International Urban Freight Conference (I-NUF)
October 21 - 23, 2015 - Long Beach, California

Urban Freight Survey Strategy to Address Skewness in Freight Activity Variables

October 23, 2015

Takanori SAKAI, University of Illinois at Chicago
Kazuya KAWAMURA, University of Illinois at Chicago
Tetsuro HYODO, Tokyo University of Marine Science and Technology
Outline

1. Introduction
2. Past Establishment Surveys
3. Tokyo Metropolitan Freight Survey
4. Analysis
   4.1 Approach
   4.2 Variable Distributions and the Contribution of Mega Shippers
   4.3 Sample Size and Sampling Variability
   4.4 Evaluation of Possible Sampling Strategies
5. Conclusion
1. Introduction

Background & Purpose

Background

• Scarcity of data is one of the key challenges in urban freight transportation planning
  - Complexity and heterogeneity of freight traffic
  - High cost of survey
  - Corporate privacy

• Accuracy of traffic estimation is critical to decision making

Study Purpose

• To examine the distributions of key shipment variables and the sampling strategies for urban freight survey
2. Past Establishment Surveys

Sample Sizes in the Past Establishment Surveys

• Allen and Browne (2008) - review of urban freight surveys around the world
  - 162 freight surveys, going back as far as the 1960s.
  - 92 establishment surveys and 5 commodity surveys covered.
    ✓ only 9 surveys have more than 500 samples since 2000;
    ✓ among them, only 3 surveys exceeded 3,000 samples.

• RSG (2012) reviewed recent establishment surveys: Phoenix (Y2007; 562 samples), Toronto (Y2007; 597 samples), Tokyo (Y2003; 30,000 samples).

• Recently, Paris survey (Y2012; 1,300 samples), Bordeaux survey (Y2013; 1,570 samples) and Lisbon survey (Y2013; 604 samples) were conducted.

Urban freight surveys exceeding 500 samples are relatively rare.
Use of Stratified Sampling Approach

- Calgary (2000) and Edmonton (2001): Strata based on **industry category**.
- Paris (2012): Strata based on **activity type**, **size** and **land-use type**.
- Lisbon (2013): Strata based on **establishment category** and **land use type**.

**Stratification is applied using information on target sample’s characteristics available in advance.**
Tokyo Metropolitan Freight Survey (TMFS)

- Decennial urban freight survey covering Tokyo Metropolitan Area

**2013 TMFS**
- Area: 23 thousand km$^2$
  - (16 thousand km$^2$ in 2003 TMFS)
- Population: 42 million
  - (37 million in 2003 TMFS)

- Mail survey (mainly)/ interviews for large companies

- Covered following components:
  1. Size & function of establishments
  2. Records on both inbound and outbound shipments
  3. Attributes of goods movement (weight, item etc.)
  4. Attributes of mode (freight mode, truck size, truck ownership etc.)

- Agriculture/forestry/fishing, mining, and construction are not covered.
3. Tokyo Metropolitan Freight Survey

**Tokyo Metropolitan Freight Survey (TMFS) (2)**

- 2013 TMFS collected 43,588 responses (response rate: 32%)
  
  (2003 TMFS collected 29,485 responses with response rate of 25%)

- Data of **manufacturing, trans. service, warehousing, wholesale, and service industry** were used.

- Expansion factors for the sample data were determined based on location, type of industry, and employment size.

---

**Distribution of Respondents (2013TMFS)**

We use expanded sample (n = 110,548) for this study.
4.1 Approach

**Analysis Approach**

Analyses of establishment survey data in three steps (focusing on *outbound truck trips and tons shipped*)

1. Distributions of outbound truck trips and tons shipped

2. Relationship between sample size and sampling variability (using *bootstrap method*)

3. Effectiveness of different stratification approaches (using *bootstrap method*)
4.2 Variable Distributions and the Contribution of Mega Shippers

Distribution of Truck Trips and Tons Shipped

Truck trips: Top 5% of the establishments account for 59% of the total.

Tons: Top 5% of the establishments account for 86% of the total.
• Outliers (responses that exceed upper quartile plus 1.5 times the inter quartile range)
  ➢ 74.1% of truck trips
  ➢ 98.7% of tons shipped

• Far outliers (responses that exceed upper quartile plus 3 times the inter quartile range)
  ➢ 65.9% of truck trips
  ➢ 97.7% of tons shipped
What does this mean?

Surveys with small N likely to miss mega shippers

Miss large portion of freight traffic
4.2 Variable Distributions and the Contribution of Mega Shippers

Share of Top 10% within Each Stratum

Share of top 10% shippers within each stratum (manufacturing industry)

**Targeting larger establishments (in terms of employment size) may not capture high volume shippers.**
## A survey with modest sample size can produce highly inaccurate estimates.

### Simulation of Random Sampling Survey

Distribution of resampling samples’ means (10,000 draws)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>n</th>
<th>truck trips/day</th>
<th>±25% of true mean</th>
<th>tons/day</th>
<th>±25% of true mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>2.6-6.1</td>
<td>77.0%</td>
<td>7.6-38.8</td>
<td>44.1%</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>2.9-5.6</td>
<td>87.8%</td>
<td>9.3-35.1</td>
<td>54.9%</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>3.2-5.1</td>
<td>96.4%</td>
<td>11.1-31.9</td>
<td>71.0%</td>
</tr>
<tr>
<td>Manufacturing Est.in top 10% employment size (4,433 samples)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>6.9-26.7</td>
<td>51.6%</td>
<td>22.2-188</td>
<td>31.9%</td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>8.5-22.4</td>
<td>69.3%</td>
<td>33.4-152</td>
<td>46.2%</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>10.6-18.9</td>
<td>91.2%</td>
<td>47.9-117</td>
<td>72.7%</td>
</tr>
</tbody>
</table>
4.4 Evaluation of Possible Sampling Strategies

Stratified Sampling with Neyman Allocation

- Tested five stratification strategies with Neyman (Optimum) Allocation (n=500, assuming the S.D. of each stratum is known).

Stratified sampling with Neyman Allocation based on **truck trips**

<table>
<thead>
<tr>
<th>Stratification strategy</th>
<th>Sampling distribution of means</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>truck trips/day</td>
<td>tons/day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>95% CI</td>
<td>S.E.</td>
</tr>
<tr>
<td>Without stratification</td>
<td>0.68</td>
<td>2.88-5.54</td>
<td>7.14</td>
</tr>
<tr>
<td>(i) Industry (5 categories)</td>
<td>0.63</td>
<td>2.96-5.40</td>
<td>6.00</td>
</tr>
<tr>
<td>(ii) Industry (13 categories)</td>
<td>0.59</td>
<td>3.02-5.30</td>
<td>5.69</td>
</tr>
<tr>
<td>(iii) Employment size (Pattern 1)</td>
<td>0.52</td>
<td>3.12-5.14</td>
<td>5.30</td>
</tr>
<tr>
<td>(vi) Employment size (Pattern 2)</td>
<td>0.52</td>
<td>3.11-5.15</td>
<td>5.25</td>
</tr>
<tr>
<td>(v) Industry (5 categories) &amp; Employment size (Pattern 2)</td>
<td>0.48</td>
<td>3.17-5.04</td>
<td>4.67</td>
</tr>
</tbody>
</table>

**Employment stratification**

**Pattern 1**: 0-20<sup>th</sup>, 20-40<sup>th</sup>, 40-60<sup>th</sup>, 60-80<sup>th</sup> and 80-100<sup>th</sup> percentiles

**Pattern 2**: 0-50<sup>th</sup>, 50-75<sup>th</sup>, 75-90<sup>th</sup>, 90-95<sup>th</sup> and 95-100<sup>th</sup> percentiles
### Stratified Sampling with Neyman Allocation (2)

Stratified sampling with Neyman (Optimum) Allocation based on tons shipped

<table>
<thead>
<tr>
<th>Stratification strategy</th>
<th>Sampling distribution of means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tons/day</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
</tr>
<tr>
<td>Without stratification</td>
<td>7.14</td>
</tr>
<tr>
<td>(i) Industry (5 categories)</td>
<td>5.67</td>
</tr>
<tr>
<td>(ii) Industry (13 categories)</td>
<td>5.12</td>
</tr>
<tr>
<td>(iii) Employment size (Pattern 1)</td>
<td>5.19</td>
</tr>
<tr>
<td>(iv) Employment size (Pattern 2)</td>
<td>5.00</td>
</tr>
<tr>
<td>(v) Industry (5 categories) &amp; Employment size (Pattern2)</td>
<td>4.09</td>
</tr>
</tbody>
</table>

**Stratification would result in modest improvement over the simple random sampling for both truck trips and ton shipped.**
4.4 Evaluation of Possible Sampling Strategies

**Targeting Strata of the Large Share of Shipment**

13,000 strata based on **industry** (13), **facility type** (10), **floor area** (10) and **employment size** (10)

Ordered by ave. trip rate

Conduct survey

- **Target Strata**
- Estimate ave. of the population

2003 TMFS, assumed to be available, is used for finding the strata for conducting a new survey. Simulate the survey process with 2013 TMFS data.
4.4 Evaluation of Possible Sampling Strategies

**Estimation of Mean based on the Sampled Data**

\[
\hat{M}_{13} = \frac{T_{ts,13} \times \frac{100}{ST_{ts,03}} \times \frac{n_{ts,03}}{n_{ts,13}}}{N_{13}}
\]

- \( \hat{M}_{13} \): The estimated mean of population
- \( T_{ts,13} \): **Estimated total no. of truck trips/tons of the target strata based on the survey** (2013 TMFS)
- \( ST_{ts,03} \): The share of the target strata in total no. of truck trips/tons based on 2003 TMFS
- \( n_{ts,03}, n_{ts,13} \): No. of the establishments in the target strata, in 2003 or 2013.
- \( N_{03}, N_{13} \): No. of the establishments in population, 2003 or 2013
4.4 Evaluation of Possible Sampling Strategies

Selection of the Strata based on truck trips per day (n = 500)

If top 1000 strata is targeted,

- 26.9% of all establishments are covered
- 95% confidence interval is 3.20 – 4.90, compared with 2.88 – 5.54 of random sampling.
4.4 Evaluation of Possible Sampling Strategies

Selection of the Strata based on tons per day (n = 500)

If top 1000 strata is targeted,

- **23.0%** of all establishments are covered
- **95% confidence interval** is **11.37 – 25.95**, compared with **9.28 – 36.68** of random sampling.
Conclusions

Analyses underscore the issue of the skewness of freight data

- The distribution of variables such as truck trips generated and tons shipped can be highly skewed even with stratification.
  - The range of the sampling errors across strata can be very large, and/or
  - The standard errors for some strata can be deceptively small
- A survey with a modest sample size can produce highly inaccurate estimates
- The effectiveness of Neyman allocation is not promising. (especially if the accurate data of the S.D. is not available)
Conclusions (2)

- Survey data from an earlier year could help improve the accuracy of the survey with small sample size. (This may not be the case for tons shipped.)

Sampling/data management strategies should be continuously pursued; especially, the ones based on **vehicle registration, business record, other surveys such as road-side intercept surveys.**
Fin...

thank you for listening
## Coverage of Top 100 Strata – 2003 vs. 2013

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>30.7%</td>
<td></td>
<td>80.0%</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>26.9%</td>
<td>0.88</td>
<td>68.8%</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>26.4%</td>
<td></td>
<td>96.5%</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>23.0%</td>
<td>0.87</td>
<td>76.6%</td>
<td>0.79</td>
</tr>
</tbody>
</table>

As for tons, the decrease on tons-covered is much more than establishments-covered; this causes the under-estimation.
Repeated sampling (with replacement) from a population to produce many “pseudo samples” that can be used to investigate the properties of estimators.

- **Population**
  - (Expanded TMFS samples)

- **Sampling**

- **Bootstrap sample**
  - (n = 50, 100, 300, 500 or 1,000)

- **Target property**
  - (Mean)

- **Distribution**
  - (Distribution of mean)

Red: applied in this study

Repeat: (10,000 times)
4.4 Evaluation of Possible Sampling Strategies

**Neyman Allocation**

Neyman allocation, sampling by allowing a higher sample size for a stratum with a higher standard deviation, minimizes the variance of sample mean.

\[
n_h = n \times \frac{N_h \times S_h}{H \sum_{i=1}^{H} (N_i \times S_i)}
\]

- \(n_h\): the optimal sample size of stratum \(h\)
- \(n\): the total sample size
- \(N_h\): the population size of stratum \(h\)
- \(S_h\): the standard deviation of stratum \(h\)
- \(H\): the number of strata.