DATA COLLECTION TECHNIQUES AND PROGRAM DESIGN

Outline

1. Summary of Current Practice
2. Framework for Data Collection
3. Data Needs
4. Manual Data Collection Techniques
5. Sampling
6. Special Considerations for Surveying

Additional readings for this block:


Summary of Transit Data Collection Programs

Great variation in data collection resources

Variation in techniques used: automated, manual, mixed

Statistical validity of sampling approach varies

Inefficient use of data

Often limits use of other analytic planning and operations methods

ADCS presents major opportunity for strengthening data to support decision-making
Summary of Data Collection Program
Design and Implementation

- Determine data needs
- Determine property characteristics
- Assemble available data

Select data collection techniques

Develop route-by-route sampling plans, checker requirements, and cost

Conduct baseline phase

Determine any desired changes in monitoring phase techniques, sampling plans, and checker requirements

Conduct periodic monitoring phase

If significant change is detected

Determine if a pretest is required

Conduct pretest, if necessary

Determine statistical inputs for estimating sample size
Data Needs in Baseline Phase

A. Route (or Stop) Specific
   Load (at peak point -- other key points)
   Running time
   Schedule adherence
   Total boardings (i.e., passenger-trips)
   Revenue
   Boardings (or revenue) by fare category
   Passenger boarding and alighting by stop
   Transfer rates between routes
   Passenger characteristics and attitudes
   Passenger travel patterns

B. System Wide
   Unlinked Passenger Trips
   Passenger-miles
   Linked Passenger trips
## Conversion Factors

<table>
<thead>
<tr>
<th>Auxiliary Data Item</th>
<th>Inferred Data Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load or Revenue</td>
<td>Boardings</td>
</tr>
<tr>
<td>Boardings, Load or Revenue</td>
<td>Passenger Miles</td>
</tr>
<tr>
<td>Point Load</td>
<td>True Maximum Load</td>
</tr>
<tr>
<td>Revenue</td>
<td>Peak Point Load</td>
</tr>
</tbody>
</table>
Passenger Counting Techniques

Operator (trip cards)

Traffic Checker (with handheld device)
- ride check (on/off and running time)
- point check (load and headway)

Fare System
- passenger counts
- revenue counts only

Automatic Passenger Counters

Passenger Surveys
# Types of Counts and Readings

<table>
<thead>
<tr>
<th>Type of Count/Reading*</th>
<th>Description</th>
<th>Corresponding Deployment Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>On/off count</td>
<td>Ons and offs by stop; also time at time points. In rare cases, ons may be by fare category</td>
<td>Ride check/APC</td>
</tr>
<tr>
<td>Boarding counts</td>
<td>Boardings by trip, by fare category, may be also by stop</td>
<td>Ride check Driver count Fare systems</td>
</tr>
<tr>
<td>Load counts</td>
<td>Load on bus as it passes a point; also time at that point</td>
<td>Point check APC</td>
</tr>
<tr>
<td>Revenue count</td>
<td>Revenue by fare type</td>
<td>Fare systems</td>
</tr>
<tr>
<td>Transfer counts</td>
<td>Count of transfer tickets sorted by original and final route</td>
<td>Fare systems</td>
</tr>
<tr>
<td>Route origin/Destination count</td>
<td>Count of passengers by O/D stop pair</td>
<td>Special</td>
</tr>
<tr>
<td>Survey</td>
<td>Passengers respond to questions, either written or verbal</td>
<td>Special</td>
</tr>
</tbody>
</table>

* See Transit Data Collection Design Manual, p. 35 in reader, to see how each of these can be used in more detail

Improving Traffic Checker Data

Point check observed 70 passengers on a trip.

- **Uncontrolled**: load could be anywhere between 41 and 81.
- **Large random variation compounded by systematic overcount**
- **Controlling error**: verification counts, immediate feedback, retraining
- **Have the checker board the bus to count**
Improving Traffic Checker Data (cont’d)

Preprinted forms:
• scheduled trips, stop lists

Handheld devices
• reduce real-time coding errors
• error detection
• load checks

Have checkers code their own data
• immediate graphical feedback

Watch for fabricated data
Designing a Data Collection Program

**BASELINE PHASE**

- Data needs, Baseline Phase
  - Check accuracy of systemwide totals
  - Sample size for individual items
    - Choose data collection techniques, schedule data collection
      - Analyze individual items
  - Choose conversion factors
    - Sample size for conversion factors
      - Analyze conversion factors

**MONITORING PHASE**

- Data needs, Monitoring Phase
  - Check accuracy of system-wide totals
  - Sample size for indirect measurement
    - Analyze data
      - Trigger need for follow-up
  - Sample size for direct measurement
    - Choose data collection techniques, schedule data collection
Sampling Strategies

Simple random sampling
Every trip has equal likelihood of selection

Systematic sampling
Sample every 6\textsuperscript{th} day – like random, but smoothes data collection load
   \textit{Example: FTA Circular}

Cluster sampling
Identify natural clusters in advance, select them at random
With passenger surveys, bus trip = cluster of passengers
   \textit{Example: on-board survey}
   \textit{Example: sample round trips, or clusters of 4 trips}
Sampling Strategies

Ratio estimation/Conversion factors
Take advantage of complete or less expensive data sources
  
  *Example:* convert farebox boardings to pass.-mi
  
  *Example:* convert load at checkpoint to load elsewhere

Stratified sampling
Separate sample for each stratum
  
  *Example:* long vs. short routes for average trip length
Accuracy of an estimate has two dimensions.

“Mean boardings per trip is 33.1.”

Exactly 33.1???

“Mean boardings per trip is 33.1, plus or minus 10%”
   – precision

“Mean boardings per trip is 33.1, plus or minus 3.3”
   – tolerance

Are you sure?

“I’m 95% confident that mean boardings per trip is 33.1, plus or minus 10%”
   – precision and confidence level

To simplify matters:
   • hold confidence level fixed (90%)
   • vary precision to reflect different levels of accuracy

*National Transit Database specification for annual boardings, pass-miles:*
+10% precision at 95% confidence level
Desired Accuracy (AET)

System boardings for management purposes:

±3% quarterly – equivalent to ± 1.5% for annual estimate

On-time performance, systemwide:

Suppose percent on time is 80%. Choose tolerance:

80% ± ? (I’ll choose 4%)

Convert to “absolute equivalent tolerance”

\[ AET = 0.5 \frac{tol}{\sqrt{p(1-p)}}, \text{ where } p = \text{expected proportion} \]

\[ AET = 0.5 \left(4\%\right) / \sqrt{0.8\times0.2} = 5\% \]
Tolerance naturally improves as the proportion moves toward the extremes (0%, 100%). AET is the tolerance you’d get if the proportion were 50%.

<table>
<thead>
<tr>
<th>Expected proportion</th>
<th>Tol. corresponding to +5% AET</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>5%</td>
</tr>
<tr>
<td>60% or 40%</td>
<td>4.9%</td>
</tr>
<tr>
<td>70% or 30%</td>
<td>4.6%</td>
</tr>
<tr>
<td>80% or 20%</td>
<td>4%</td>
</tr>
<tr>
<td>90% or 10%</td>
<td>3%</td>
</tr>
<tr>
<td>95% or 5%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>
# Recommended Tolerances

## Peak Load (also boardings):

<table>
<thead>
<tr>
<th>Routes with</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3 buses</td>
<td>±30%</td>
</tr>
<tr>
<td>4-7 buses</td>
<td>±20%</td>
</tr>
<tr>
<td>8-15 buses</td>
<td>±10%</td>
</tr>
<tr>
<td>&gt;15 buses</td>
<td>± 5%</td>
</tr>
</tbody>
</table>

## Vehicle trip time:

<table>
<thead>
<tr>
<th>Routes with trip time</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 20 mins</td>
<td>±10%</td>
</tr>
<tr>
<td>≥20 mins</td>
<td>± 5%</td>
</tr>
</tbody>
</table>

## On-time performance

±10% AET
## Default Values for Coefficient of Variation of Key Data Items

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Time Period</th>
<th>Route Classification</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Load</td>
<td>Peak</td>
<td>&lt; 35 pass./trip</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≥ 35 pass./trip</td>
<td>0.35</td>
</tr>
<tr>
<td>Off- Peak</td>
<td>&lt; 35 pass./trip</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>35-55 pass./trip</td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>&gt; 55 pass./trip</td>
<td></td>
<td>0.35</td>
</tr>
<tr>
<td>Evening</td>
<td>All</td>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>Owl*</td>
<td>All</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Sat, 7 AM-6 PM</td>
<td>All</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>Sat, 6 PM-1 AM</td>
<td>All</td>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>Sun, 7 AM-1 AM</td>
<td>All</td>
<td></td>
<td>0.75</td>
</tr>
</tbody>
</table>

*Owl default values are the same for weekdays and weekends
# Default Values for Coefficient of Variation of Key Data Items

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Time Period</th>
<th>Route Classification</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boardings, Passenger Miles</td>
<td>Peak</td>
<td>&lt; 35 pass./trip</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≥ 35 pass./trip</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Off- Peak</td>
<td>&lt; 35 pass./trip</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35-55 pass./trip</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt; 55 pass./trip</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Evening</td>
<td>All</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Owl*</td>
<td>All</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Sat, 7 AM-6 PM</td>
<td>All</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Sat, 6 PM-1 AM</td>
<td>All</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Sun, 7 AM-1 AM</td>
<td>All</td>
<td>0.73</td>
</tr>
<tr>
<td>Running Time</td>
<td>All</td>
<td>short (≤ 20 min.)</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>long (&gt; 20 min.)</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Owl default values are the same for weekdays and weekends*
Step-by-Step Data Collection Program Design Procedure

1. **Determine data needs and acceptable tolerances** based on uses of data.

2. **Select statistical inputs** (i.e. coefficient of variation) based on preliminary data analysis and/or default values.

3. **Select data collection techniques** based on data needs and efficiency of each technique for property.
   
   e.g.  
   - **Baseline:** ridechecks + supplementary point checks  
   - **Monitoring:** pointchecks  
   - **Update:** ride checks

4. **Determine constraining sample sizes** for each technique by route and time period by applying formula.

5. **Determine detailed checker requirements** for each route and time period.

6. **Estimate ratios** (e.g. average fare, trip length, peak load/total passengers) using baseline data for possible use in monitoring.

7. **Revise monitoring plan** (techniques and sample sizes) based on data analysis.
Sample Size Equations

Simple Random Sample:

\[ n = \frac{3.24v^2}{d^2} \quad \text{or} \quad d = \frac{1.8v}{\sqrt{n}} \]

Where

- \( n \) = sample size (number of trips)
- \( d \) = tolerance (e.g. \( d = .05 \) means ± 5% tolerance)
- \( v \) = coefficient of variation

90% confidence level assumed

Notes: assuming 90% confidence level

\( v = \text{coefficient of variation} \)

Required Sample Size for Estimating Averages

<table>
<thead>
<tr>
<th>( v )</th>
<th>0.5</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.3</th>
<th>0.35</th>
<th>0.4</th>
<th>0.45</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>52</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>117</td>
<td>30</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.4</td>
<td>208</td>
<td>52</td>
<td>24</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.5</td>
<td>324</td>
<td>82</td>
<td>36</td>
<td>21</td>
<td>13</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>0.6</td>
<td>467</td>
<td>117</td>
<td>52</td>
<td>30</td>
<td>19</td>
<td>13</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>0.7</td>
<td>636</td>
<td>159</td>
<td>71</td>
<td>40</td>
<td>26</td>
<td>18</td>
<td>13</td>
<td>10</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>0.8</td>
<td>830</td>
<td>208</td>
<td>93</td>
<td>52</td>
<td>34</td>
<td>24</td>
<td>17</td>
<td>13</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>0.9</td>
<td>1050</td>
<td>263</td>
<td>117</td>
<td>66</td>
<td>42</td>
<td>30</td>
<td>22</td>
<td>17</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>1296</td>
<td>325</td>
<td>144</td>
<td>82</td>
<td>52</td>
<td>37</td>
<td>27</td>
<td>21</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>1.25</td>
<td>2025</td>
<td>507</td>
<td>225</td>
<td>127</td>
<td>82</td>
<td>57</td>
<td>42</td>
<td>32</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>1.5</td>
<td>2917</td>
<td>730</td>
<td>324</td>
<td>183</td>
<td>117</td>
<td>82</td>
<td>60</td>
<td>46</td>
<td>37</td>
<td>30</td>
</tr>
</tbody>
</table>
Determine Sample Size in Monitoring Phase Using Conversion Factor

Where

- \( n_2 \) = sample size of auxiliary item in monitoring phase
- \( d_m \) = desired tolerance of the inferred data item

\[
n_2 = \frac{V_x^2(1 + V_R^2)}{0.31d_m^2 - V_R^2}
\]

b. Desired Tolerance of Inferred Item = ±10%

<table>
<thead>
<tr>
<th>( V_x )</th>
<th>0.0001</th>
<th>0.0005</th>
<th>0.001</th>
<th>0.0015</th>
<th>0.002</th>
<th>0.00225</th>
<th>0.0025</th>
<th>0.00275</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>0.20</td>
<td>14</td>
<td>16</td>
<td>20</td>
<td>26</td>
<td>37</td>
<td>48</td>
<td>67</td>
<td>115</td>
</tr>
<tr>
<td>0.30</td>
<td>31</td>
<td>35</td>
<td>43</td>
<td>57</td>
<td>82</td>
<td>107</td>
<td>151</td>
<td>258</td>
</tr>
<tr>
<td>0.40</td>
<td>54</td>
<td>62</td>
<td>77</td>
<td>101</td>
<td>146</td>
<td>189</td>
<td>268</td>
<td>459</td>
</tr>
<tr>
<td>0.50</td>
<td>84</td>
<td>97</td>
<td>120</td>
<td>157</td>
<td>228</td>
<td>295</td>
<td>418</td>
<td>717</td>
</tr>
<tr>
<td>0.60</td>
<td>121</td>
<td>139</td>
<td>172</td>
<td>226</td>
<td>328</td>
<td>425</td>
<td>602</td>
<td>1032</td>
</tr>
<tr>
<td>0.70</td>
<td>16</td>
<td>189</td>
<td>234</td>
<td>307</td>
<td>447</td>
<td>578</td>
<td>819</td>
<td>1404</td>
</tr>
<tr>
<td>0.80</td>
<td>214</td>
<td>247</td>
<td>306</td>
<td>401</td>
<td>583</td>
<td>755</td>
<td>1070</td>
<td>1834</td>
</tr>
</tbody>
</table>

Notes: assuming 90% confidence level

- \( V_x \) = coefficient of variation of auxiliary item
- \( V_R^2 \) = square of coefficient of variation of conversion factor

Nigel Wilson

1.258J/11.541J/ESD.226J

Spring 2010, Lecture 8
Sample Size for Proportions

Using absolute equivalent tolerance (AET),

\[ n = \frac{.96}{AET^2} \]

Recall conversion from tolerance around an expected proportion \( p \) to AET:

\[ AET = \frac{0.5 \, tol}{\sqrt{p*(1-p)}}, \]

where \( p = \) expected proportion

<table>
<thead>
<tr>
<th>p</th>
<th>4%</th>
<th>6%</th>
<th>8%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AET</td>
<td>600</td>
<td>267</td>
<td>150</td>
<td>96</td>
</tr>
<tr>
<td>p = 50%</td>
<td>4%</td>
<td>6%</td>
<td>8%</td>
<td>10%</td>
</tr>
<tr>
<td>p = 60%</td>
<td>3.9%</td>
<td>5.9%</td>
<td>7.8%</td>
<td>9.8%</td>
</tr>
<tr>
<td>p = 70%</td>
<td>3.7%</td>
<td>5.5%</td>
<td>7.3%</td>
<td>9.2%</td>
</tr>
<tr>
<td>p = 80%</td>
<td>3.2%</td>
<td>4.8%</td>
<td>6.4%</td>
<td>8.0%</td>
</tr>
<tr>
<td>p = 90%</td>
<td>2.4%</td>
<td>3.6%</td>
<td>4.8%</td>
<td>6.0%</td>
</tr>
<tr>
<td>p = 95%</td>
<td>1.7%</td>
<td>2.6%</td>
<td>3.5%</td>
<td>4.4%</td>
</tr>
</tbody>
</table>

Rule of thumb:

LARGE sample size needed to estimate a proportion accurately!
Sample Size for Passenger Surveys

- Determine needed sample size for proportion
e.g., proportion of passengers who are pleased, who own a car, etc.

- Multiply SS if proportions are desired for various strata
e.g., proportion of passengers car-owning passengers who are pleased

- Multiply by “clustering effect”
e.g., in on-board survey, 4 responses from same bus may be equivalent to 1 response from a randomly selected rider; clustering effect depends on question
  *if so, expand SS by 4*

- For origin-destination matrix,
  SS = 20 * number of cells (rule of thumb)
  level of detail determined number of cells

- Expand by 1/(response rate)

- Be prepared to revise your expectations when you see how large the needed sample is!
Along with getting correct answers, your primary concern should be getting a high response rate

- Cost: lower response rate means more surveying to get the needed number of responses
- Non-response bias: non-responders may be different from responders, and you’ll never know!
Some non-response bias is predictable and insidious:

- standees are less likely to respond, making close-in origins underrepresented
- low literacy, teens, & non-native population respond less
- predicable biases can be modeled and corrected by numerical procedures

Ways to improve response rate:

- shorten the questionnaire
- quick oral survey: “What station are you going to?”
- get info from counts whenever possible (e.g., fare type)
- distribution method, surveyor training, supervision
- believe and experiment!
The Survey Design Process

1. Define survey objectives
2. Define the population to be surveyed
3. Determine data requirements
4. Specify precision required
5. Select survey instrument
6. Define sampling unit
7. Select sampling procedure and sample size
8. Pretest the survey
9. Develop the survey management process
10. Determine analysis methods
11. Develop data storage and management system