Data Collection Techniques and Program Design

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

Summary of Transit Data Collection Practice
- Transition from manual to automated data collection systems
- Variation in techniques used: automated, manual, mixed
- Statistical validity of sampling approach is often weak
- Inefficient use of data often limits use of analytic planning methods
- ADCS presents major opportunity for strengthening data to support decision-making

Manual vs. Automatic Data Collection

<table>
<thead>
<tr>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>low capital cost</td>
<td>higher capital cost</td>
</tr>
<tr>
<td>high marginal cost</td>
<td>low marginal cost</td>
</tr>
<tr>
<td>small sample sizes</td>
<td>large sample sizes</td>
</tr>
<tr>
<td>aggregate/disaggregate</td>
<td>aggregate/disaggregate</td>
</tr>
<tr>
<td>quantitative/qualitative</td>
<td>quantitative</td>
</tr>
<tr>
<td>unreliable</td>
<td>errors and biases can be estimated and corrected</td>
</tr>
<tr>
<td>limited spatially and temporally</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>not immediately available</td>
<td>available in real-time or quasi-real-time</td>
</tr>
</tbody>
</table>

Key Automated Data Collection Systems
- **Automatic Fare Collection Systems (AFC)**
  - increasingly based on contactless smart cards with unique ID
  - provides entry (exit) information (spatially and temporally) at the individual passenger level
  - traditionally not available in real-time
- **Automatic Vehicle Location Systems (AVL)**
  - bus location based on GPS
  - train tracking based on track circuit occupancy
  - real-time availability of data
- **Automatic Passenger Counting Systems (APC)**
  - bus systems based on sensors in doors with channelized passenger movements
  - passenger boarding (alighting) counts for stops/stations with fare barriers
  - train load weight systems can estimate number of passengers on board
  - traditionally not available in real-time
Data Collection Program Design Process

1. Determine data needs
   - 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

2. Determine property characteristics
   - Assemble available data
   - Determine if a pretest is required
   - Conduct pretest, if necessary
   - Determine statistical inputs for estimating sample size

   If significant change is detected

Data Needs

- **Route (Route Segment) Level**
  - 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

- **System Level**
  - Unlinked Passenger Trips
  - Passenger-miles
  - Linked Passenger trips

Data Inference

<table>
<thead>
<tr>
<th>Auxiliary Data Item</th>
<th>Inferred Data Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFC Boardings</td>
<td>Total Boardings</td>
</tr>
<tr>
<td>APC Counts</td>
<td></td>
</tr>
<tr>
<td>AFC Boardings</td>
<td>Passenger Miles</td>
</tr>
<tr>
<td>APC Counts</td>
<td></td>
</tr>
<tr>
<td>AFC Boardings</td>
<td>Peak Point Load</td>
</tr>
<tr>
<td>APC Counts</td>
<td></td>
</tr>
</tbody>
</table>

Passenger Counting Techniques

- **Manual**
  - Checker (with handheld device)
    - ride check (on/off counts and running time)
    - point check (load on board and headway)

- **Automated**
  - Fare System
    - passenger counts
    - transaction data
  - Automatic Passenger Counters
Manual Data Collection Roles

- If ADCS systems are implemented with analysis and inference tools, the main role for manual data collection is for survey data used to supplement the count-based ADCS information
  - traditional survey methods
  - web-based surveys

Sampling Strategies

- Simple random sampling
  - Every trip has equal likelihood of being included in sample
- Systematic sampling
  - Sample every 6th day – quasi-random, but smooths data collection load
- Cluster sampling
  - Identify natural clusters in advance, select among them at random
  - With passenger surveys, bus trip = cluster of passengers
    - Example: on-board survey
    - Example: sample round trips, or clusters of 4 trips
- Ratio estimation/Conversion factors
  - Estimate relations between data items from comprehensive (baseline) data set
  - Take advantage of complete or less expensive data sources
    - Example: convert AFC 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

Tolerance and Confidence Level

- Accuracy of an estimate has two dimensions
  - "Mean boardings per trip is 33.1."
  - Exactly 33.1?
    - relative tolerance
      - "Mean boardings per trip is 33.1, plus or minus 10%"
    - absolute tolerance
      - "Mean boardings per trip is 33.1, plus or minus 3.3"
      - "Mean percentage of students is 23%, plus or minus 5%"
  - Are you sure?
    - tolerance and confidence level
      - "I'm 95% confident that mean boardings per trip is 33.1, plus or minus 10%"
- To simplify matters
  - hold confidence level fixed (90% or 95%)
  - vary tolerance to reflect different levels of accuracy
  - National Transit Database specification for annual boardings, pass-miles:
    ± 10% relative tolerance at 95% confidence level

\[ t \text{ probability distribution} \]

- Arises when estimating the mean of a normally distributed population with unknown mean and variance, from a sample of size \( n \)

\[
\bar{x} = \frac{\sum x_i}{n} \quad \text{sample mean}
\]
\[
\sigma^2 \approx \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{sample variance}
\]
\[
\sigma^2 \approx \frac{\sigma^2}{n} \quad \text{variance of the distribution of the mean}
\]
\[
\nu = n - 1 \quad \text{degrees of freedom}
\]
Tolerance and Confidence Level

$$d_a = t \frac{\sigma}{\sqrt{n}} = t \frac{\sigma_x}{\sqrt{n}}$$

- Absolute tolerance
- $t$-statistic: number of standard deviations from the mean

Estimating Averages

- Relative tolerances are typically used for averages
  - 5720 boardings ± 5%
  - in this case equivalent to an absolute tolerance of ± 286
- the coefficient of variation is typically easier to guess than the mean and variance separately
  - e.g. with a relative tolerance of ± 5%, a confidence level of 95%, a coefficient of variation of 0.3, and assuming a large sample,
    $$n = \frac{t^2}{d^2} \left( \frac{\sigma_x}{\bar{x}} \right)^2 = \frac{1.96^2}{0.05^2} (0.3)^2 \approx 140$$
  - sample is not that large, with d.f. = 139, $n = 140.732$, so let $n = 150$

Calculating Required Sample Size

When an absolute tolerance is specified:

$$d_a = t \frac{\sigma}{\sqrt{n}} = t \frac{\sigma_x}{\sqrt{n}}$$

$$n = \frac{t^2 \sigma^2}{d^2}$$

When a relative tolerance is specified:

$$d_r = \frac{d_a}{\bar{x}}$$

$$n = \frac{t^2}{d_r^2} \left( \frac{\sigma_x}{\bar{x}} \right)^2$$

Estimating Averages

Simple Random Sample: $n = \frac{3.24\sigma^2}{d^2}$  
$d = \frac{1.8\sigma}{\sqrt{n}}$  
90% confidence level and $n > 11$

Where

- $n = \text{sample size (number of trips)}$
- $d = \text{tolerance (e.g. } d = .05 \text{ means ± 5% tolerance)}$
- $\sigma = \text{coefficient of variation (i.e. ratio of standard deviation to mean)}$

Required Sample Size for Estimating Averages

<table>
<thead>
<tr>
<th>$d = \text{tolerance}$</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu$</td>
<td>1.25</td>
<td>2.02</td>
<td>2.57</td>
<td>3.10</td>
<td>3.65</td>
<td>4.17</td>
<td>4.68</td>
<td>5.19</td>
<td>5.71</td>
<td>6.25</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>1.41</td>
<td>1.77</td>
<td>2.17</td>
<td>2.60</td>
<td>3.08</td>
<td>3.55</td>
<td>4.04</td>
<td>4.56</td>
<td>5.10</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.29</td>
<td>1.64</td>
<td>2.06</td>
<td>2.51</td>
<td>3.00</td>
<td>3.50</td>
<td>4.02</td>
<td>4.56</td>
<td>5.15</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1.10</td>
<td>1.48</td>
<td>1.92</td>
<td>2.39</td>
<td>2.88</td>
<td>3.40</td>
<td>3.93</td>
<td>4.50</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>1.00</td>
<td>1.40</td>
<td>1.85</td>
<td>2.35</td>
<td>2.87</td>
<td>3.41</td>
<td>3.96</td>
<td>4.56</td>
<td>5.17</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.94</td>
<td>1.36</td>
<td>1.82</td>
<td>2.33</td>
<td>2.86</td>
<td>3.41</td>
<td>3.97</td>
<td>4.59</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.90</td>
<td>1.32</td>
<td>1.79</td>
<td>2.31</td>
<td>2.84</td>
<td>3.39</td>
<td>4.00</td>
<td>4.63</td>
<td>5.26</td>
</tr>
</tbody>
</table>
Estimating Proportions

- Consider sampling a group of passengers to estimate the proportion of passengers who are students.
- Each observation is a Bernoulli trial, with probability \( p \) that the passenger is a student.
- We are estimating probability \( p \) of the Bernoulli distribution
  - variance \( \sigma^2 = p(1-p) \) from 0 to 0.25
- Tolerance is typically specified in absolute terms
  - 23% ± 5%

\[
n = \frac{t^2 \sigma^2}{d_e^2} = \frac{t^2}{d_e^2} p(1-p)
\]

Estimating Proportions Using Absolute Equivalent Tolerance

- Collecting data to estimate \( p \) to calculate sample size can be expensive, so often \( p \) is initially assumed
  - worst-case scenario: \( p = 0.5, \sigma^2 = 0.25 \)
- Appropriate tolerance depends on \( p \)
  - smaller tolerances and larger samples are typically needed for extreme \( p \)
    - 32% ± 5% probably OK
    - 1.2% ± 5% probably not OK
- Conveniently, these two factors work in opposite directions
  - if \( p \) is extreme, tolerance must be small but variance will also be small
  - if \( p \) is closer to 0.5, variance will be larger, but tolerance can be larger
- Practical solution: assume \( p = 0.5 \) and work with absolute equivalent tolerance.

<table>
<thead>
<tr>
<th>Expected proportion ( p )</th>
<th>Tolerance corresponding to ±5% AET</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>5.0%</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>4.0%</td>
</tr>
<tr>
<td>90% or 10%</td>
<td>3.0%</td>
</tr>
<tr>
<td>95% or 5%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

\[
d_e = \frac{0.5d_a}{\sqrt{p(1-p)}}
\]

Sample Size for Proportions

- A large sample size is 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 sample sizes 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 sample size by 4
- For origin-destination matrix,
  - sample size = 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!

**Non-Response Bias**

- Non-responders may be different from responders
- Some non-response bias is predictable and insidious
  - standees are less likely to respond, making close-in origins underrepresenting
  - low literacy, teens, and non-native population respond less
  - predictable 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 information from counts whenever possible (e.g. fare type)
  - distribution method, surveyor training, supervision

**Response Rate**

- 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!

**Suggested Tolerances**

- Peak Load (also boardings)
  - Routes with 1-3 buses ±30%
  - Routes with 4-7 buses ±20%
  - Routes with 8-15 buses ±10%
  - Routes with >15 buses ±5%
- Vehicle trip time
  - Routes with trip time ≤ 20 mins ±10%
  - Routes with trip time ≥ 20 mins ±5%
- 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>0.50</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>0.60</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>1.00</td>
</tr>
<tr>
<td>Saturday</td>
<td>All</td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>Sunday</td>
<td>All</td>
<td></td>
<td>0.75</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>0.10</td>
</tr>
</tbody>
</table>

**Conversion Factor Equations**

Compute conversion factor and its coefficient of variation:

\[ R = \frac{\bar{y}}{\bar{x}} \]

where

- \( R \) = conversion factor
- \( \bar{y} \) = average of inferred data item (e.g. boardings) in paired sample
- \( \bar{x} \) = average of auxiliary data item (e.g. load) in paired sample

\[ v_R^2 = \frac{1}{n-1} \left( v_x^2 + v_y^2 - 2 v_x v_y r_{xy} \right) \]

\[ r_{xy} = \frac{S_{xy}}{S_x S_y} \]

\[ S_{xy} = \text{Cov}(x, y) = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{n-1} \]

where

- \( v_R \) = coefficient of variation of conversion factor
- \( v_x \) = coefficient of variation of auxiliary item (e.g. load)
- \( v_y \) = coefficient of variation of inferred item (e.g. boardings)
- \( r_{xy} \) = correlation coefficient between auxiliary and inferred items
- \( n \) = number of paired observations in sample

**Step-by-Step Manual Data Collection Program Design**

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.
   - 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.

**Determine Sample Size in Monitoring Phase**

\[ n_2 = \frac{v_x^2 (1 + v_R^2)}{0.31 d_{yu}^2 - v_R^2} \]

(90% confidence level assumed)

where

- \( n_2 \) = sample size of auxiliary item in monitoring phase
- \( d_{yu}^2 \) = desired tolerance of the inferred data item

**Desired tolerance of the inferred data item = ±10%**

<table>
<thead>
<tr>
<th>( V_x^2 )</th>
<th>0.0001</th>
<th>0.0005</th>
<th>0.0010</th>
<th>0.0015</th>
<th>0.0020</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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>0.60</td>
<td>123</td>
<td>139</td>
<td>173</td>
<td>226</td>
<td>328</td>
<td>425</td>
<td>602</td>
</tr>
<tr>
<td>0.70</td>
<td>161</td>
<td>189</td>
<td>234</td>
<td>307</td>
<td>447</td>
<td>578</td>
<td>819</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>
</tr>
</tbody>
</table>