Exploring New Methods of Data Gathering in Long-Distance Passenger Travel Data

Ben Pierce,
Battelle
June 8, 2011
Why Are New Methods Needed?

• Declining Response Rates
  – Survey saturation in US
  – Growing cell-only households (coverage issues)
  – Telephone Numbers no longer tied to geography

• Tightening of Available Resources
  – Telephone only recruitment no longer cost efficient
  – Increasing use of Internet, social media, telephone as communication (versus telephone)

• Frequency of long-distance trips
  – Impacts recall period
  – Impacts respondent burden
A Successful Long-Distance Travel Survey Requires Many Pieces to Fit Perfectly

- Questionnaire Design
- Data Cleaning/Editing
- Managing Non-Response
- Data Collection Methods
- Sample Frame/Design
- Weighting and Statistical Estimation
A Successful Long-Distance Travel Survey Requires Many Pieces to Fit Perfectly

- Questionnaire Design
- Managing Non-Response
- Data Collection Methods
- Sample Frame/Design
- Weighting and Statistical Estimation
- Data Cleaning/Editing
Key Technical Challenge: Data Collection

- Frequency of Long-Distance Trips
  - 1995 ATS implies >75% of households did not take a long distance trip in their assigned year
    - Long Distance trip = 100 miles
  - 2001 NHTS ~50% of households did not take a long distance trip in their 4 week recall period
    - Long Distance trip = 50 miles

- Historical data collection approach cannot adjust for this infrequency of travel without introducing (significant) recall bias
How can we solve this problem?

• Collect data more frequently!
  – Use event driven data collection (pulse survey)
  – Leverage technology
  – Effectively shorten recall period

• The difficulty is that you have to do this….
  – Without increasing respondent burden
  – Within limited resource constraints
  – Without introducing any bias
  – In a fashion so that you can maximize the amount and nature of the data you are collecting
Potential Solutions #1: Trip Memory Jogger

- Person-Based GPS Trip Memory Jogger
  - Passive or active GPS tracking of distances traveled
  - Smartphone or similar device

- Advantages:
  - Has proven successful in capturing trips in daily travel surveys
  - Regardless of GPS accuracy, distance still more accurate than personal recall

- Disadvantages:
  - May miss trips if technology
  - Could still incur recall bias depending upon when data sorted and gathered
  - May require extensive post-processing of data
Potential Solution #2: Event Triggered Data Collection

- Person-Based GPS or Cellular Geofence
  - Trigger notification to data collection center when distance threshold exceeded
  - Distance threshold established upon home address

- Advantages:
  - Less invasive; not reliant on GPS, Cellular geofence might be sufficient
  - Less post-processing of data
  - Triggers survey shortly after event occurs (efficiently identifies event of interest)
  - May require only “simple” cell phone, not “smart phone”

- Disadvantages:
  - May miss events if technology not used
  - Could miss segments of population without access to technology
  - Cellular geofence may require access to cellular information
Potential Solution #3: Use of Social Media to Reduce Recall Period

• Use Social Media to “connect” to participants
  – Data mine social content to identify past, current, or future long-distance trip events
  – Use social media for self-reporting of trips/events
  – Use social media as data collection mechanism

• Advantages
  – Already established industries keyed to mining this media for events
  – Well suited to most mobile population (young adults)
  – Relatively inexpensive; could improve response rates because of widespread acceptance/use

• Disadvantages
  – May miss trips that are not discussed; or impose a reporting bias
  – May not be suitable for all segments of population (viewed as privacy invasion)
What is the Reality?

• Combination of historical and new data collection methods will be needed
  – Telephone follow-up
  – Social Media follow-up, communication, and data collection
  – Event-driven surveys
    - GPS triggered
    - Search Engine/ticket purchase triggered
    - Etc.
Key Technical Challenge: Sample Frame/Design

• Lack of an efficient method for identifying and selecting a probabilistic sample
• Incomplete frames/non-inclusive sampling designs increase risk of coverage bias
Traditional Sampling Frames/Design for Household Travel Surveys

- Area Sampling/Address Frame
- Telephone Frame
- Hybrid Frame/Design
- E-mail List Augment
- Cell-Phone Augment
Alternatives to Traditional Sampling Designs: Link Trace Sampling

• Network Sampling (Multiplicity Sampling)
  – Establish a network of individuals who are related in some manner to an individual selected for a particular survey
  – The nature of the relationship of the selection unit to the other individuals in the network are referred to as the counting rules
  – Technique used since 1980’s for rare incidents (typically medical conditions)
  – Requires that data received from everyone “in-network”
Alternatives to Traditional Sampling Designs: Link Trace Sampling

• Respondent Driven Sampling
  – Recruit a small set of “seed” respondents
  – Provide an incentive for the seed to recruit others
    — Recruitment has been shown to be equivalent to a Markov Chain (asymptotically unbiased estimates)
    — Can be challenging to estimate the probabilities of selection accurately, could introduce bias in estimation

• Adaptive Cluster Sampling – similar
Alternatives to Traditional Sampling Designs: Abandon Household Frames

- Sample trips in process
  - Roadside, Rest Station or other roadway, airport, Train terminal intercepts
  - Sample ticket purchases
  - Intercepts along corridors
  - Area-based sampling, but of facilities not households
  - Possible to augment with sensor data
    - Bluetooth detection
    - License plate reads
    - etc.

- National Level O/D Corridor Survey
Weighting and Statistical Estimation

• May need to increase the sophistication of the weighting and analysis methods used for long-distance travel

  – Weighting
    - More of an issue because of lack of Census
    - Weighting a sample to a sample
    - Rare long-distance travel events can be very influenced by extreme weights (more so than daily travel)
    - New methods for weighting, trimming, and validating survey weights need to be developed

  – Statistical Estimation
    - Better estimates need to be obtained through more sophisticated models at the tradeoff of more assumptions needed
Summary

• Traditional sampling design and data collection methods **will not** be suitable for a modern long-distance travel survey

• Alternatives must be investigated including:
  – Technology alternatives
  – Methods for improving response rates/sampling frame coverage
  – Data weighting and estimation

• The days of a “simple” address/telephone design are gone; there will be no quick-fix solution to these issues