A SURVEY DESIGN SYSTEM FOR THE MEASUREMENT OF TRUCK CARGO FLOWS IN PERU[1]

A. Satin and R. Ryan[2]

This paper describes a survey design established to measure truck commodity flows in Peru. The article addresses the conceptual and operational features of the survey design as well as describing its elements and implementation techniques in the context of a pilot project. Finally, the paper illustrates how the results of this pilot might be used to design and implement a full-scale national survey.

1. INTRODUCTION

This article documents elements of a proposed survey design system which will allow for the collection of data (product-origin-destination, i.e. P-O-D) on cargo movements by truck in Peru. The basic design was established during a three-week field trip to the Office of Sectorial Statistics within the Ministry of Transport and Communications in Lima, Peru. It was decided that the design would be tested in a pilot study during the summer of 1978. If successful, the design would serve as a potential base for the conduct of a full-scale, continuous national roadside survey. The techniques and procedures may also be adapted to similar P-O-D studies conducted at the regional level. Background material used in the development of the system is included in the papers [1], [2], [3], [4] and [5].

2. DATA SOURCES AND REQUIREMENTS

With respect to the development of a design to collect P-O-D information, three potential data sources were reviewed by the authors:

---

[1] This project was commissioned by the Canadian Transportation Commission as part of an overall technical assistance program for the Canadian International Development Agency in the transportation field.

[2] A. Satin, Household Surveys Development Division and R. Ryan, Special Surveys Co-ordination Division, Statistics Canada.

(a) the consignor - consignee (the shipper)
(b) the carrier company
(c) the driver.

Under alternative (a), one would be required to locate and obtain data from the shipper. There are many shippers without known addresses, especially farmers in the Sierra, many of whom sell their produce (even in the field before harvesting) to a middleman-agent-carrier; the middleman transports and markets the goods and the original producer may not know how much he had to ship or where it was going. For these reasons (a) is considered an unlikely source of reliable P-O-D information.

The carrier company (alternative (b)) is also viewed as an unlikely source. A high proportion of the carriers (more than 50% and perhaps as much as 85%) are one vehicle-one driver operations. The location of many of these small operators is unknown so they would be unavailable for mail or personal contact. In addition, it is felt that the records kept by these individuals are poor (i.e. virtually non-existent), so their movements and cargo over any but the immediately preceding period would be uncertain.

The remaining data source under alternative (c), namely the truck driver, is viewed for now as the best alternative. It is understood that they are an unco-operative group as a whole, viewing police and government as an undesirable intrusion into their private affairs. They are suspected of going to some lengths to avoid such questions as may be required to collect the data for a survey.

The advantage of the truck driver as the data source is that certain data is automatically available, namely the commodity, origin-destination and the vehicle description. This approach has the additional advantage that it has been used before, and its risks are known and, to some extent, the drivers are used to it.

One of the frequently mentioned "solutions" to the problems of data collection is the "Planica Unica" or waybill document (mandatory) used for all shipments with extra copies automatically provided for control and statistical purposes. It is apparently in successful use at this time in Colombia. Occasional planning for its use in Peru goes back at least to 1970. In this type of system, the shipper and carrier provide all data needed, prior to moving the goods, and copies of the documents are collected at various control points in the course of a trip.

The introduction of such a system of documents is likely to be an involved process, time consuming and costly. It is estimated that about four years were necessary before statistics started to flow from the Colombian model.

Such a system does have obvious advantages, but in light of the long lead time, and the lack of success in reaching agreement on it in Peru in the past, it would be more appropriate, for the present, to look for another solution.

Finally, the use of models might be viewed as having the potential to satisfy data requirements. The basic drawback here is the lack of reliable statistics (consumption of products, production of products) to render such an approach workable at this time.

In summary, it was concluded that, for the present, the only viable alternative for collecting product-origin-destination (P-O-D) statistics is by means of a roadside survey, the details of which are described in the following sections of this report.

With respect to data requirements, the most basic information required on a regular basis and likely to be available from a truck driver is:

(1) Commodity(ies): The code system suggested is the European Classification. Another system under consideration is the U.N. Commodity Classification.

(2) Origin(s) / Destination(s): A place name, associated with the specific area in the country, to be coded ultimately to the provincial (regional) level.

Only major flows will be measured as the measurement of minor flows would require that stations be placed along all road segments, which is not practical.

(3) Quantity: In metric tons, cubic metres, or units (where other measures are unavailable) to be converted to weight as appropriate.

(4) Vehicle: Licence plate, number of axles, type of body.

The basic unit of the study is a commodity trip. Thus, for each loading and/or unloading of cargo, a separate commodity origin-destination data set would be prepared. The vehicle trip would be defined by the point of first loading of cargo to the point of ultimate emptying of the vehicle, for the commodities on the truck at the survey point. Cargo collected and discharged prior to the interview is not to be considered, nor is cargo to be picked up and delivered later in the vehicle trip.

3. THE PILOT STUDY

In designing the frame for the pilot, two major areas of consideration rested with the determination of (1) the geographical or spatial coverage of the pilot, and (2) the temporal coverage associated with the sampling methodology. It was determined that both the pilot and the extension for the future would involve the measurement of road freight traffic for major flows in the country. The measurement of all flows would not be financially or operationally feasible or practical.

In terms of spatial coverage, it was important that the locations selected for the pilot and the associated survey points (control stations) simulate as much as possible the conditions that might be experienced in the rest of the country. In this way, all methodologies associated with the pilot could be realistically tested and evaluated in terms of the possibilities of extending to a national survey. The criteria used to select the area for the pilot were determined to be:

(1) high volume of traffic,
(2) a diversity of product mix,
(3) testing data collection methods under extremes in climatic conditions, and
(4) an area where external data sources can be used for validation.

Peruvian officials listed three areas for consideration; namely, (1) the Ancash area in the north, (2) the Lima-La Orolla route in the central area, and (3) the area around Pisco in the south.

The three areas were discussed in light of the selection criteria and it was determined that the region of Ancash around Chimbote would best meet the criteria and so was chosen for the pilot to be run during the course of a full trimester. A secondary objective of the pilot was to furnish data for the region, if possible. This would depend upon the extent to which the sample design, field procedures, etc. were to be modified and improved during the course of the pilot.

Five control points were established to measure the major flows in the region.

Prior to discussing the basic design, it should be noted that the following constraints had to be incorporated into the design and estimation procedures. The field staff (brigade) consisting of 4-5 men at each control station could work a maximum of 12 hours per day, 20 days per month with a minimum rest period of 6 hours between interview periods.

In order that traffic flows not be unduly disrupted during interview periods, a systematic sampling scheme could not be adopted (particularly in heavy traffic flows). The field procedure which is viable to implement can be briefly described as stopping the first truck, obtaining a complete interview, then stopping the next truck which passes the station when the interview is completed. During a selected interview period, a second member of the brigade (classifier) independently records the movements of all trucks passing the station according to truck type.

4. SAMPLE DESIGN (TEMPORAL)

4.1 Introduction

At each control station an independent sample is constructed. The design can be described as a stratified 3-stage replicated probability sample of trucks which pass a control station in a calendar year (trimester in the pilot).

4.2 Stratification

Stratification in a time frame is a process of classifying time units into certain collections called strata. An advantage of stratified sampling is the possible increase in efficiency per unit cost in estimating the population characteristics. In the context of a future national survey, stratification provides the flexibility of redesigning the sample of a specified stratum or groups of strata, without affecting the design in the remaining strata.

It is important that the variables selected for stratification meet certain requirements if the stratification is to result in the increase of design efficiency. The requirements for the P-O-D survey were:

(i) The stratification variables selected should be highly correlated with road cargo movements. Accounting for major sources of variation in cargo flows should result in the creation of relatively homogeneous strata which increases design efficiency.

(ii) For the purpose of the national survey, the variables selected should result in strata which maintain their homogeneity for the life of the survey.

(iii) The design is to be constructed so as to allow for the production of data at the annual level (national survey) as well as for each of the four trimesters. This constraint must also be accommodated in the stratification.

Sources of temporal variation in cargo flows are dependent specifically on the location of the interview stations (control points). For the purposes of the pilot, the following sources of temporal variation were determined on the basis of existing data, a field trip to several control stations, and subject matter knowledge of O.S.E. officials.

(a) time of day - morning, afternoon, evening, night

(b) day type - workday, non-workday

workday - (1) Mondays, Fridays

(2) Tuesdays, Wednesdays, Thursdays

non-workday - (3) Saturdays, Sundays, Holidays

(c) months

(d) trimesters.

The stratification variables selected were trimester (4) x months (3) x day type (3). Time of day is incorporated in the design through the sampling of time periods and ratio estimation.

In total, there are 9 strata for each trimester and hence 36 for the calendar year.

4.3 Sample Size

In view of financial constraints and the lack of reliable historical data, it was decided to select a maximum of 20 days per month for the 3 months of the trimester. The data resulting from the pilot will serve as input towards the determination of sample size requirements for the national survey. The sample size for the national survey will be based upon (1) data reliability requirements, (2) the sample design and estimation procedures used, (3) the systems implemented for geographical and commodity classifications and (4) financial constraints.

4.4 Sample Allocation

The following sample allocation strategies were discussed in detail:

(i) Neyman allocation,
(ii) X-proportional allocation (proportional to traffic volume), and
(iii) proportional to size (no. of time units in strata).

The Neyman allocation approach was eliminated due to the fact that (a) no historical reliable information base exists (i.e. no variance statistics), (b) the allocation could vary considerably with the commodity being measured and (c) the flow patterns could change over time which would adversely affect the design efficiency.

X-proportional allocation could be considered when reliable information on traffic flows becomes available. Traffic flow information was not considered sufficiently reliable to use this approach in the pilot. Further, although this approach may improve the reliability of many commodity estimates, the measurement of those commodity movements which are negatively correlated with total traffic flow would be adversely affected. In addition, it would be preferable to have truck traffic rather than total traffic (trucks and other vehicles) flows if such information were available. Again, if the flow pattern were unstable over the life of the survey, the design would require updating or possible redesign to accommodate such changes.

The strategy chosen for the pilot survey was proportional to size (i.e. proportional to the number of days in each stratum). This approach is conservative and relies least on reliable historical information. When the pilot is completed, a decision will be made as to whether or not to choose a compromise between this allocation and X-proportional for the national survey in view of the considerations outlined above.

4.5 Sampling Stages and Sample Selection Methods

Having determined the sample sizes (days) and allocation strategy, the first stage of sample selection is carried out. In each stratum an independent sample of days (20/month) is selected systematically. To facilitate variance estimation, in view of the complex survey design, the sample of days is systematically selected (without replacement) within each of two replicates for each stratum.

The second stage of selection involves the random selection of one six-hour period (morning, afternoon, evening or night) within each of the selected days. The selection of six hour periods in the pilot was such as to satisfy the field constraint requiring a minimum 6 hour rest period between interview periods.

Having selected the days and the time period within each of the days, the final stage of selection involves the selection of trucks within the selected time periods.

4.6 The Problem of Direction

To handle the problem of direction, it was decided to conduct interviews in <u>each</u> direction for 3 hours within the six hour time period. (In the development of a national survey, light flows may be accommodated with interviewing in both directions simultaneously without unduly disrupting traffic flows or statistical efficiency).

4.7 Truck Classification

During the six hour interview period, a second member of the field brigade records the truck traffic by type. This is required for the purpose of estimation (expansion of the sample to the population). To accommodate the interview methodology, this classification information should be recorded within sufficiently small time periods for statistical accuracy. For the pilot, classification is to be carried out for each 1-hour period instead of the entire 6-hour periods. With sufficiently small time periods, the sampling method approaches systematic sampling.

The time of interview is recorded by the interviewer on each questionnaire, for the purpose of estimation. It can also be noted that should the estimates for a major flow not be sufficiently reliable, two options are available to adjust the design, (1) adjust sample allocation, (2) increase the sample size (i.e. a second brigade at the same control station).

5. THE FORMS

5.1 Forms Design

The two major forms are: (1) Sample Control Document and (2) interview schedule.

(1) The Sample Control Document (SCD) contains all the information required to weight the sample to the population. A separate SCD is completed, corresponding to each 6-hour interview period. All the necessary control information is inserted before field operations begin. Such control information refers to location, date, time period, stratum, replicate, direction and sub-weight (see weighting). The SCD is then transferred to the brigade member responsible for truck classification. For each hour, he inscribes the total number of trucks by type which pass by the control station.

(2) The interview form is to be completed by the brigade member responsible for interviewing "Cammioneros". The control information outlined above is transferred onto the batched questionnaires before the field operations begin. The questionnaire presently allows for multi-commodity shipments. The total number of commodities is recorded and up to three are enumerated with "major" origin and "major" destination. The time of interview is recorded on the questionnaire for estimation purposes.

The following points should also be noted:

(i) Truck types and a commodity classification scheme have yet to be finalized. It was indicated that the greater the detail of the commodity classification, the lower will be, in general, the reliability of commodity estimates. As well, there will be an increase in the complexity of field operations, coding, quality control, and data processing procedures.

Since the commodities carried by different types of trucks vary by type and weight, defining major truck types for the purpose of estimation will improve the statistical efficiency of the design.

(ii) It is recognized that there is some information loss, particularly with agricultural products, for which it is probable that some trucks pick up the same commodity at different points along the route and drop off a commodity at several destinations along the route. To lessen the complexity of the field data collection, coding, quality controls and data processing, only one origin and one destination will be required.

(iii) A standardized system for recording weight as well as a standardized geographical code structure must be developed and clearly workable at the field level and by the data processing staff.

5.2 Batching, Colour Coding

It is suggested that the forms be colour coded corresponding to the interview stations to increase control of the survey documents.

The questionnaires are to be bound and enclosed with the SCD corresponding to each six hour period selected in the sample. Upon their completion, the forms are to be bound in a similar fashion and then forwarded.

5.3 Clerical Operations

The forms are to be checked for their completeness and accuracy upon their return from the field. Instructions need to be prepared for the staff to check the data and resolve discrepancies (see section 10).

After the data has been checked and corrected, the effective sample size (i.e. usable questionnaires) is determined on the basis of the time of interview which has been recorded on the questionnaires. This number is entered onto the SCD beside the corresponding truck type total which has been recorded by the truck classification officer in the field.

The clerical staff then checks the SCD for completeness, performs the necessary collapsing (when the effective sample take is zero) and, finally, adjusts the sub-weight according to eight pre-determined categories within each trimester defined by time period (4) and direction (2).

Having adjusted the sub-weight, the final weight (see weighting) is determined by the clerical staff and inserted onto the completed, checked questionnaires. The questionnaires are then ready for data entry.

6. TRAINING OF FIELD STAFF

It had been recognized that the success of the statistical system fundamentally rests with the reliability of the data collection. The importance of centralized or regional training programs - training sessions, interviewers' manuals, procedures manuals were discussed in the context of the data collection exercise. Provision has been made in the launching of the training program for the implementation of all these aspects.

7. ESTIMATION

7.1 Introduction

In a probability sample, the sample design itself determines the weights which may be used to produce unbiased estimates. Each record may be weighted by the inverse of the probability of selecting the unit to which the record refers.

The file to be created for tabulation purposes contains one record per commodity for each selected truck in the sample. Instead of physically duplicating the sample records, an overall weighting factor is entered in each record. For example, if the total weight of copper shipped by truck between a given origin and destination is required, this is done by sorting the records referring to those trucks in the sample which carry copper between the origin and destination and summing the product of the sample weights by the copper weights entered on these records.

Since objective information concerning the time periods and directions is available for the universe, the reliability of estimates can be improved by utilizing such auxiliary information. Ratio estimation is one of the most prevalent techniques of utilizing relevant information.

Upon comparing the estimates derived from the survey with those obtained externally, the estimates from outside sources are divided by the sample estimates for each classification and the weights of the records in each classification are adjusted by multiplying the weights by this factor. After the adjustment of the weights, the estimated aggregates will now agree with the estimate from the independent source for each classification. Ratio estimation is quite simple as compared to other methods of using external information and at the same time results in increased efficiency. The choice of external information is, however, very crucial to the procedure as it leads to higher efficiency only if such information is highly correlated with the characteristics of interest in the survey.

7.2 Weighting

In the cargo freight study, the final weight attached to each record is derived in five steps and is the product of five factors. These are referred to as the day weight, time period weight, direction weight, correction factor for non-response, and truck weight.

(i) Day Weight

This weight corresponds to the inverse sampling ratio of days selected for each replicate within each stratum. Since each replicate corresponds to a 1/2 sample, the weight is divided by 2.

(ii) Time Period Weight

Since one of four time periods corresponding to either 0:01-6:00; 6:01-12:00; 12:01-18:00; 18:01-24:00 is selected for each sampled day, this weight is four in all cases.

(iii) Direction Weight

Since classification and interviewing in the pilot is carried out in one direction only for each of three consecutive 1-hour periods, this weight is two in all cases. For classification and interviewing in 2 directions, the weight is 1.

The product of the above three factors will be referred to as the sub-weight. This sub-weight is entered into the SCD before the field operations commence.

(iv) Correction Factor for Non-Response

Since some SCDs will be lost (unusable, etc.) or since in some cases interviewing trucks was not possible, the information corresponding to the sample period will be lost unless the weights for other sample periods are adjusted to compensate for this. The sub-weights for completed SCDs are adjusted within eight classes defined by time period (4) and direction (2) based on the assumption that the volume and commodity characteristics of trucks that have been successfully interviewed represent the volume and commodity characteristics of trucks that should but were not interviewed within the above classes. However, if this assumption is not true, the estimates will be biased and the bias will be large with a high rate of non-response. The exact magnitude of bias introduced by the adjustment for non-response is impossible to calculate. Consequently, such an adjustment should be viewed as a last resort and every effort should be made to reduce it in the field.

The sub-weights multiplied by the correction factor provides the corrected sub-weight.

(v) Truck Weight

After the questionnaires have been checked and corrected, the effective sample take is recorded on the SCD according to truck type beside the total count of trucks. The ratio provides the sampling fraction for trucks and the inverse provides the truck weight.

The final weight corresponding to an interviewed truck corresponds to the product of the above five factors. This weight is calculated manually and is attached to each commodity of a sampled truck.

The estimate of total weight of a given commodity moved between a given origin and destination in a time period is obtained by first sorting all records referring to this commodity in the time period (trimester or year) having the corresponding origin and destination codes, and then aggregating the product of the final weight and the commodity weight.

A ratio adjustment was considered using as an auxiliary source, information supplied by traffic counters. The idea was dropped for three reasons, namely:

(1) traffic counters record number of axles not vehicles,
(2) procedures for incorporating the information were complex operationally and hence quite expensive, and
(3) a validation of the physical reliability of the counters would be necessary before any serious attempt to use the information is made.

7.3 Description of the Estimation

The estimation procedure outlined above can be stated algebraically and the following notation can be used for that purpose.

h - stratum (h = 1,2, ... 36)
i - replicate (i = 1,2)
j - day (j = 1,2,... 20)
k - time period (k = 1,2,3,4)
ℓ - direction (ℓ = 1,2)
m - hour (m = 1,2, ... 6)
r - type of truck (r = 1,2,3,4)
o - interviewed truck
t - trimester (t = 1,2,3,4)

The final weight for an interviewed truck as represented by $W_{hijk\ell mro}$ can be expressed as the produce of five weighting factors as follows:

$$W_{hijk\ell mro} = W_{hij}\ W_{hijk}\ W_{hijk\ell}\ W_{k\ell t}\ W_{hijk\ell mr}$$

Each of the component weighting factors may be expressed as follows:

(i) $W_{hij} = \dfrac{N_h}{2n_{hi}}$

N_h : Number of days in stratum h

n_{hi}: Number of days selected from stratum h in replicate i

(ii) $W_{hijk} \equiv 4$ (one of four time periods is selected)

(iii) $W_{hijk} \equiv 2$ (selection of one direction as in the pilot; heavy flows in the national survey)

$\equiv 1$ (two directions for light flows)

(iv) $W_{k\ell t} = \dfrac{N_{k\ell t}}{\hat{N}_{k\ell t}}$

$N_{k\ell t}$: Number of time periods of type k corresponding to direction ℓ in the t^{th} trimester in the universe

$$\hat{N}_{k\ell t} = \sum_{h \epsilon t} \sum_i \sum_j W_{hij}\ W_{hijk}\ W_{hijk\ell}$$

i.e. $\hat{N}_{k\ell t}$ is the sum of the sub-weights of usable SCDs referring to the k^{th} time period type corresponding to direction ℓ in the t^{th} trimester.

(v) $W_{hijk\ell mr} = \dfrac{N_{hijk\ell mr}}{n_{hijk\ell mr}}$

$N_{hijk\ell mr}$ = Total number of trucks of type r passing the control station during the m^{th} hour corresponding to direction ℓ in the k^{th} time period type of the j^{th} selected day belonging to replicate i in the h^{th} stratum as recorded by the truck classification officer.

$n_{hijk\ell mr}$ = Corresponding number of interviewed trucks.

The final weight is replicated for each commodity of a sampled truck.

7.4 Pooling of Hours and Truck Types in Case of Zero Observations

Let $T_{k\ell mr}$ refer to the total truck traffic of truck type r in the m^{th} hour in the ℓ^{th} direction within time period type k and $t_{k\ell mr}$ to the corresponding number of interviews.

If it happens that, for some r and m, $T_{k\ell mr} = 0$, there is no problem; the truck weight is defined to be 0 and simply does not appear in the estimate.

If, however, $T_{k\ell mr} \neq 0$ and $t_{k\ell mr} = 0$, the above procedure does not apply since a portion of the traffic would be omitted from all estimates. It is,therefore, necessary to develop a strategy for pooling over hours and/or truck types to account for zero observations.

The procedure to be followed in the P-O-D survey involves the collapsing of a given truck type over adjacent hours. If, for example, $t_{k\ell 11} = 0$ and $t_{k\ell 21} \neq 0$, the first two hours are collapsed for truck type one to form the following quantity necessary for the calculation of truck weights: $\frac{T_{k\ell m^*1}}{t_{k\ell m^*1}}$ where m^* refers to a pooled two hour time segment.

If pooling over all the hours in one direction still results in zero observations, pooling is then carried out between adjacent truck types. Finally, should zero observations remain after such pooling, the number of interviews within a direction in a given time period is zero and the corresponding period is treated as non-response. This component of non-response is handled by an adjustment to the sub-weights discussed in the estimation section.

8. SPECIAL CONSIDERATIONS

8.1 Empty Trucks

The code value corresponding to 'empty' is inserted onto the questionnaire. The commodity carried is 'no commodity' and no distinction is made for the purpose of weighting. Empty trucks are then handled through domain estimation.

8.2 Duplication

Since more than one control point may be stationed along a given route between a defined origin and destination, the problem of duplication arises. The problem of duplication stems from the fact that a truck has a non-zero probability of being stopped at each station along the route. The problem cannot be overcome by sampling each station at different times or rejecting duplicate trucks in the sample. The following four methods are proposed to resolve this problem.

1) Select the control station in advance which will provide the P-O-D estimate. This option is easy to apply. The principal drawback rests with the fact that a station is chosen arbitrarily and that information from other control stations is lost.

2) Select the control station for which the P-O-D estimate is based on the largest sample size. The drawback here again is that information from other control stations is lost and that record counts must be determined for each estimate to select the station.

3) Create a simple average of the estimates from the control stations along an O-D route. The advantage here is that all information is used. The drawbacks are that each station's contribution to the estimate is the same (i.e. equal weight) regardless of the sample size and that programming is required to construct such estimates.

4) Create a weighted average of the estimates from the control stations along an O-D route on the basis of relative sample size. This method is the one which is the most statistically sound. The advantage again is that all information is used in a manner which minimizes the variance of P-O-D estimates. The drawback stems from the fact that record counts must be determined for all estimates and extra programming to construct such estimates is necessary.

8.3 Alternative Routes from an Origin to a Destination

There are some (not many) alternative routes from a major origin to a major destination in Peru. One such road network has been included in the pilot test. Estimates of P-O-D obtained from alternative routes must be aggregated. After aggregation, the situation reduces to that corresponding to one route.

9. VARIANCE ESTIMATION

On the basis of sample data, estimates of the sampling variability of P-O-D estimates can be calculated.

The methodology adopted for estimating variances in the pilot test is pseudo-replication. For the purpose of variance estimation, the two replicates selected from within each stratum are assumed to have been selected independently. In fact, the units for each of the two replicates were selected without replacement. In addition, the adjustment for non-response is carried out for both replicates together rather than independently for each replicate. This adjustment is carried out at the

trimestrial level rather than the stratum level. The replicates are, therefore, somewhat correlated but this has been assumed negligible in the pilot. A more precise formulation for the construction of variance estimates requires more complex programming. Such a system may be developed for the national survey as explained in a subsequent section of this report. The advantage of the present formulation is that it should be relatively easy to program and should provide useful approximations with respect to data reliability.

Letting $\hat{X}_t$ represent a given P-O-D estimate for trimester t, the variance estimate denoted by $\hat{V}(\hat{X}_t)$ may be determined as follows:

$$\hat{V}(\hat{X}_t) = \sum_{\substack{h=1 \\ h \epsilon t}}^{9} (\hat{X}_{h1} - \hat{X}_{h2})^2$$

where X_{h1} = half sample estimate corresponding to replicate 1 in stratum h

X_{h2} = half sample estimate corresponding to replicate 2 in stratum h

and $\sum_{\substack{h=1 \\ h \epsilon t}}^{9}$ refers to summation over all strata in trimester t.

The variance estimate of the corresponding annual P-O-D estimate $\hat{X}_A$ denoted by $\hat{V}(\hat{X}_A)$ may be determined as follows:

$$\hat{V}(\hat{X}_A) = \sum_{t=1}^{4} \hat{V}(\hat{X}_t)$$

where $\sum_{t=1}^{4}$ refers to the summation over the 4 trimesters.

10. EDITING AND IMPUTATION

10.1 Introduction

The purpose of editing and imputation is to identify and correct invalid entries or codes, to reconcile conflicting data and as far as possible to fill in missing values in information fields of records with partial

non-response. This procedure is distinct from weighting adjustments for complete non-response which do not take the form of changes within 'individual' records but may change an entire record's relative weight in the estimates.

Editing and imputation must be done according to a specified set of steps and decision rules which are based mainly on external knowledge and on logical rules.

All the editing and imputation to be carried out on the pilot responses are to be done at the level of the individual record, bypassing the necessity of a "hot-deck" frequency imputation approach.

The following sections will illustrate the general approach to edit and imputation since exact specifications will depend upon the final questionnaire content.

10.2 Coding and Transcription Errors

All editing will be carried out manually. Invalid codes may appear in single fields of identification. Correction must be made using other available identification information. Alternatively, several fields may individually contain allowable codes, but taken in combination, indicate a non-existent case. For example, it is not enough simply to ensure that 'Location' is one of the allowable codes; it must correspond to a selected day and time period combination. This sort of conflict is easy to specify by means of a list of valid code combinations.

10.3 Edit/Imputation Specifications

All edits must be specified before manual editing can be carried out. The first step involves the identification of all allowable responses. This is handled by means of a coded list of control stations, origins, destinations, commodities, weight classification, etc. For actual commodity weight, any non-negative integer, for example, may be acceptable

but usually a limit must be placed. Failures may essentially take one of three forms: a single field containing an invalid code which must be corrected; a single field containing an invalid blank; or two or more fields in conflict, that is, each having a legal code but forming an invalid combination. For variables which appear on the SCD and questionnaire, one would normally demand agreement on information appearing on these records. In case of disagreement, the problem is to decide which record contains the correct codes.

Further, there are some logical edits. Consider, for example, the situation where the total number of commodities recorded is less than the number of commodities listed.

Finally, one may impose "reasonableness" edits. Could a two-axle truck carry 10 commodities, 3 of which are listed and whose total weight is 1,000 tons? Subject matter knowledge will determine how many of this type of edit will be worth imposing. Since the editing is to be done manually, the complexity of the edit structure should be kept to a minimum.

Having identified those records which fail one or more edit rules, the problem is how to correct them. The first task is to separate questionnaires into three groups: (a) records which pass all edit rules, (b) records which require some correction and (c) unusable returns. The first two groups will provide the data for further processing. Rules need to be established concerning the manner in which edit failures are corrected for group (b).

Single field correction is to be resolved by referring to the SCD for the same information or deterministically on the basis of a defined set of imputation actions. Most difficult to resolve is the situation involving multiple field edits. In some cases, one may just change the minimum number of fields required to obtain a valid combination. In other cases, one field or a relationship between fields may be determined

to be of over-riding importance. In any event, all imputation rules must be developed which can be easily and consistently applied manually.

11. COMPUTER PROCESSING AND TABULATION

From the resulting clean and weighted set of questionnaires, a clean file can be created subject to errors involved in the data entry which can be minimized by means of key-edit, etc. Construction of any desired estimate is then relatively straightforward: the estimate for any O-D is found by just summing the product of the final weights and the commodity weights for the period under consideration. Tabulation is a matter of deciding what estimates to produce and arranging a suitable output format.

12. CONSIDERATIONS FOR THE SURVEY DESIGN OF A NATIONAL CONTINUOUS P-O-D SURVEY

Outlined below are recommendations concerning an analysis of the pilot survey data and an evaluation of the survey operations to serve as input towards the development of a national or regional P-O-D survey.

12.1 Concepts and Definitions

The adequacy of the concepts and definitions used in the pilot to obtain origin, destination, single and multi-commodity shipments should be assessed in terms of statistical and operational effectiveness.

In particular, the manner in which the problem associated with several origins/destinations corresponding to the same commodity is handled in the pilot, should be evaluated. Further, decisions must be made on the basis of the pilot as to the maximum number of commodities that can be reliably obtained during the course of a roadside interview.

12.2 Classification Systems

The commodity, truck, weights and measures, and geographical classification systems adopted in the pilot, should be evaluated in terms of their ability to satisfy data requirements, and their operational and statistical efficiency. Revised systems can then be implemented at the regional and national levels.

12.3 Field Operations

The methods and procedures used with respect to hiring and training field brigades in the pilot should be evaluated and possibly revised. In particular, the field procedures should be reviewed with respect to their ability to handle heavy traffic flows and surveying under adverse climatic conditions. It should be possible to determine how many interviews can be successfully handled by a brigade as a function of traffic volume. Such information can be used to monitor the performance of brigades in the national survey. Systematic programs for field observation should, as well, be an integral part of a national data collection exercise.

12.4 Sample Design

(a) Stratification

On the basis of an analysis of major commodity estimates, the efficiency of the stratification variables selected for the pilot can be assessed. Those variables which are found to be important can be incorporated into the revised sample design.

(b) Sample Size/Sample Allocation

Analysis of the variance estimates for selected commodity estimates should guide decisions with respect to adjustments of the sample size and/or allocation. For example, heavy flows might be handled by two brigades instead of one.

The allocation strategy which should be considered in the context of a national survey, is a compromise allocation falling between proportional to size of strata and proportional to estimated traffic volume of strata.

(c) Selection of Time Periods

If large differences in traffic flow are found between the four periods of the day, the periods within selected days may be selected with probability proportional to estimated traffic flows. Such a selection procedure should be considered only if traffic flow information is judged to be sufficiently accurate and reasonably stable over time.

The truck weight (fourth factor in the overall weight) is very large if a truck is selected during a heavy traffic flow time period. The selection of time periods by PPS would result in smaller time period weights (second factor in the overall weight) corresponding to such periods of heavy traffic. The overall weights for trucks would then be more even resulting in more stable estimates.

The selection of time periods was carried out in the pilot to satisfy field operating constraints. It is suggested that the selection of time periods be carried out on a random basis for each selected day and subsequently adjusted to satisfy field constraints.

12.5 Estimation

It is recommended that the adjustment of the sub-weights for non-response be carried out at the stratum level within each replicate rather than at the trimestrial level over both replicates. This procedure, of course, adds to the complexity of the operation and might be considered when experience has been gained and the development of automated procedures becomes feasible and practical.

The assumption underlying the adjustment for non-response is that information obtained for certain time periods represent other time periods for which information is not available. The assumption is more valid if the correction is carried out at the stratum level.

For the purpose of variance estimation outlined below, the replicates constructed for variance estimation purposes will be correlated unless the non-response adjustment is carried out independently for each replicate. The variance estimates would otherwise be subject to bias, the magnitude of which could be high if each of the replicates have very different response rates.

12.6 Variance Estimation

To account for the ratio weight adjustment by time period type and direction to improve the statistical accuracy of P-O-D estimates, more precise variance estimates than those used in the pilot can be constructed on the basis of the following formulation.

Letting $\hat{X}_t$ and $\hat{X}_A$ refer to the trimestrial and annual P-O-D estimates of commodity weight X and $\hat{V}(\hat{X}_t)$ and $\hat{V}(\hat{X}_A)$ to their variance estimates, the following is calculated for each stratum.

$$\Delta_h = \hat{X}_{h1} - \hat{X}_{h2} - \sum_{k=1}^{4} \sum_{l=1}^{2} \frac{\hat{X}_{tk\ell}}{T_{tk\ell}} (\hat{T}_{h1k\ell} - \hat{T}_{h2k\ell})$$

where $\hat{X}_{hi}$: refers to the half sample estimate of the total weight to commodity X derived from the i^{th} replicate in stratum h (i=1,2)

$\hat{X}_{tk\ell}$: refers to the total weight of commodity X in trimester t corresponding to the k^{th} time period type in the ℓ^{th} direction

$T_{tk\ell}$: refers to the total number of time periods of type k in the ℓ^{th} direction in trimester t

$\hat{T}_{hik\ell}$: refers to the half sample estimate of the number of periods of type k in the ℓ^{th} direction obtained from the i^{th} replicate in stratum h (i = 1,2).

The variance estimate corresponding to $\hat{X}_t$ is calculated as follows:

$$\hat{V}(\hat{X}_t) = \sum_{h=1}^{9} \Delta_h^2$$

The corresponding variance estimate for the annual estimate $\hat{X}_A$ is calculated as follows:

$$\hat{V}(\hat{X}_A) = \sum_{t=1}^{4} \hat{V}(\hat{X}_t).$$

12.7 Data Handling

The problems encountered in the manual data processing cycle: coding, clean up of forms, weighting, etc. should be evaluated in terms of cost, timeliness and accuracy. Certain procedures (e.g. part of the edit/imputation or weighting) may be more effectively and efficiently handled through the development and implementation of automated packages.

12.8 Time and Cost Analysis

The time and cost of all phases of the pilot survey should be recorded and later evaluated to estimate such parameters in the proposed national survey. This will serve as input into all phases of the development and implementation of the national survey.

ACKNOWLEDGEMENTS

The research presented in this paper was carried out under the general direction of Dr. K.W. Studnicki-Gizbert, Executive Director of the Research Branch of the CTC. This project is part of an overall technical assistance program in the transportation field which has been commissioned by the Canadian International Development Agency in Peru.

The authors wish to acknowledge the helpful suggestions of Messrs M. Nargundkar, H. Gough, and Miss J. Forgie of Statistics Canada and the support and direction of Miss M. Fleming (CTC) and Mr. R. Platek, Statistics Canada and Messrs A. Gemmell and D. Napier, the CTC representatives in Peru for the background research and their support to us while in Lima.

RESUME

Cet article décrit un plan d'enquête qui a été élaboré pour mesurer le flot de marchandises transportées par camion au Pérou. L'article considère les traits conceptuels et opérationnels du plan d'enquête, et en décrit les éléments et les techniques d'exécution dans le contexte d'un projet pilote. Enfin, on démontre comment on pourrait utiliser les résultats de ce projet pilote pour élaborer et exécuter une enquête nationale à grande échelle.

REFERENCES

[1] Gough, J.H., Ghangurde, P.D. (1976), "An Alternative Method of Surveying International Travellers At Frontier Points - Methodology Report", Statistics Canada.

[2] Platek, R. Singh, M.P. (1976), "Methodology of Canadian Labour Force Survey", Household Surveys Development Division, Statistics Canada.

[3] "For-Hire Trucking Survey (1976)", Transportation and Communications Division, Statistics Canada.

[4] Koop, J.C. (1960), "On Theoretical Questions Underlying the Technique of Replicated or Interpenetrating Samples", Institute of Statistics, North Carolina State College.

[5] Gemmell, A., Napier, D. (1977), "Background Document On Inter-Urban Cargo Statistics (Road)", Research Branch, Canadian Transportation Commission.