Using USDA Fresh Fruit and Vegetable Arrivals to Determine the Distribution of a State's Production

Richard Beilock and Kenneth M. Portier

This paper examines the problem of transforming information on fresh fruit and vegetable Arrivals to U.S. Metropolitan Statistical Areas into a distribution of products to larger geographical regions. Three methods for the regional distribution of Florida-grown produce are compared. A new method, which takes into account regional population sizes in the allocation of an area's produce to the region, is shown to produce allocations similar to those obtained through trucker surveys. Finally, the new approach is applied to produce from other areas, and allocations to regions compared to that obtained using the Arrivals data only.

Introduction

For almost 70 years the USDA has collected data on volumes, origins, and transport modes of produce arriving in selected Metropolitan Statistical Areas (MSA's) in the US and Canada (Table 1). The data are published in *Fresh Fruit and Vegetable Arrivals in Western Cities* (USDA, 1985b–1987b) and *Fresh Fruit and Vegetable Arrivals in Eastern Cities* (USDA, 1985a–1987a). There are two potentially important uses for these data, hereafter called "Arrivals." The first is in tracking the volume of a commodity arriving at one or more specific locations. For example, California grape growers might use Arrivals information to evaluate the effectiveness of advertising in a specific market. The second use is in determining the distribution of produce from an origin. For example, Michigan apple growers might wish to know how much of their product is sold in different regions and how this compares to their competitors. One limitation faced in using the data for these purposes is uncertainty regarding the best method to transform information on Arrivals into data on the distribution of a product to all points.[1]

This paper examines this problem and compares two alternative approaches (denoted the "Direct" and "Indirect" approaches) for generating distribution estimates with information from interviews with produce truckers as they exited the Florida Peninsula (denoted the "Objective" approach). In this paper, possible problems associated with estimating a distribution of produce shipments from Arrivals are discussed. Florida produce is used as a case study for illustrative purposes.

Some Considerations in Using Arrivals Data to Develop a Distribution

The 27 Arrival cities used by the USDA, do not represent uniform coverage of the populations in their respective regions. These cities account for 21%, 45%, 25% and 35% of the total population for South, Northeast, Lake and West regions respectively (see Table 1).[2] If comparable proportions of the produce arriving in each city stay in that city, the proportion of all produce covered by Arrivals data would differ markedly across the regions.

The study was funded in part by the ERS, USDA and AMS, USDA. The authors acknowledge the assistance of Larry Summers and Richard Overheim of AMS, USDA. The authors are responsible for any remaining inaccuracies. Florida Experiment Station Journal Series No. 9631.

[1] The total volume of reported arrivals is roughly equivalent to 50% that of the total interstate produce shipments (USDA 1986c and 1987c). This comparison is often made to gauge the degree of coverage for arrivals. However intrastate movements are not reported in the Shipments data, while intrastate Arrivals are. For example, California produce arriving at San Francisco/Oakland or Los Angeles would be counted in the Arrivals data, but not in Shipments.

[2] The regions are defined as follows:

Northeast—the New England states, New York, New Jersey, Pennsylvania, Delaware, the Maritime Provinces, Quebec, and Ontario.

South—Maryland, Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi, Tennessee, Louisiana, and Arkansas.

Lake—West Virginia, Ohio, Indiana, Michigan, Kentucky, Illinois, Wisconsin, Missouri, Iowa, and Minnesota.

West— Texas, Oklahoma, Kansas, Nebraska, South Dakota, North Dakota, New Mexico, Colorado, Wyoming, Montana, Arizona, Utah, Idaho, California, Nevada, Oregon, Washington, Manitoba, Saskatchewan, Alberta, British Columbia, Northwest Territories, and the Yukon.

Table 1. Arrivals by City and Per Capita Arrivals by City, 1986.

City	Arrivals (1,000 cwt)	Population (1,000,000)	Per Capita Arrivals	Percent[a] Coverage	Adjusted Per Capita[b] Arrivals
Atlanta	10,000	2.3	4.34	90	4.82
Baltimore–Wash.	10,829	5.6	1.93	90	2.15
Columbia	3,173	.4	7.41	90	8.24
New Orleans	3,261	1.3	2.48	85	2.92
SOUTH TOTAL	27,263	9.7	2.82	NA	NA
Boston	21,922	2.8	7.82	95	8.23
Buffalo	3,362	1.0	3.36	80	4.19
N.Y.–Newark	35,696	17.7	2.01	85	2.37
Philadelphia	12,669	5.7	2.21	90	2.45
Pittsburgh	7,684	2.4	3.20	90	3.56
Montreal	15,552	1.0	15.28	NA	NA
Ottawa	3,098	.3	10.16	NA	NA
NORTHEAST TOTAL	118,449	31.6	3.75	NA	NA
Chicago	20,830	6.1	3.40	90	3.78
Cincinnati	8,065	1.4	5.70	90	6.38
Detroit	7,292	4.6	1.58	70	2.26
St. Louis	7,436	2.4	3.10	95	3.27
LAKE TOTAL	43,623	14.5	3.00	NA	NA
Dallas	8,948	2.2	4.12	80	5.15
Denver	4,800	1.6	3.07	80	3.84
Los Angeles	37,938	12.2	3.11	85	3.66
S. Fran.–Oak.	19,300	5.6	3.43	90	3.81
Seattle–Tacoma	8,986	2.2	4.11	90	4.57
Vancouver	7,765	.4	18.76	NA	NA
Winnipeg	3,163	.5	5.61	NA	NA
WEST TOTAL	90,900	24.7	3.68	NA	NA
US/CANADA TOTAL	280,235	80.4	3.48	NA	NA

[a]Estimate of percent of city's arrivals recorded by USDA. The estimate in by the USDA Officer in Charge.
[b]Adjusted for the estimate described in Note a.
Source: USDA 1987a & b and US Bureau of the Census.

Next, appreciable amounts of produce reported in the Arrivals data are trans-shipped to points outside these MSA's. Per capita arrival volumes suggest that the extent of this trans-shipment differs across MSA's. On average for each city, 348 pounds of Arrivals per capita were reported in 1986 (Table 1). For U.S. MSA's this varied from 782 pounds per capita for Boston to 158 pounds per capita for Detroit. If the data is adjusted for USDA estimates of the percent of all Arrivals covered in the reports, per capita Arrivals varies from 824 pounds in Columbia, S.C. to 237 pounds for New York-Newark. The per capita estimates across the Canadian cities are also quite variable. They tend to be higher than for the U.S., suggesting that the Canadian city population data does not encompass entire metropolitan regions to the extent that U.S. MSA population data does.

The salient point, however, is that there are large variations in per capita Arrivals across cities in both Canada and the U.S. This indicates the existence of correspondingly large variations in per capita consumptions or in reshipment rates or both. However, it does not seem likely to the authors that consumption differences can explain a major share of the differences in per capita Arrivals. For example, there are no apparent reasons for hypothesizing differences in per capita produce consumption between New York-Newark and Boston, yet the per capita Arrivals for the latter are nearly four times that of the former (Table 1). It seems more likely that there are significant differences across cities in the percentages of reported Arrivals that are trans-shipped.

The cities at which Arrivals data are gathered act as distribution points. Therefore, it seems likely that the total volumes of Arrivals in these cities are larger, per capita, than for other locations. If this is true, then an expansion of the Arrivals data by the ratio of the regional population to the Arrivals

city populations would overestimate the total volume of produce shipped to points in the region. If the percentages that those overestimates represent of the total are consistent across regions, then they could be employed to estimate the distribution of Arrivals across regions. The likelihood of this occurring, however, appears remote.

There are no data available to determine if cities are representative of their regions with respect to the mix of Arrivals (i.e., the types and origins of the produce received). However, there are no apparent reasons for assuming that there should be such differences. Indeed, it will be argued below that a reasonable approach would be to use Arrivals to determine the mix, but not the level, of Arrivals in a region.

As previously noted, the USDA is aware that the Arrivals data do not cover all Arrivals in each city and that the percentages missed in each city are not comparable. For the U.S. cities, the USDA publishes the "Officers in Charge" estimates of the percent of Arrivals captured in the reports (Table 1). With the exceptions of Boston, St. Louis, and Detroit, all of these estimates fall between 80% and 90%. Therefore, the percentages of all Arrivals in the reports appear to be reasonably consistent across cities.

While there is no way to confirm that the characteristics of the missed Arrivals are the same as for those which are covered in the USDA reports, there are no apparent reasons for suspecting that there are major differences.

Approaches for Generating Interregional Distributions

The most straightforward methodology for generating an interregional distribution from the Arrivals data is to equate the proportion of all Arrivals from an origin point (usually a state) reported by the USDA that go to a region with the proportion of all shipments from the origin that go to the region. This will be referred to as the Direct Approach and is defined in equation 1:

$$(1) \qquad X_{itj}^{(D)} = \text{estimated proportion of shipment from origin i to region j at time t} = A_{itj} \Big/ \sum_{j=1}^{n} A_{itj}$$

where: A_{itj} = reported Arrivals in region j from origin i at time t, and

n = number of regions.

The preceding discussion, however, raises serious concerns regarding the direct use of Arrivals data for generating interregional distributions. Yet, this is almost always what has been done (see for example, Auburn and Sperling, Dow, Pavlovic et al., and Manalytics). However, it is reasonable to use Arrivals for information regarding the market share in a Region of produce from an Origin. As argued above, due to differing intensities of coverage across regions by the USDA and differences across cities and regions regarding apparent reshipment rates, the validity of expanding the volume estimates generated by the USDA Arrivals is highly questionable. However, there are no apparent reasons for assuming that the points covered in the Arrivals are atypical with respect to the mix of commodities and origins for those commodities.

A simplifying assumption is made that per capita consumption is consistent across Regions. Based upon this, the Indirect Approach is defined in equation 2:

$$(2) \qquad X_{itj}^{(I)} = B_{itj} \Big/ \sum_{j=1}^{n} B_{itj}$$

where:

$$B_{itj} = (A_{itj} \times P_j) \Big/ \sum_{j=1}^{n} A_{itj}, \text{ and}$$

P_j = Population in region j.

Again, this approach only depends upon the Arrivals data for market share information (the RHS term in parentheses in equation 2). It is based upon the philosophy that Arrivals are more properly viewed as a sampling to establish market shares, rather than as a direct barometer of the levels of all produce shipments.

As the measure of comparison, consider survey data regarding the destinations of produce from a specific origin (Florida). The distributions derived from these data (the Objective Approach) are assumed to be reasonably accurate approximations of the actual distribution. This is based upon the implicit assumption that the distribution of shipments not captured by the survey does not differ from those in the survey. The derivation for this distribution is presented in equation 3:

$$(3) \qquad X_{itj}^{(O)} = s_{itj} \Big/ \sum_{j=1}^{n} s_{itj}$$

where s_{itj} = volume shipped to region j from origin i at time period t from the trucker survey.

Data

The principal data sources employed in this study are the Arrivals data from USDA (1985a–1987a and 1985b–1987b), and survey data from an ongoing project described in Beilock, MacDonald, and Powers (1988). The sample period is from November, 1984, through November, 1987. Within that period, data were collected for January, March, May/June, and November; a total of 9 points in time. The Origin is Florida, and the Regions are as defined in Footnote 2. For the truck Arrivals data, both Arrivals from Florida and total Arrivals from all origins were collected.

The survey data consist of interviews with truckers at the Florida Agricultural Inspection Stations (FAIS). These stations are arrayed along the 16 possible routes into the Florida Peninsula. The stations are always open and all trucks must stop to be inspected. It should be noted that the FAIS system does not cover movements from the Florida Panhandle, save for shipments routed through Northeast Florida. Produce production in the Florida Panhandle is insignificant relative to Peninsular Florida. In each survey month, interviews were conducted for two consecutive days and from 6:00 PM to 1:00 AM at the FAIS stations on US I-10, US I-75, and US I-95. These routes account for between 80% and 90% of all produce shipped from the Florida Peninsula and the hours selected coincide with the highest flows. Across the thirteen survey periods, 5,568 interviews were completed.[3] Finally, the three interstates provide good geographic coverage.

Refusal rates normally were less than 5%, and often as low as 2%. There are no apparent reasons to suspect that those refusing differ from the participants with respect to commodity mix or destinations. Nor are there apparent reasons for suspecting differences from those passing through other FAIS stations and for those passing the stations on other days of the same month. Therefore, the interregional distributions derived from the survey data are assumed to be reasonably accurate approximations of the actual distributions.

The survey data include shipments bound for all areas of the U.S. (except the Florida Peninsula) and Canada. The population of the Florida Peninsula was excluded from the total for the South for calculating the Indirect Approach estimate facilitating the comparison of the Indirect and Objective Approaches. However, it could be argued that it is improper to compare the Indirect, Objective, Direct Approaches as only the last includes (implicitly) the Florida Peninsula. That is, even if the three approaches were identical, the inclusion of the Direct Approach should assign larger shares of produce to the South due to that approach's inclusion of the Florida Peninsula. In the results, however, the Direct method always assigned the lowest shares to the South, indicating that if the Indirect and Objective Approaches had included the Florida Peninsula, the differences between them and the Direct Approach would have been even greater.

Three produce groupings, accounting for over 90% of all produce truck movements from Florida, were identified for the analysis. They are:

CITRUS	Includes oranges, grapefruit, tangerines, tangelos, limes and lemons. In 1985, this grouping accounted for 31% of all produce shipments from the state by truck (USDA, 1986c).
TOMATOES	Includes tomatoes and cherry tomatoes. In 1985 this grouping accounted for 18% of all produce shipments from the state by truck (USDA, 1986c).
MIX	Includes snap beans, peppers, strawberries, celery, sweet corn, cucumbers, lettuce, squash, potatoes, and watermelons. In 1985, this grouping accounted for 43% of all produce shipments from the state by truck (USDA, 1986c).
ALL	Includes all produce truck Arrivals.

Methodology for Comparing Approaches

Multivariate tests

The comparison of the three approaches to generating interregional distribution is done initially using all aspects of the multivariate proportion vectors.

[3] The number of interviews in each survey period were:

Year	January	March	May/June	November
1984	—	—	—	410
1985	445	280	603	361
1986	445	403	386	471
1987	448	469	554	293

$$\underset{\sim}{X}_{it}^{(D)} = \left\{X_{it1}^{(D)}, X_{it2}^{(D)}, X_{it3}^{(D)}, \ldots, X_{itn}^{(D)}\right\} \quad t = 1, \ldots, T_D$$

$$\underset{\sim}{X}_{it}^{(I)} = \left\{X_{it1}^{(I)}, X_{it2}^{(I)}, X_{it3}^{(I)}, \ldots, X_{itn}^{(I)}\right\} \quad t = 1, \ldots, T_I$$

$$\underset{\sim}{X}_{it}^{(O)} = \left\{X_{it1}^{(O)}, X_{it2}^{(O)}, X_{it3}^{(O)}, \ldots, X_{itn}^{(O)}\right\} \quad t = 1, \ldots, T_O$$

Comparison here is equivalent to testing if the three approaches produce vectors that come from the same multivariate distribution. Usually such tests of multivariate distributions are performed using a generalized likelihood ratio statistic and assuming multivariate normality (Anderson, 1958).

Because the proportions in the vectors sum to one, a strong dependency among the components exists. This sum-to-one constraint on the proportion vector precludes use of a normal distribution and hence use of standard multivariate tests for comparing these vectors. Aitchison (1986) argues that such comparisons are more appropriately performed on a transformed data vector of the form

$$\underset{\sim}{Z}_{it}^{(D)} = \left\{Z_{it1}^{(D)} = \log_e\left[\frac{X_{it1}^{(D)}}{X_{itJ}^{(D)}}\right], \ldots \right.$$

$$Z_{itJ-1}^{(D)} = \log_e\left[\frac{X_{itJ-1}^{(D)}}{X_{itJ}^{(D)}}\right],$$

$$Z_{itJ+1}^{(D)} = \log_e\left[\frac{X_{itJ+1}^{(D)}}{X_{itJ}^{(D)}}\right], \ldots$$

$$\left. Z_{itn}^{(D)} = \log_e\left[\frac{X_{itn}^{(D)}}{X_{itJ}^{(D)}}\right]\right\}$$

with $\underset{\sim}{Z}_{it}^{(I)}$ and $\underset{\sim}{Z}_{it}^{(O)}$ defined similarly. This transformation reduces the dimension of the vectors to $n-1$, and eliminates the effect of the sum-to-one constraint. These data are also closer to being normally distributed. The choice of the component to use in the denominator of the ratio, i.e., $X_{itJ}^{(D)}$, is not critical to the analysis, but usually a component that is not close to zero is used.

The strategy for testing for differences in multivariate populations begins with the specification of the most complex model that could explain the data. This model usually has the largest number of parameters. The appropriateness of models of lesser complexity are tested by comparing the fit of this model to the fit of the most complex model. This comparison can, and is, usually performed using a generalized likelihood ratio test (Anderson, 1958). A series of hypothesized models can be examined in this manner until a model is found that cannot be further simplified.

The most complex model considered for the log ratio data vectors has each approach vector being multivariate normal with different mean vectors and variance-covariance matrices. Let $\underset{\sim}{\mu}_i^{(D)}$, $\underset{\sim}{\mu}_i^{(I)}$, $\underset{\sim}{\mu}_i^{(O)}$ represent the mean vectors and $\Sigma_i^{(D)}$, $\Sigma_i^{(I)}$, $\Sigma_i^{(O)}$ represent the variance-covariance matrices for the three approaches. The most complex model assumes

$$H_O: \underset{\sim}{\mu}_i^{(D)} \neq \underset{\sim}{\mu}_i^{(I)} \neq \underset{\sim}{\mu}_i^{(O)} \text{ and } \Sigma_i^{(D)} \neq \Sigma_i^{(I)} \neq \Sigma_i^{(O)}$$

where two of the means may be equal but not all three, and/or two of the correlation matrices may be equal but not all three. Less complex models are

$$H_1: \underset{\sim}{\mu}_i^{(D)} = \underset{\sim}{\mu}_i^{(I)} = \underset{\sim}{\mu}_i^{(O)}$$

and variance matrices are different,

$$H_2: \underset{\sim}{\Sigma}_i^{(D)} = \underset{\sim}{\Sigma}_i^{(I)} = \underset{\sim}{\Sigma}_i^{(O)}$$

and mean vectors are different,

and

$$H_3: \underset{\sim}{\mu}_i^{(D)} = \underset{\sim}{\mu}_i^{(I)} = \underset{\sim}{\mu}_i^{(O)} \text{ and } \underset{\sim}{\Sigma}_i^{(D)} = \underset{\sim}{\Sigma}_i^{(I)} = \underset{\sim}{\Sigma}_i^{(O)}.$$

Maximum likelihood estimation of $\underset{\sim}{\mu}_i^{(D)}$, $\underset{\sim}{\mu}_i^{(I)}$, $\underset{\sim}{\mu}_i^{(O)}$, $\underset{\sim}{\Sigma}_i^{(D)}$, $\underset{\sim}{\Sigma}_i^{(I)}$, $\underset{\sim}{\Sigma}_i^{(O)}$ under the full model and under the three alternative models follows Mardia, Kent, and Biddy (1979, Section 5.5.3). In matrix notation these are:

The mean estimate

$$\underset{\sim}{m}_i^{(D)} = \frac{1}{T_D}\sum_{t=1}^{T_D} \underset{\sim}{Z}_{it}^{(D)};$$

$$\underset{\sim}{m}_i^{(I)} = \frac{1}{T_I}\sum_{t=1}^{T_I} \underset{\sim}{Z}_{it}^{(I)}; \quad \underset{\sim}{m}_i^{(O)} = \frac{1}{T_S}\sum_{t=1}^{T_O} \underset{\sim}{Z}_{it}^{(O)};$$

The separate variance-covariance estimates,

$$S_i^{(D)} = \frac{1}{T_D}\sum_{t=1}^{T_D} \left[\underset{\sim}{Z}_{it}^{(D)} - \underset{\sim}{m}_i^{(D)}\right]\left[\underset{\sim}{Z}_{it}^{(D)} - \underset{\sim}{m}_i^{(D)}\right]',$$

$$S_i^{(I)} = \frac{1}{T_I}\sum_{t=1}^{T_I}\left[\underset{\sim}{Z}_{it}^{(I)} - \underset{\sim}{m}_i^{(I)}\right]\left[\underset{\sim}{Z}_{it}^{(I)} - \underset{\sim}{m}_i^{(I)}\right]',$$

$$S_i^{(O)} = \frac{1}{T_O}\sum_{t=1}^{T_O}\left[\underset{\sim}{Z}_{it}^{(O)} - \underset{\sim}{m}_i^{(O)}\right]\left[\underset{\sim}{Z}_{it}^{(O)} - \underset{\sim}{m}_i^{(O)}\right]',$$

The pooled variance-covariance estimate,

$$S_i^{(P)} = (T_D + T_I + T_O)^{-1}(T_D S_i^{(D)} + T_I S^{(I)} + T_O S^{(O)})$$

and the combined sample mean and variance-covariance estimates,

$$m_i^{(c)} = (T_D + T_I + T_O)^{-1}(T_D m_i^{(D)} + T_I m_i^{(I)} + T_O m_i^{(O)})$$

and

$$S_i^{(c)} = (T_D + T_I + T_O)^{-1}\left[\sum_{t=1}^{T_D}\left(Z_{it}^{(D)} - m_i^{(c)}\right)\left(Z_{it}^{(D)} - m_i^{(c)}\right)' + \sum_{t=1}^{T_I}\left(Z_{it}^{(I)} - m_i^{(c)}\right)\left(Z_{it}^{(I)} - m_i^{(c)}\right)' + \sum_{t=1}^{T_O}\left(Z_{it}^{(O)} - m_i^{(c)}\right)\left(Z_{it}^{(O)} - m_i^{(c)}\right)'\right]$$

The problem of testing H_1 is the multivariate version of the awkward Behrens-Fisher problem. No explicit form for the maximum likelihood estimate exists but the following simple iterative procedure will result in the appropriate estimate.

1. Let $S_i^{(DH)} = S_i^{(D)}$, $S_i^{(OH)} = S_i^{(O)}$, and $S_i^{(IH)} = S_i^{(I)}$.
2. Compute

$$m_i^{(H)} = \left(T_I S^{(IH)^{-1}} + T_D S_i^{(D)^{-1}} + T_O S^{(OH)^{-1}}\right)^{-1}\left(T_I S^{(IH)^{-1}} m_i^{(I)} + T_D S^{(DH)^{-1}} m_i^{(D)} + T_O S^{(OH)^{-1}} m_i^{(O)}\right)$$

3. Compute new estimates for $S_i^{(DH)}$, $S_i^{(IH)}$, $S_i^{(OH)}$ using

$$S_i^{(OH)} = S_i^{(D)} + (m_i^{(D)} - m_i^{(H)})(m_i^{(D)} - m_i^{(H)})'.$$

4. Repeat steps 2 and 3 until convergence. Convergence usually occurs after seven to ten iterations.

To test the model H_1, the test statistic is

$$T_D\log_e(|S_i^{(DH)}|/|S_i^{(D)}|) + T_I\log_e(|S_i^{(IH)}|/|S_i^{(I)}|) + T_O\log_e(|S_i^{(OH)}|/|S_i^{(O)}|)$$

which is compared to a chi-square distribution with $n-1$ degrees of freedom.

To test the model H_2, the test statistic is

$$T_D\log_e(|S_i^{(P)}|/|S_i^{(D)}| + T_I\log_e(|S_i^{(P)}|/|S_i^{(I)}|) + T_O\log_e(|S_i^{(P)}|/|S_i^{(O)}|)$$

which is compared to a chi-square distribution with $n(n-1)$ degrees of freedom.

To test the model H_3, the test statistic is

$$T_D\log_e(|S_i^{(c)}|/|S_i^{(D)}|) + T_I\log_e)|S_i^{(c)}|/|S_i^{(I)}|) + T_O\log_e(|S_i^{(c)}|/|S_i^{(O)}|)$$

which is compared to a chi-square distribution with $(n+2)(n-1)$ degrees of freedom.

If all three hypothesized models are rejected, then the most complex model, H_O, remains the best explanation of the data. Failure to reject a model indicates its appropriateness in explaining the data.

Univariate tests

The multivariate analysis approach provides no easy way to examine the regional differences between the methods in detail nor whether year-to-year or season-to-season differences exist. For this more detailed analysis, standard analysis-of-variance models were applied to the proportions as well as to the arcsine-square-root transformed proportions.

With each region being examined separately the sum-to-one constraint is not directly accounted for. Because of this equality, significant differences in the three approaches for one region will also show up as significant differences in one or more of the other regions.

Post-hoc comparison of estimated mean percentages assigned by the different approaches will be performed using the Waller-Duncan Bayesian k-ratio (LSD) procedure with k-ratio = 100 (Chew, 1972).

Table 2. Results of Multivariate Analysis of Method Differences Using Log Rates Transformed Compositional Data.

	MODEL		
	H_1 $\mu^{(D)} = \mu^{(I)} = \mu^{(O)}$ $\Sigma^{(D)} \neq \Sigma^{(I)} \neq \Sigma^{(O)}$	H_2 $\mu^{(D)} \neq \mu^{(I)} \neq \mu^{(O)}$ $\Sigma^{(D)} = \Sigma^{(I)} = \Sigma^{(O)}$	H_3 $\mu^{(D)} = \mu^{(I)} = \mu^{(O)}$ $\Sigma^{(D)} = \Sigma^{(I)} = \Sigma^{(O)}$
CITRUS	42.79 (1) 12 (2) <.001 (3)	63.84 3 <.001	115.64 18 <.001
TOMATOES	62.50 12 <.001	107.95 3 <.001	190.46 18 <.001
MIX	53.11 12 <.001	24.95 3 <.001	85.03 18 <.001
ALL	65.04 12 <.001	14.44 3 <.002	122.21 18 <.001

(1) log likelihood value
(2) degrees of freedom
(3) significance probability

Results

Results for the test of the different models are given in Table 2. It is clear that there are significant differences between at least two of the three approaches in mean vectors and covariance matrices. Differences in distribution of total produce by region and method are illustrated in Figure 1.

In Figure 2, a three-dimensional representation of the proportion of all vegetables for the three estimation approaches is presented. The four corners of this tetrahedron represent a pure composition, that is, a point near the corner labeled South has a very high proportion of the produce going to the South. The central point of the tetrahedron is the composition {.25, .25, .25, .25} which allocates equally to all four regions. Because very little of Florida produce goes to the Western region, there is very little variability in the West fraction. The tetrahedron is orientated to best present the variability among the regions, and hence the West node is presented at the rear of the image. As can be seen from this diagram, the Direct approach allocates more produce to the Northeast and South, whereas the Indirect and Objective methods allo-

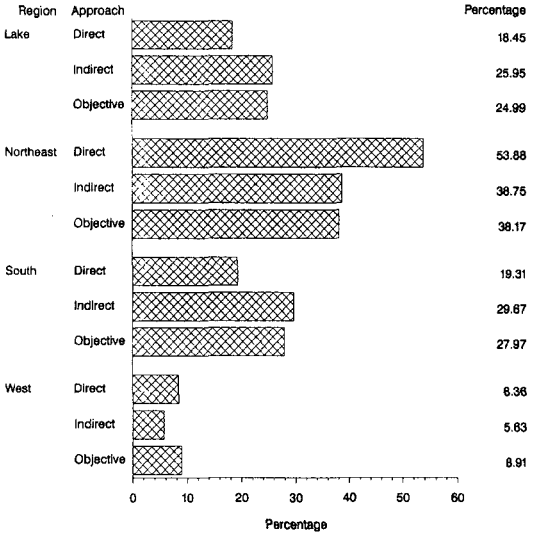

Figure 1. Distribution of all Produce from Florida 1984–1987.

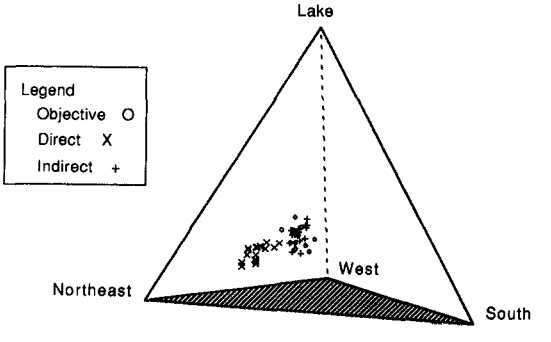

Figure 2. Three-dimensional representation of the proportion of All vegetables for the three estimation approaches.

cate more to the Lake region. The Direct approach also produces less dispersed estimates than do the other methods, one reason the multivariate analysis rejects the model of equal variance-covariance matrices.

In Table 3, the probabilities of significant differences among the three approaches after adjusting for year and season effects are presented. Notice that in every case there is very strong evidence ($p<.001$) of estimation approach differences. Only occasionally are year effects significant, but seasonal differences are almost always present. The analysis results reported here are on the raw proportions, but the conclusions are the same if the arcsine square-root transformation data are used.

The mean percentage assigned to each region by the three estimation approaches are given in Table 4 along with an indication of which means are statistically different.

In the Southern region, the Direct approach tends to estimate a lower proportion than either the Indirect or Objective approaches. The Indirect and Objective approaches are very similar for all four crop categories and are not statistically different for Citrus and Tomatoes.

For the Northeast, the Direct approach produces estimates that are statistically higher, from 9% to 15%, than those of the Indirect and Objective approaches. For all four crops the Indirect and Objective approaches are statistically equivalent.

For the Lake region, the Direct approach tends to produce lower estimates than the Indirect and Objective approach. For tomatoes, the Indirect approach estimates are significantly higher than the other two methods. Due primarily to high variability, no differences between the methods can be shown for Mixed Vegetables even though the Direct approach is nearly 8% below the Objective method.

For the Western region, the Indirect approach estimates significantly lower proportions than the other two methods although for Citrus the three methods are very close. This is due primarily to the smaller populations in the Western U.S. which force the Indirect approach to allocate less of the produce to this region.

Application of the Direct and Indirect Approaches to Other Origin Regions

Distributions across four destination regions using both the Direct and Indirect approaches were cal-

Table 3. Probabilities of Significant Differences for Year, Time Period, and the Three Estimation Approaches[a] as Sources of Variation.

Commodity	Time[b] Period	Year	Estimation Approach[a]	Mean[c] Square Error
SOUTH				
Citrus	.132	.386	<.001	.0007
Tomatoes	.004	.967	<.001	.0012
Mix	.021	.402	<.001	.0017
All	.022	.097	<.001	.0006
NORTHEAST				
Citrus	.054	<.001	<.001	.0017
Tomatoes	<.001	.284	.002	.0049
Mix	<.001	.241	<.001	.0014
All	<.001	<.001	<.001	.0005
LAKE				
Citrus	.490	<.001	<.001	.0010
Tomatoes	.068	.315	<.001	.0016
Mix	<.001	.041	<.001	.0008
All	<.001	.003	<.001	.0004
WEST				
Citrus	<.001	.137	.026	.0005
Tomatoes	<.001	.251	.019	.0025
Mix	<.001	.668	.001	.0003
All	.540	.086	<.001	.0002

[a]The Direct, Indirect, and Objective approaches described in the text.
[b]January, March, May/June, and November.
[c]Based on untransformed proportion data.

Table 4. Average Percentage Assigned to Each Region by the Three Estimation Approaches and Statistical Differences Among Them.

Commodity	Approach		
	Direct	Indirect	Objective
	1	2	3
SOUTH			
Citrus	16.49	27.67(a)	25.66(a)
Tomatoes	18.58	24.94(a)	24.29(a)
Mix	20.36	31.42	27.72
All	19.31	29.67	27.97
NORTHEAST			
Citrus	54.07	36.62(a)	39.23(a)
Tomatoes	46.77	37.93(a)	38.96(a)
Mix	58.24	43.65(a)	42.31(a)
All	53.88	38.75(a)	38.17(a)
LAKE			
Citrus	22.08	30.59(a)	28.79(a)
Tomatoes	19.25(a)	26.39	21.73(a)
Mix	15.61	21.03	23.58
All	18.45	25.95(a)	24.99(a)
WEST			
Citrus	7.36(b)	5.12(a)	6.31(a,b)
Tomatoes	15.40(a)	10.73	15.02(a)
Mix	5.79(a)	3.89	6.39(a)
All	8.36(a)	5.63	8.91(a)

NOTE: The "(a)" connotes that the methods are not statistically different from one another for the region/commodity represented by that row, using Waller-Duncan Bayesian LSD procedure.

culated employing 1987 Arrivals data, for the following origins and commodities:

ORIGIN	COMMODITIES			
CALIFORNIA	All Produce	Lettuce	Oranges	Grapes
PACIFIC NORTHWEST[1]	All Produce	Potatoes	Apples	Onions
TEXAS	All Produce	Onions	Watermelons	Cabbage
NORTHEAST[2]	All Produce	Potatoes	Apples	Onions

[1]Washington, Oregon, and Idaho. [2]Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, Pennsylvania, New Jersey, Delaware, Maryland, and West Virginia.

The commodities selected for each origin were the three with the highest interstate movement volumes (USDA 1987b). Including Florida, these five origins account for nearly 80% of all interstate movement (USDA 1987b) of agricultural produce.

As expected, the results between the two approaches differ considerably (Table 5). For California, Pacific Northwest, and to a lesser extent, Texas the main difference is a reallocation of produce in the Indirect approach away from the West in favor of the three other regions. For example, employing the Direct approach, 58.2% of the West's truck produce movements are to destinations in the West. Using the Indirect method, the share going to the West drops by a fifth to 46.8%, the South and Northeast gain slightly, while the share going to Lake increases by two thirds from 14.5% to 24.3%. Considering the populations of these regions, and that California is known to market large volumes east of the Mississippi, the distribution using the Indirect approach seems much more plausible than that for the Direct approach.

By either approach, the majority of the Northeast produce is allocated to the Northeast, and most of the remainder to South. For Apples, Onions, and All Produce from the Northeast, the Indirect approach allocates less to South than does the Direct approach. For All Produce, this is primarily the result of reallocations to Lake (4.1% versus 2.7%). For Apples and Onions the reallocation from South

Table 5. Distributions of Produce from California, Pacific Northwest, Texas, and Northeast via the Direct and Indirect Approaches, 1987[1]

	PERCENT TO							
	NORTHEAST		SOUTH		LAKE		WEST	
ORIGIN/COMMODITY	DIR	IND	DIR	IND	DIR	IND	DIR	IND[2]
CALIFORNIA								
Grapes	19.3	18.5	22.8	18.9	17.6	27.2	40.3	35.4
Lettuce	14.2	14.2	19.9	19.4	22.3	30.6	43.6	35.7
Oranges	15.4	23.5	18.7	16.5	16.2	24.3	49.7	35.8
All Produce	12.3	13.6	14.9	15.2	14.5	24.3	58.2	46.8
PACIFIC NORTHWEST								
Apples	12.6	9.8	13.1	9.5	17.1	34.5	57.2	46.2
Onions	9.0	11.4	12.1	10.4	15.6	30.7	63.3	47.4
Potatoes	6.1	8.5	4.9	8.9	14.1	22.4	74.9	60.2
All Produce	10.6	11.8	9.6	9.9	14.9	25.2	64.9	53.0
TEXAS								
Cabbage	15.5	12.6	23.9	16.3	35.9	50.4	24.8	20.8
Onions	31.0	33.3	29.7	21.5	20.0	33.1	19.4	12.2
Watermelons	30.2	26.3	10.2	8.2	8.6	19.4	51.0	46.1
All Produce	24.9	18.2	18.9	19.6	18.4	31.2	37.9	31.0
NORTHEAST								
Apples	78.7	79.8	21.3	20.2	0.0	0.0	0.0	0.0
Onions	82.6	86.4	15.8	11.1	1.6	2.5	0.0	0.0
Potatoes	85.6	82.1	14.2	17.6	.3	.3	0.0	0.0
All Produce	81.7	81.6	15.2	14.0	2.7	4.1	.4	.3

Notes: 1. See text for definitions of origins and Figure 1 for definitions of destination regions.
2. DIR and IND denote Direct and Indirect approaches, respectively.
Source: USDA (1987a)

is primarily in favor of the Northeast. The Indirect approach seems to correct for the much more intense sampling in the Northeast and lower intensity of sampling in the South relative to the other regions (as indicated by the proportions of each region's population in the Arrivals cities). Therefore, the Indirect approach should allocate relatively more to the Northeast, *ceteris paribus*. This seemingly perverse reallocation of Apples and Onions reflects low market shares for these products in the South. That is, the proportions of all Apples and all Onion Arrivals from the Northeast to the arrival cities in the South are sufficiently low relative to their shares in the Northeast to counteract the sampling intensity correction. An examination of allocations for these products from Texas and Pacific Northwest reveals the same pattern.

Summary and Conclusions

This paper has explored the proper use of USDA's Arrivals data in determining the distribution of produce from an origin to various regions of the U.S. and Canada. Two approaches were employed: the direct use of Arrivals, equating its distribution with that for all produce; and an indirect approach that relies upon Arrivals only for information regarding market shares. To the authors' knowledge, the former approach has always been used. However, the latter approach was hypothesized to be superior as it avoids problems related to differing percentages of regional populations accounted for by the cities used for Arrivals, less than universal coverage within those cities, and trans-shipments out of the cities.

Estimates of the distribution of produce from Florida to four regions of the U.S. and Canada using the two approaches were compared with the results obtained from interviews with truckers as they exited the Florida Peninsula. The results suggest that the proposed indirect approach is far superior. Assuming the survey data (Objective approach) represents the correct allocation of produce, the Direct approach misassigned 32% whereas the proposed Indirect approach only misassigned 7%. Comparison of the Direct and Indirect approaches was also made for major commodities from three other sources in the United States. In each case the Indirect approach provided a more plausible allocation of produce.

These findings imply that use of the indirect approach could enhance knowledge of interstate

movements of produce. In addition, if information regarding regional differences in per capital produce consumption were available, the method could be further improved.

References

Aitchison, J. *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York, 1986, p. 141–182.

Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958, p. 90.

Auburn, J. and D. Sperling. "Transportation Patterns for California Fresh Produce: An Initial Investigation" presented at the 1987 Meeting of the Transportation Research Forum.

Beilock, R., J. MacDonald, and N. Powers. *An Analysis of Product Transportation: A Florida Case Study*, ERS Agr. Econ. Report 579, 1988.

Chew, V. *Comparisons Among Treatment Means in an Analysis of Variance*, ARS/H/6, Agricultural Research Service, USDA, October, 1977.

Dow, K. *Transportation of Florida Perishables—Problems and Research Needs*, Food and Resource Economics Department Economic Information Report 106, University of Florida, 1979.

Manalytics, *A Long-Term Study of Produce Transportation: A Profile of Fresh Fruit and Vegetable Shipments From Selected Growing Areas, Volume 3* report prepared for the US Department of Transportation and the National Bureau of Standard, 1977.

Mardia, K.V., Kent, J. T. and Biddy, J. M. *Multivariate Analysis*, Academic Press, New York, 1979.

Pavlovic, K., D. Reaves, G. Long, and T. Maze. *Domestic Transportation for Florida Perishable Produce*, Transportation Research Center Report, Department of Civil Engineering, University of Florida, 1980.

Shoemyen, A., et al. *1985 Florida Statistical Abstract*, University Presses of Florida, Gainesville, Florida, 1985.

US Bureau of the Census, *Statistical Abstract of the United States, 1986*, Washington, 1986.

USDA, *Fresh Fruit and Vegetable Arrivals in Eastern Cities*, Agricultural Marketing Service, USDA, 1985a–1987a (annual).

USDA, *Fresh Fruit dnd Vegetable Arrivals in Western Cities*, Agricultural Marketing Service, USDA, 1985b–1987b.

USDA, *Fresh Fruit and Vegetable Shipments*, Agricultural Marketing Service, USDA, 1986c and 1987c.