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:

- **Furth, P., B. Hemily, T.H.J. Muller, and J.G. Strathman, "Using Archived AVL-APC Data to Improve Transit Performance and Management." Transportation Research Board, TCRP Report 113, 2006.**
- **Zhao, J., A. Rahbee, and N.H.M. Wilson, "Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems." Computer-Aided Civil and Infrastructure Engineering 22 (2007), 376–387.**

Summary of Transit Data Collection Programs

Great variation in data collection resources

Variation in techniques used: automated, manual, mixed

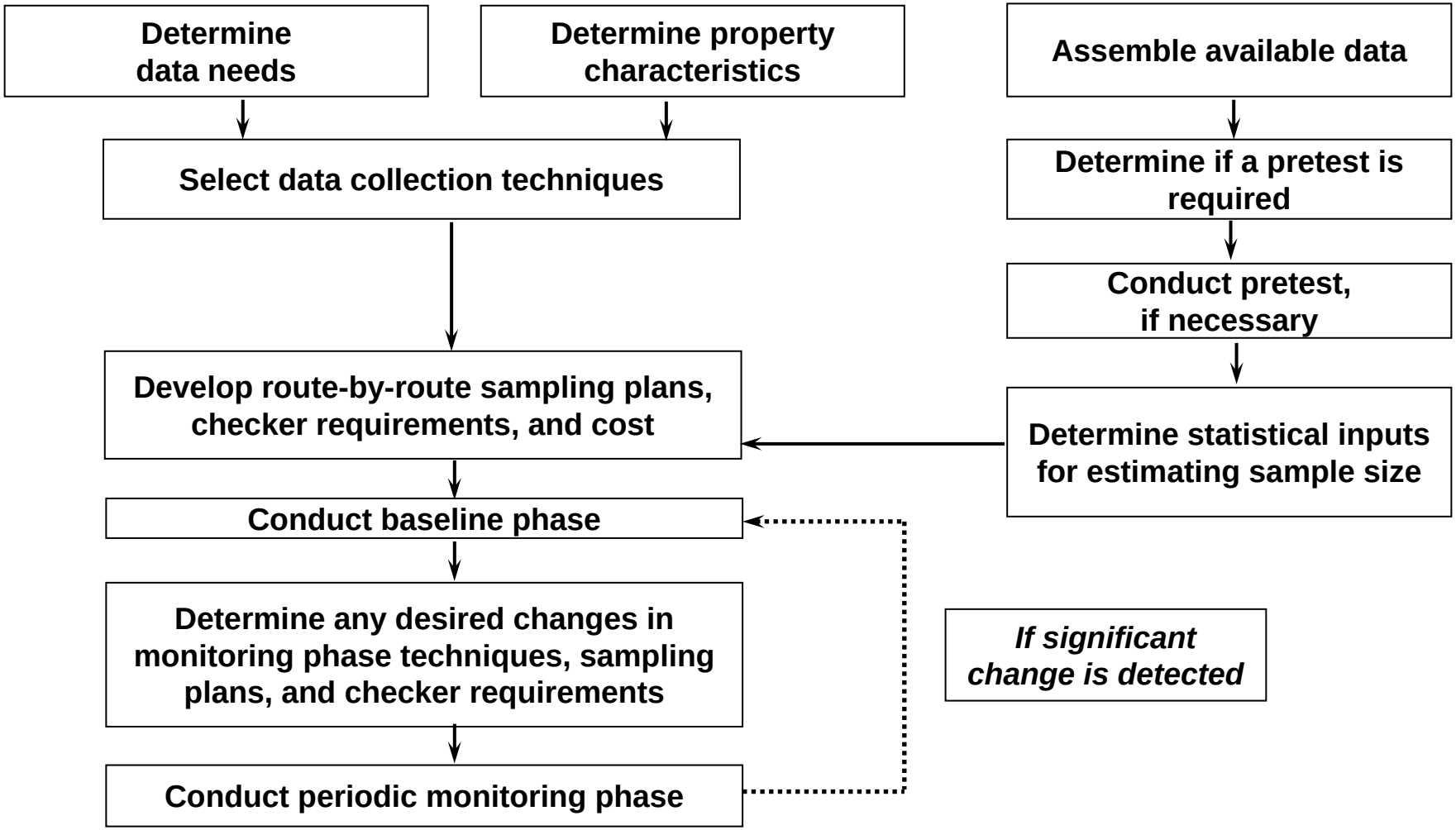
Statistical validity of sampling approach varies

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



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

Auxiliary Data Item	Inferred Data Item
Load or Revenue	Boardings
Boardings, Load or Revenue	Passenger Miles
Point Load	True Maximum Load
Revenue	Peak Point Load

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

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

* *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

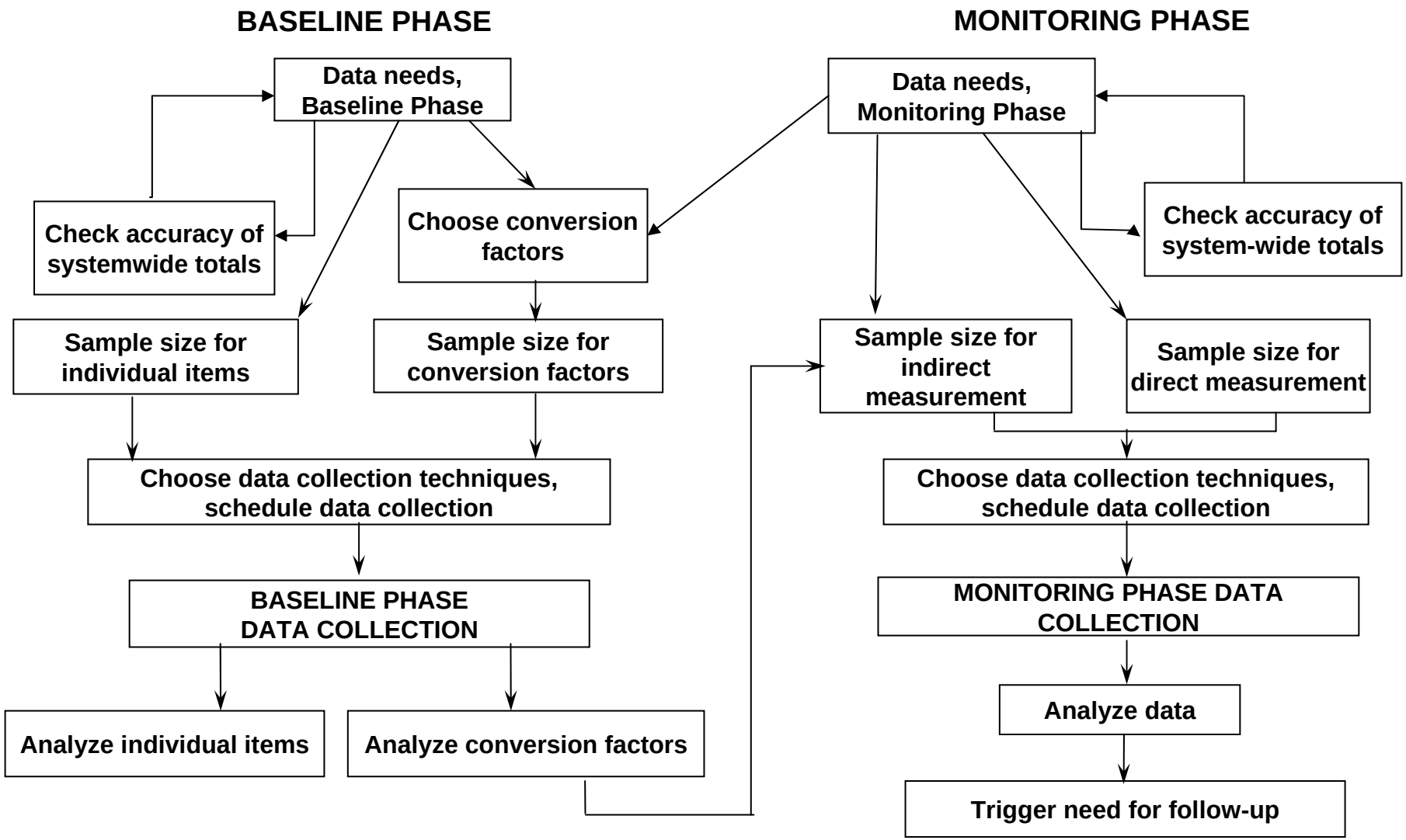
- reduce real-time coding errors
- error detection
- load checks

Have checkers code their own data

- immediate graphical feedback

Watch for fabricated data

Designing a Data Collection Program



Sampling Strategies

Simple random sampling

Every trip has equal likelihood of selection

Systematic sampling

Sample every 6th day – like random, but smoothes data collection load

Example: FTA Circular

Cluster sampling

Identify natural clusters in advance, select them at random

With passenger surveys, bus trip = cluster of passengers

Example: on-board survey
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

Precision and Confidence Level

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 tol / sqrt[p*(1-p)], where p = expected proportion

AET = 0.5 (4%) / sqrt[0.8*(0.2)] = 5%

Desired Accuracy (AET)

Tolerance naturally improves as the proportion moves toward the extremes (0%, 100%). AET is the tolerance you'd get if the proportion were 50%.

Expected proportion	Tol. corresponding to ±5% AET
50%	5%
60% or 40%	4.9%
70% or 30%	4.6%
80% or 20%	4%
90% or 10%	3%
95% or 5%	2.2%

Recommended 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

Data Item	Time Period	Route Classification	Default Value
Maximum Load	Peak	< 35 pass./trip	0.5
		≥ 35 pass./trip	0.35
	Off- Peak	< 35 pass./trip	0.6
		35-55 pass./trip	0.45
		> 55 pass./trip	0.35
	Evening	All	0.75
	Owl*	All	1
	Sat, 7 AM-6 PM	All	0.6
	Sat, 6 PM-1 AM	All	0.75
	Sun, 7 AM-1 AM	All	0.75

*Owl default values are the same for weekdays and weekends

Default Values for Coefficient of Variation of Key Data Items

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

*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:

$$n = \frac{3.24y^2}{d^2} \quad or \quad d = \frac{1.8v}{\sqrt{n}}$$

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 = coefficient of variation

Required Sample Size for Estimating Averages

	d = tolerance									
v	**0.5**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**
0.1	13	4	2	1	1	1	1	1	1	1
0.2	52	13	6	4	3	2	2	1	1	1
0.3	117	30	13	8	5	4	3	2	2	2
0.4	208	52	24	13	9	6	5	4	3	3
0.5	324	82	36	21	13	10	7	6	5	4
0.6	467	117	52	30	19	13	10	8	6	5
0.7	636	159	71	40	26	18	13	10	8	7
0.8	830	208	93	52	34	24	17	13	11	9
0.9	1050	263	117	66	42	30	22	17	13	11
1	1296	325	144	82	52	37	27	21	17	13
1.25	2025	507	225	127	82	57	42	32	25	21
1.5	2917	730	324	183	117	82	60	46	37	30

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

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

b. Desired Tolerance of Inferred Item = ±10%

V_x \ V_R^2	.0001	.0005	.001	.0015	.002	.00225	.0025	.00275
0.10	4	4	5	7	10	12	17	29
0.20	14	16	20	26	37	48	67	115
0.30	31	35	43	57	82	107	151	258
0.40	54	62	77	101	146	189	268	459
0.50	84	97	120	157	228	295	418	717
0.60	121	139	172	226	328	425	602	1032
0.70	16	189	234	307	447	578	819	1404
0.80	214	247	306	401	583	755	1070	1834

Notes: assuming 90% confidence level

V_x = coefficient of variation of auxiliary item V_R^2 = square of coefficient of variation of conversion factor

Sample Size for Proportions

Using absolute equivalent tolerance (AET),

$$n = .96/AET^2$$

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

$$AET = 0.5\ tol\ /\ sqrt[p*(1-p)],$$

where *p* = expected proportion

Rule of thumb:

LARGE sample size needed to estimate a proportion accurately!

n	600	267	150	96
AET	4%	6%	8%	10%
p = 50%	4%	6%	8%	10%
p = 60%	3.9%	5.9%	7.8%	9.8%
p = 70%	3.7%	5.5%	7.3%	9.2%
p = 80%	3.2%	4.8%	6.4%	8.0%
p = 90%	2.4%	3.6%	4.8%	6.0%
p = 95%	1.7%	2.6%	3.5%	4.4%

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!**

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!***

Response Rate

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

1.258J / 11.541J / ESD.226J Public Transportation Systems

Spring 2010