are brushed with red and for $trans.p = 2$ or 3 , the data are brushed with blue. In Figure 8 the yellow, cyan and white all reflect data from the chain that ceased promotion. For this chain, sales quantities were not large in either year 1 or year 2 and remained about the same whether they engaged in promotional activities or not. Also clear is that this chain's prices were about the same in both years so that price relatives were constrained to a fairly tight region. Apparently this chain made a sensible business decision by avoiding the extra expenses incurred by extensive promotional activities.

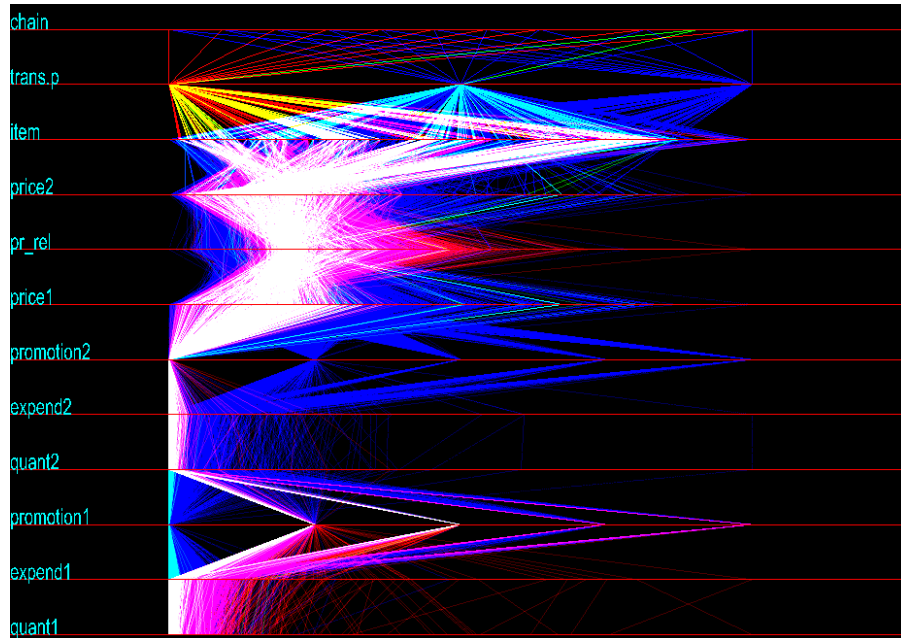


Figure 8. One Chain Ceased Promotions. The color coding is as follows: 'red' means promotion the first year, but not the second, 'blue' means either no change or promotion second year but not the first, 'green' identifies the one chain that stopped promotion. Recall green+red = yellow and green+red+blue = white. The yellow and white areas indicate the impact of that single chain.

Churning Effects - Our third data analysis focused on investigating churning effects. Churning occurs because an item that was sold in one time period is not sold in the second time period. This is not a problem over a relatively short time period, but over a longer time period many items may be lost and so it is impossible to compute their price relatives. In the 1996 to 2001 time period covered by this third data analysis only about 35% of the items were available in both years. The data examined in this setting were now 18 dimensional. The variables examined this time were the store code (Store), the item code (Item), the probability sampling unit (PSU), the quantity sold in ounces in year 1 (Quant1), the unit price in year 1 (Price1), the promotion code in year 1 (Promo1), the total expenditure in year 1 (Expend1), the quantity sold in ounces in year 2 (Quant2), the unit price in year 2 (Price2), the promotion code in year 2 (Promo2), the total expenditure in year 2 (Expend2), the price relative (PrcRel), the promotion transition (Trans.p) a code for the type of cereal (Type), a manufacturer code (Manufact), store birth or death code (StoreB&D), an item birth or death code (ItemB&D), and a code for the chain (Chain2).

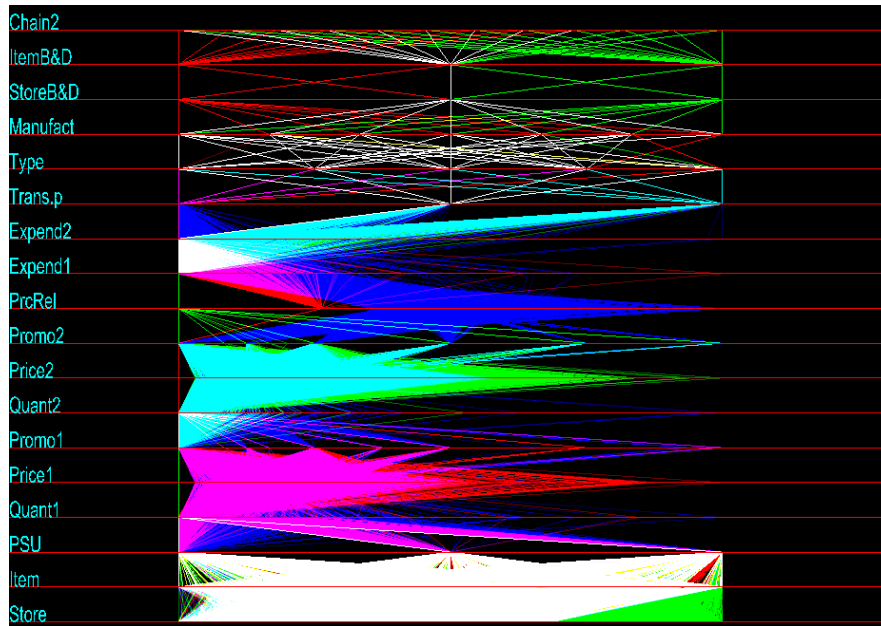


Figure 9. Churning effects. Price relative = 0 is coded as red (item not available in 2001) while price relative = ∞ is coded as green (item not available in 1996).

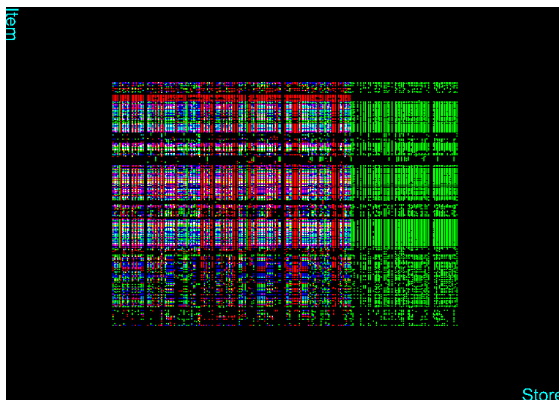


Figure 10. The vertical axis is item code, the horizontal axis is store code.

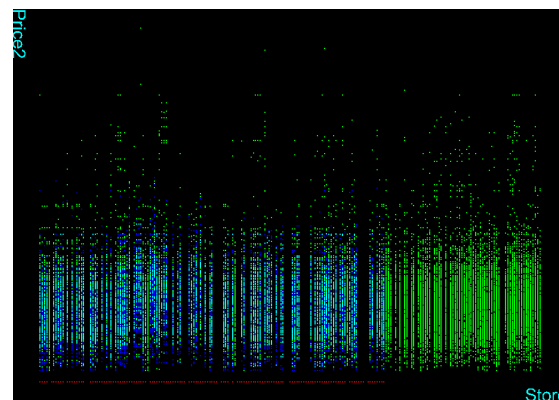


Figure 11. The vertical axis price year 2, the horizontal axis is store code.

As can readily be seen from Figure 9, an item can be unavailable for two reasons. First of all the item itself may have been discontinued or newly introduced (ItemB&D) or the store, which might have sold the item, either left the sample or was newly included in the sample (StoreB&D). Also note that newly included stores have simply been added with larger codes (Store).

Figure 10 is a plot of Item versus Store. The interesting aspect of Figure 10 is that the vertical banding in green indicates that the store came into the sample between 1996 and 2001. The vertical banding in red indicates that the store went out of the sample. Horizontal banding in green indicates a newly introduced product whereas horizontal banding in red indicates a discontinued product.

Figure 11 is a plot of second year price versus store code. While many newly introduced products overlap in prices with the existing products, it is clear that all of the most expensive items are new products. Thus one potential effect of churning is that new products are introduced as a method for raising prices. Since there is no price in the earlier period, there is no direct way to measure price relative and so these new, higher priced products will not have a direct effect on the price index. This implies that they are a hidden cost of inflation.

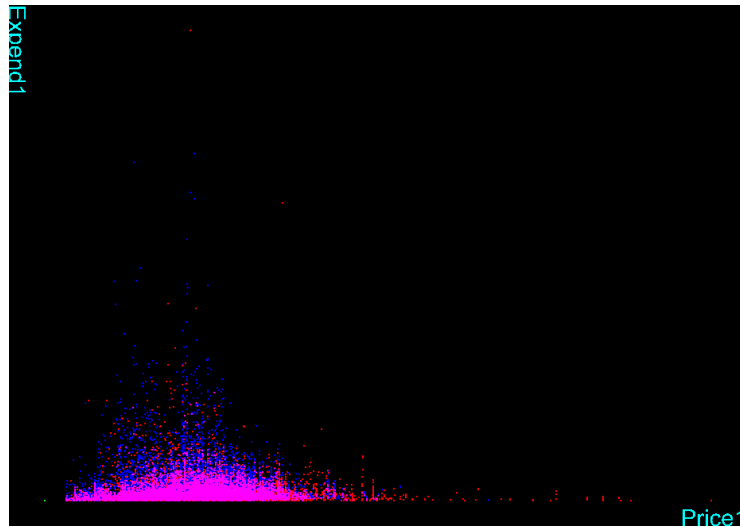


Figure 12. The vertical axis is Expenditures in Year 1, the horizontal axis is Prices in Year 1.

Finally, consider Figure 12 which is a plot of expenditures in year 1 versus price in year 1. Expenditures are the product of price and sales volume and hence essentially linearly related to sales volume. In Figure 12, the red items are the discontinued items. As might be expected high priced items attracting low levels of consumer expenditures are items that are likely to be discontinued. However, Figure 12 also illustrates that some moderately priced items that are attracting relatively high levels of consumer expenditures are also discontinued. This could, of course, be because the raw materials for these items are no longer available. More likely however is the conjecture that these items were replaced by substantially similar items possibly introduced at higher prices. This again is an effect of churning that suggests that churning is not limited to marginal items with low consumer expenditures.

5. Conclusions

This exploratory analysis of the cereal data suggests several important potential impacts on the computation of the consumer price index. One dramatic and persistent effect we have discovered is that aggressive promotion greatly enhances sales volume and total consumer expenditure. It also tends to depress prices. Because scanner-based indices can be sensitive to small price relatives and large relative expenditures, use of scanner data potentially yields price index results which are substantially different from those obtained through traditional survey methods. This may or may not be a disadvantage.

Scanner data would tend to track actual expenditures more closely than would data derived by survey-based methods. However, because of the potential increased volatility of scanner-based indices, issues of time reversal and chaining become more important.

It is also clear that long-term effects of product churning lead to many items with price relatives either 0 or ∞ , which consequently must be excluded from the price index computation. This of course is true for survey-based data as well as scanner data. However, with the massive datasets generated by scanner data, anomalous data are easier to overlook and hence have a potentially disproportional impact on the price index computation. This is especially true if price relatives are not exactly zero, but very close to zero because of roundoff or coding errors. In principal, the move to scanner data would appear to be an obviously desirable move. However, as illustrated by this data analysis and discussion, there are a number of subtle issues that suggest caution.

Acknowledgment

The work of Dr. Wegman was conducted initially while he was an ASA/NSF/BLS Senior Faculty Fellow at the Bureau of Labor Statistics. Additional work was completed while Dr. Wegman was under contract to the Bureau of Labor Statistics. We express our thanks to Dr. Janice Lent and to Dr. Sylvia Leaver for insightful discussions.

References

Lent, Janice (2001) "A note on the effect of extreme price values on price indexes," unpublished manuscript.

Wegman, E. J. (1990) "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, 85, 664-675

Wegman, E. J. and Solka, J. L. (2002) "On the mathematics of visualizing high dimensional data," to appear *Sankhya (A)*