Transportation Data

CIVE 461: Urban Transportation Planning
Supplemental Notes

Three Basic Data Types

Transportation supply

Transportation demand

Land use

Transportation Supply Data

Road link attributes (**OSM** and local government databases)

- Length
- Capacity
- Free flow speed
- Downstream conditions
- Surface type and quality
- Tolls (if any)

Transit link attributes (**GTFS**)

- Lengths
- Capacities (vehicle frequency x vehicle speed)
- Stations/stops
- Fares and fare structures

Operating costs (**AAA** and other sources)

- Gas price
- Maintenance
- Repairs
- Insurance

	Small Sedan	Medium Sedan	Large Sedan	Small SUV (FWD)	Medium SUV (4WD)
Operating Costs					
fuel	7.15¢	8.31¢	11.43¢	8.27¢	11.25¢
maintenance	8.39¢	9.56¢	9.69¢	9.48¢	10.11¢
cost per mile	**15.54¢**	**17.87¢**	**21.13¢**	**17.76¢**	**21.37¢**
Ownership Costs					
full-coverage insurance	$1,342	$1,245	$1,264	$1,087	$1,118
license, registration, taxes	$560	$730	$934	$712	$968
depreciation (15k mi/yr)	$2,769	$3,394	$4,914	$3,234	$3,824
finance charge	$513	$684	$900	$665	$921
cost per year	**$5,185**	**$6,054**	**$8,013**	**$5,699**	**$6,831**
cost per day	**$14.20**	**$16.59**	**$21.95**	**$15.61**	**$18.72**
Total Cost Per Mile – 10k mi/yr					
cost per mile driven	$1,554	$1,787	$2,113	$1,776	$2,137
cost per year	$5,185	$6,054	$8,013	$5,699	$6,831
depreciation[1]	-$217	-$233	-$304	-$286	-$382
total cost per year	**$6,521**	**$7,607**	**$9,822**	**$7,188**	**$8,586**
total cost per day	**$17.87**	**$20.84**	**$26.91**	**$19.69**	**$23.52**
total cost per mile[2]	**$0.6521**	**$0.7607**	**$0.9822**	**$0.7188**	**$0.8586**
Total Cost Per Mile – 15k mi/yr					
cost per mile driven	$2,331	$2,681	$3,169	$2,663	$3,205
cost per year	$5,185	$6,054	$8,013	$5,699	$6,831
total cost per year	**$7,516**	**$8,734**	**$11,182**	**$8,362**	**$10,036**
total cost per day	**$20.59**	**$23.93**	**$30.64**	**$22.91**	**$27.50**
total cost per mile[2]	**$0.5010**	**$0.5823**	**$0.7455**	**$0.5575**	**$0.6691**
Total Cost Per Mile – 20k mi/yr					
cost per mile driven	$3,108	$3,574	$4,226	$3,551	$4,273
cost per year	$5,185	$6,054	$8,013	$5,699	$6,831
depreciation[1]	+$235	+$252	+$328	+$310	+$414
total cost per year	**$8,528**	**$9,880**	**$12,566**	**$9,560**	**$11,518**
total cost per day	**$23.36**	**$27.07**	**$34.43**	**$26.19**	**$31.56**
total cost per mile[2]	**$0.4264**	**$0.4940**	**$0.6283**	**$0.4780**	**$0.5759**

	Minivan	1/2 Ton/Crew Pickup (4WD)	Hybrid Vehicle	Electric Vehicle	2020 Weighted Average
Operating Costs					
fuel	11.22¢	15.23¢	5.20¢	3.67¢	10.66¢
maintenance	9.42¢	8.80¢	8.19¢	7.46¢	9.12¢
cost per mile	**20.64¢**	**24.02¢**	**13.38¢**	**11.13¢**	**19.79¢**
Ownership Costs					
full-coverage insurance	$1,096	$1,242	$1,212	$1,227	$1,202
license, registration, taxes	$855	$1,127	$726	$74	$851
depreciation (15k mi/yr)	$4,250	$4,267	$3,519	$5,323	$3,721
finance charge	$803	$1,069	$684	$826	$819
cost per year	**$7,004**	**$7,705**	**$6,141**	**$7,450**	**$6,593**
cost per day	**$19.19**	**$21.11**	**$16.82**	**$20.41**	**$18.06**
Total Cost Per Mile – 10k mi/yr					
cost per mile driven	$2,064	$2,402	$1,338	$1,113	$1,979
cost per year	$7,004	$7,705	$6,141	$7,450	$6,593
depreciation[1]	-$264	-$454	-$255	-$370	$335
total cost per year	**$8,805**	**$9,653**	**$7,224**	**$8,193**	**$8,236**
total cost per day	**$24.12**	**$26.45**	**$19.79**	**$22.45**	**$22.57**
total cost per mile[2]	**$0.8805**	**$0.9653**	**$0.7224**	**$0.8193**	**$0.8236**
Total Cost Per Mile – 15k mi/yr					
cost per mile driven	$3,096	$3,603	$2,007	$1,669	$2,968
cost per year	$7,004	$7,705	$6,141	$7,450	$6,593
total cost per year	**$10,101**	**$11,308**	**$8,148**	**$9,119**	**$9,561**
total cost per day	**$27.67**	**$30.98**	**$22.32**	**$24.98**	**$26.19**
total cost per mile[2]	**$0.6734**	**$0.7539**	**$0.5432**	**$0.6079**	**$0.6374**
Total Cost Per Mile – 20k mi/yr					
cost per mile driven	$4,129	$4,804	$2,677	$2,225	$3,957
cost per year	$7,004	$7,705	$6,141	$7,450	$6,593
depreciation[1]	+$286	+$495	+$277	+$395	$364
total cost per year	**$11,419**	**$13,004**	**$9,094**	**$10,070**	**$10,914**
total cost per day	**$31.29**	**$35.63**	**$24.91**	**$27.59**	**$29.90**
total cost per mile[2]	**$0.5710**	**$0.6502**	**$0.4547**	**$0.5035**	**$0.5457**

Transportation Demand Data

Roadway traffic counts

- Vehicle types
- Instantaneous flows (15 minutes, hourly, etc.) – permanent counters, spot counts, screenlines, and cordons
- Intersection counts
  - Turning movements
  - Through movements

Transit person counts

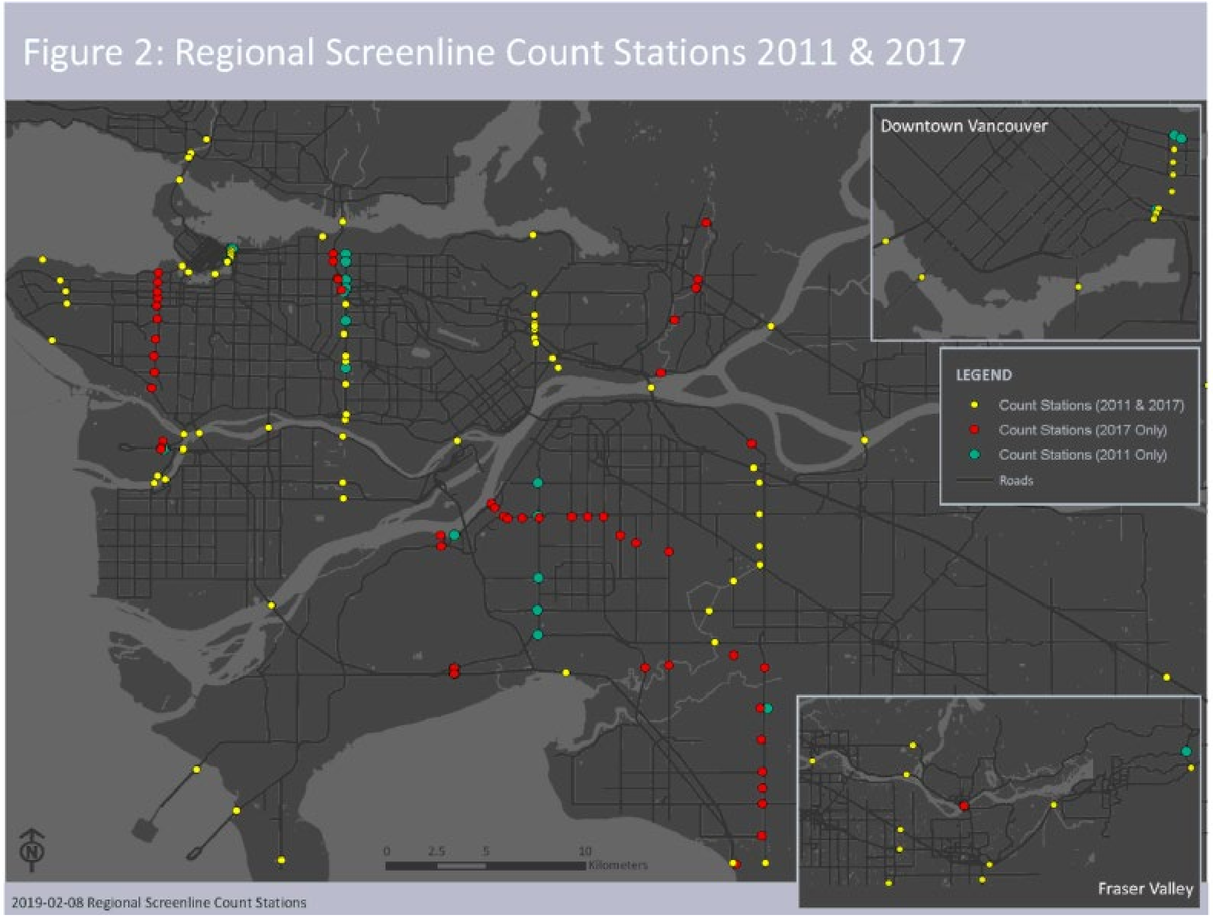
- Boarding and alighting
- Origin-destination (**Bus stop? True origin?)**
- Fare-based estimation
- Transit roadside estimations

Origin-Destination Studies

- A form of intercept survey where respondents are asked their origin and destination
- Not as detailed as home travel surveys
- For road safety reasons, often requires police presence, making respodents more nervous

Cordon Counts

Screenline Counts



Household Travel Surveys

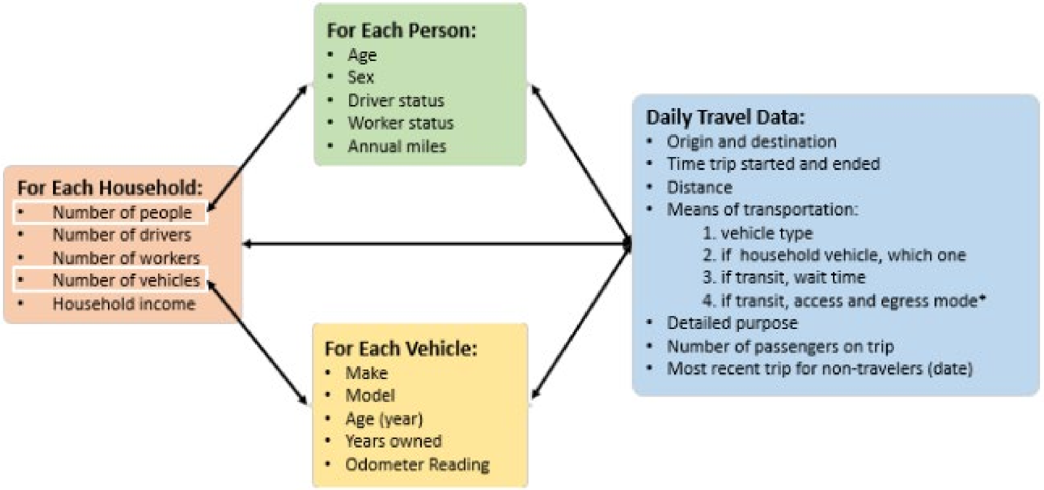
- All travel or activity by household members
- Assign each household a survey day
- **Collection methods**:
  - Home/personal interview
  - Telephone interview
  - Mail survey
- **Person data**: age, gender, work status, school status, transit pass, driver's license, parking
- **Household data**: number of members, number of vehicles, income, location and type of dwelling
- **Vehicle data:** make, model, year, powertrain, odometer reading

Household Travel Surveys

- Questionnaires typically designed to minimize respondent resistance
  - **Income questions** put at the end
  - Questions should be **simple** and **direct**
  - **Activity-recall recommended**: think about activities and when they started/ended rather than trips
  - Assigning a day in advance can help with recall and respondents can be provided a paper diary to fill out and enter online/via phone later

NHTS Travel Diary

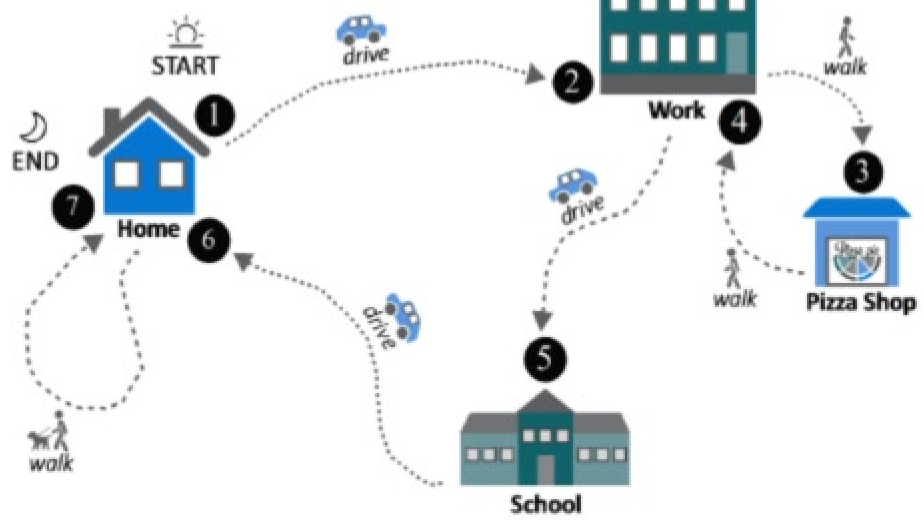
- Most recently available survey conducted in **2017 (current cycle began Jan. 18 2022)**
- Sponsored by **FHWA**
- Collects data on diary days throughout the year (both weekdays and weekends)
- Collected from **stratified random sample** of households in all 50 states and District of Columbia
- 26,000 households and 103,112 additional **Add-on samples** by 13 states and MPOs
- Data distributed in four files: households, persons, trips, and vehicles
- Files linked by common **IDs** and include sample **weight columns**



NHTS Travel Diary - 2017

University of Nebraska Lincoln

A Full Travel Day Example

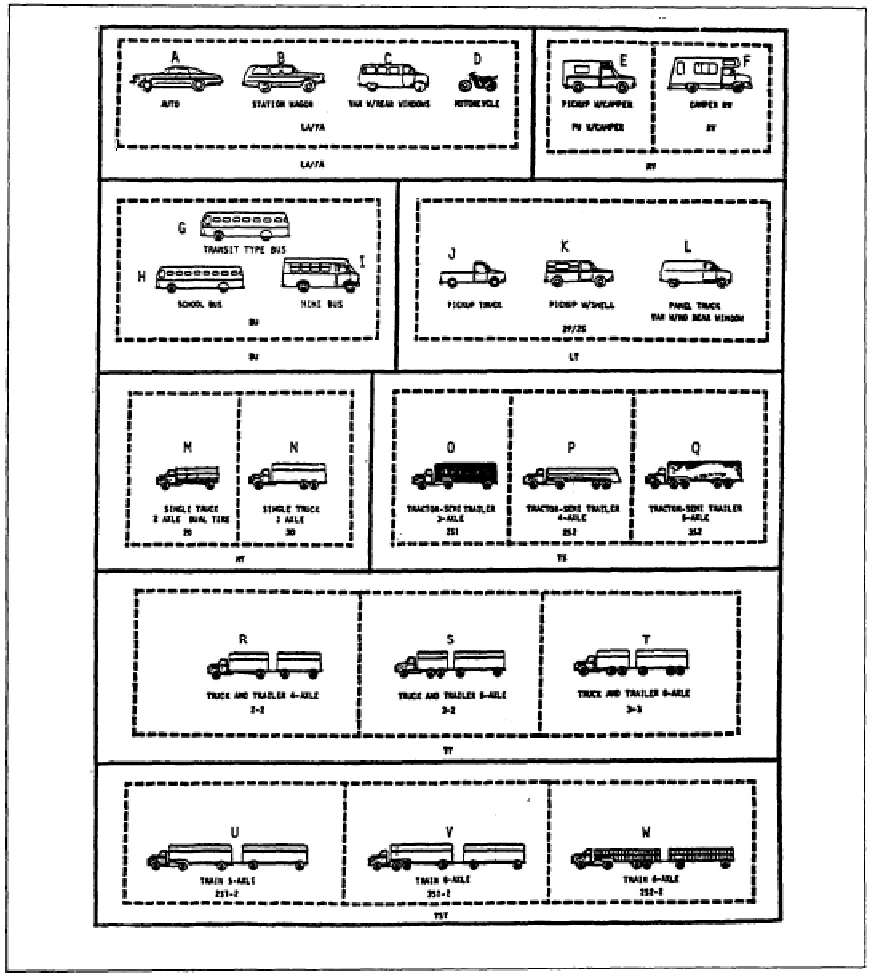


1 Where did you go?	2 How did you get there?				3 What did you do?
START HERE	What time did you arrive at this place?	How did you get to this place? (for example, walk, car, bus, train, etc.)	How many people went with you to this place?	What time did you leave this place?	What did you do at this place? Use the Activity List.
Place 1: Where were you at 4:00 AM on your assigned travel day? Provide place name and address/intersection: Home				7:31 ☒ AM ☐ PM ☐ Did not leave	01. Ate breakfast and got ready for work
Place 2: Where did you go next? Provide place name and address/intersection: Work — Arbor Law Firm 990 Central Ave, Chicago, IL 60639	7:54 ☒ AM ☐ PM	Drove my car	0	12:45 ☐ AM ☒ PM ☐ Did not leave	03. Work
Place 3: Where did you go next? Provide place name and address/intersection: Gustoso's Pizza 1800 Kerry Lane, Chicago, IL 60639	12:58 ☐ AM ☒ PM	Walked	2	1:45 ☐ AM ☒ PM ☐ Did not leave	13. Buy and eat lunch
Place 4: Where did you go next? Provide place name and address/intersection: Work — Arbor Law Firm 990 Central Ave, Chicago, IL 60639	2:02 ☐ AM ☒ PM	Walked	2	4:52 ☐ AM ☒ PM ☐ Did not leave	03. Work
Place 5: Where did you go next? Provide place name and address/intersection: Fairview Elementary 7590 North Rd. Chicago IL 60639	5:08 ☐ AM ☒ PM	Drove my car	0	5:20 ☐ AM ☒ PM ☐ Did not leave	06. Pick up daughter from school
Place 6: Where did you go next? Provide place name and address/intersection: Home	5:54 ☐ AM ☒ PM	Drove my car	1	7:30 ☐ AM ☒ PM ☐ Did not leave	01. Ate dinner and relaxed
Place 7: Where did you go next? Provide place name and address/intersection: Home	8:04 ☐ AM ☒ PM	Walked	0	☐ AM ☐ PM ☒ Did not leave	16. Walk the dog and exercised

Freight Surveys

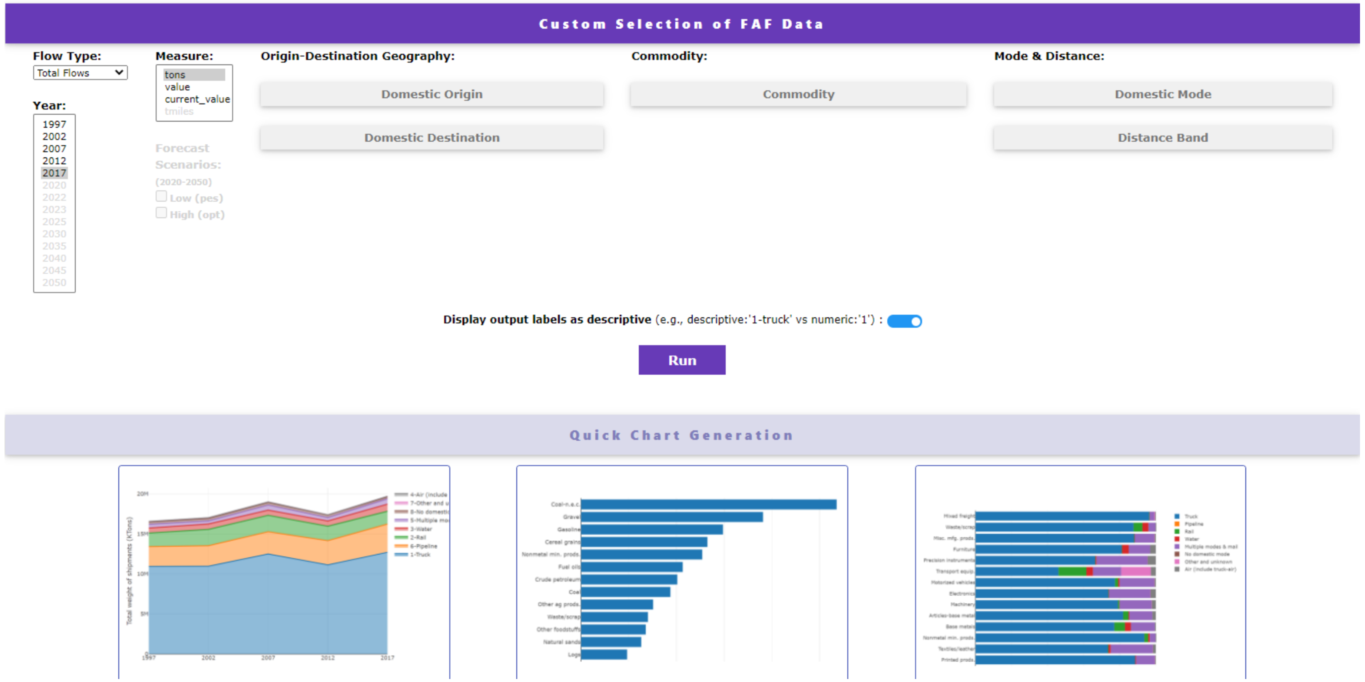
- **Truck intercept**: stop trucks on roadway for interview
- Business establishment interview (not as common)
- Business location and industry classification (according to NAICS or SIC)
- Number of employees and distribution of occupations
- Vehicle classification

Figure 9.1 Sample Commercial Vehicle Survey Travel Diary Package from Phoenix (continued)



Freight Analysis Framework (FAF)

- Integrates multiple datasets to provide comprehensive database of freight movements among states and major metropolitan areas
- Major data sources are commodity flow survey and international trade data from Census Bureau
- Data provided for origin-destination pairs by tonnage, dollar value, and miles by mode



Zone System

- Zoning used to aggregate individual households and establishments into manageable chunks for modeling
- Need to determine number and size of zones
- Ideally, zones should be designed to accommodate future population growth – e.g., smaller zones in areas where they will be needed in the future to accommodate new population
- Regional extent often dictated by political boundaries: MPO, city, or county limits

Location	Population	Number of zones
London (2006)	7.2 million	2252
		~1000
		~230
		52
Montréal (2008)	3.4. million	1425
Leeds UK (2009)	0.7 million	~560
Santiago (2009)	5.5 million	~700
Dallas-Forth Worth (2004)	6.5 million	4875
Washington DC (2008)	6.5 million	~2200
		463
Bogotá (2000)	6.1 million	637
Dublin (2010)	1.7 million	~650
Sydney (2006)	3.6 million	2690

Land Use Data

- **Population** and **employment** totals by zone
- Employment according to **SIC** or **NAICS** classification
- **Floorspace** and **land prices** from real estate transactions and tax assessment evaluations
- **Land** and **space development history** from development permit records
- Land use classification from **zoning** ordinances

Passive Data – GPS Traces

- Provides **start/end location with timestamp**
- **Significant** data processing task
- Variable data quality
  - Cellular communications: linked to cell tower locations – 100 meters to 2 km
  - Global position system (GPS): based on line-of-sight with satellites and primary from in-vehicle devices (9-12% of trucks and <1% of passenger vehicles) – precision of 1-10 meters
  - Location-based services (LBS): based on best available location information (mix of GPS, wifi, Bluetooth, and cellular data) – 10 to 100 meters
- **No sociodemographics**
- **No trip purposes**

Passive Data – Smart Cards

- Transit agencies using electronic fare payment systems
- Data provides start/"tap on" point and may provide end/"tap off" point with timestamp
- Not true trip start/end because lacks transit access data
- **No sociodemographics**
- **No trip purposes**
- Can use **data fusion** methods in combination with travel diary

Passive Data – Ride-Hailing and Micro-Mobility

- Provides trip start and end with timestamp
- Often available on local government open data portals
- **No sociodemographics**
- **No trip purposes**

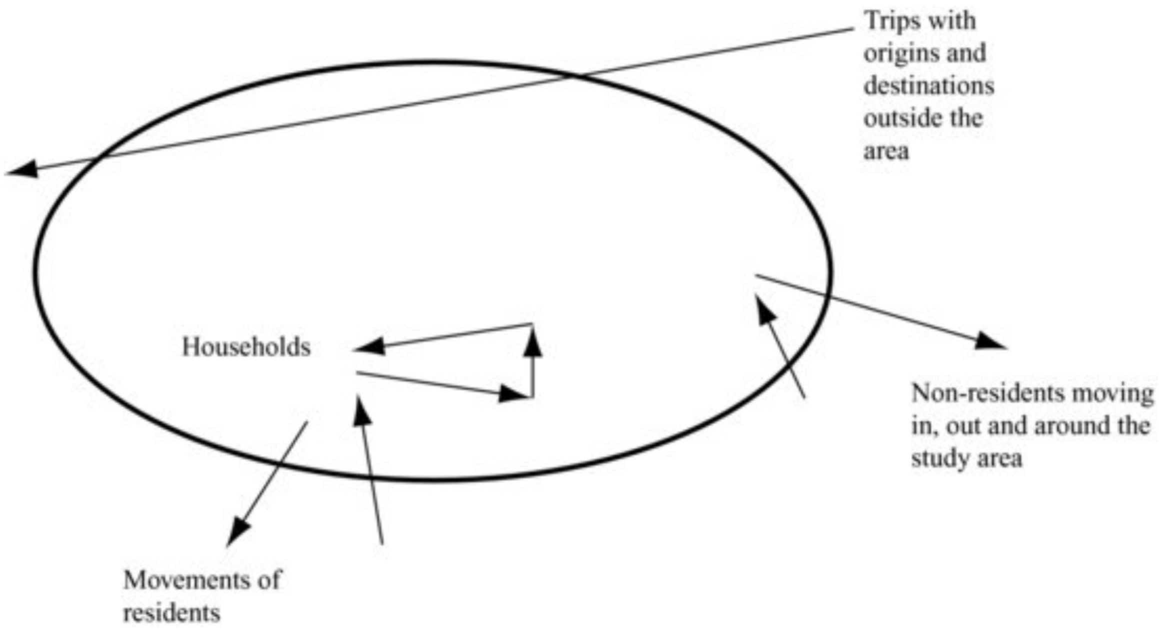
Other Useful Datasets

- Decennial U.S. census
- American Community Survey (ACS)
- American Time Use Survey (ATUS)
- EPA Smart Location Mapping
- Local open data portals
- TIGER/Line shapefiles
- IPUMS
- INRIX, Streetlight Data, and other private data providers

Useful Tools for Data Processing

- **Data processing:** R and Python common alternatives to Excel (particularly for larger datasets)
- **GIS:** QGIS a free alternative to ArcGIS, MapInfo, etc.
- **Data storage:** Postgresql, MySQL, SQLite, etc.
- Web scraping with R or Python
- **“Big Data” cluster computing tools**: Apache Hadoop, Apache Spark, Google BigQuery

Data Scope for an Urban O-D Study



Household survey: captures trips by residents in/out study area

Intercept/external cordon surveys: people crossing study area border, particularly non-residents

Internal cordons/screenlines: validate model trip estimates against "ground truth"

Travel time surveys: travel time by mode to calibrate and validate models

Other related data:

- **Land-use inventory**: residential zones (housing density), commercial and industrial zones (by type of establishment), parking spaces
- **Infrastructure and existing services inventories:** public and private transport networks, fares, frequency, traffic signal locations and timings

Data Types & Models

Qualitative data:
- **Categorical** like color or region

Quantitative:
- Categorical (**discrete**) data and **continuous** data
  - Most data not 100% continuous, such as trips, cars, sites to visit, etc. or on a limited range like a proportion between 0 and 1

- **Discrete data models**: choice models, nonchoice models, models of count processes

- **Nonchoice models**: contingency tables or cross-classification data analysis

- **Count models**: Poisson, negative binomial, etc.

- **Ordered response models**: logit and probit

Statistical Considerations

- Data often consist of a sample of observations from a population of interest – often not economically (or even technically) feasible to observe population
- How to ensure representative sample?
- How to extract valid conclusions from a sample satisfying statistical conditions?
- **Sample**: Collection of units that is selected to **represent** a larger population with certain attributes (e.g., height, age, income)
  - What population does it represent?
  - How large should the sample be?
  - What is meant by “selected”?

Statistical Considerations

- **Population of interest**: complete group about which information is sought
  - May be selected using a **sampling unit** that does not correspond to the unit of interest
  - E.g., select households but interested in individuals

- **Sampling method**:
  - Random sampling
    - **Simple random sampling**: assign a number to each sampling unit and randomly sample X units
    - **Stratified random sampling**: a priori information first used to subdivide population into homogenous strata then simple random sampling applied within strata. E.g., households with 1, 2, 3, 4, 5+ members

Sampling Considerations

- **Sampling error**: error due to dealing with a sample rather than full population
- **Sampling bias**: caused by mistakes made either when defining population of interest or selecting sampling method, data collection technique, or another part of process
- **Sampling bias** differs from **sampling error** in two ways
  - Can affect not only variability around mean but values themselves
    - More severe distortion of results
  - **Sampling error** cannot be avoided (only reduced through collecting a larger sample), **sampling bias** can be virtually eliminated through careful experimental design
- **Sample size**: no simple formula to decide sample size
  - Too large and it's **expensive** but too small and subject to **larger variability** in statistical results

Sample Size to Estimate Population Parameters

Based on three factors:

1. Variability of parameters in population under study

2. Required degree of accuracy

3. Population (least important, as we shall see)

- Central limit theorem (CLT) postulates that estimates of mean from a sample tend towards **Normal** as sample size (n) increases

- Holds for **any population distribution** if n is greater than 30 (and smaller if sample has normal-like properties)

Sample Size to Estimate Population Parameters

- Consider a population N and specific property with mean μ and variance σ^2
- Sample mean $\bar{x}$ distribution from **successive samples** is distributed Normal

$$se(\bar{x}) = \sqrt{\frac{(N-n)\sigma^2}{n(N-1)}}$$

- If we have only **one sample**, best estimate for μ is $\bar{x}$ and best estimate for σ^2 is s^2 (sample variance) and we have

$$se(\bar{x}) = \sqrt{\frac{(N-n)s^2}{nN}}$$

Sample Size Determination

- For large populations and small samples (most common case), (N-n)/N approaches 1 and we have

$$se(\bar{x}) = \frac{s}{\sqrt{n}}$$

- Thus, 4× sample size will only decrease error by half – i.e., **diminishing returns to scale**

- Standard way to determine sample size

$$n' = \frac{s^2}{se(\bar{x})^2}$$

- Then correct for finite sample size

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

Sample Size Determination

- We don't have either s^2 or $se(\bar{x})$ - cannot determine sample size without sample error but cannot determine sample error without a sample. **What to do??**
  - s^2 must be estimated from other sources
  - Must determine a confidence level (e.g., 95%, implying acceptance of 5% error)
  - Need to determine as interval around mean – useful option is expressing sample size as function of expected coefficient of variation ($CV = \frac{\sigma}{\mu}$)
  - For Normal distribution and 95% CI, $\mu \pm 1.96\sigma$
  - If we accept 10% error ($\mu \pm 0.1\mu$)

$$se(\bar{x}) = \frac{0.1\mu}{1.96} = 0.051\mu$$

$$n' = \left(\frac{s}{0.051\mu}\right)^2 = 384CV^2$$

  - We can often assume CV = 0.5 (i.e., middle point of standard deviation relative to mean) giving **96**

Errors in Modeling and Forecasting

- We assume the correct model specification is known *a priori* ("from the earlier", or from theory) and data used to estimate models have no errors
- These conditions are **always violated** in practice
- Even if true, forecasts will contain errors due to inaccuracies in assumed explanatory variables in design year

Errors in Modeling and Forecasting

- **Measurement error**: due to inaccuracies in base year data
  - Questions badly registered by the interviewee, answers badly interpreted by the interviewer, network measurement errors, coding and digitizing errors

  - We often use different units than travelers
    - Modeler works in seconds and meters while traveler finds it hard to measure at this granularity and uses minutes and miles

- **Self-selection bias**: attributes of chosen alternative perceived as better and those of unchosen alternatives worse than their true value
  - Reinforce rationality of choice
  - Ex. Driving time is 10 minutes and transit time is 25 minutes (including walking and waiting time) but a driver perceives transit time as 1 hour

Errors in Modeling and Forecasting

- **Sampling error**: Arise from using finite datasets
  - Sampling error **approx. inversely proportion to sample size** – to halve error you would need to quadruple sample size (expensive)

- **Computational errors**: Error due to iterative computational procedures
  - Generally **small relative to other errors**
  - Except for **traffic assignment to congested networks** and **supply/demand equilibrium** problems where they can be large

Specification Errors in Modeling and Forecasting

- Arise from **simplification** or **lack of understanding** of the process

  - **Inclusion** of irrelevant variable – e.g., number of computers in the household (actual variables are household size and income)

  - **Omission** of relevant variable (most common error) – e.g., travel time from a mode choice model

  - Not allowing **taste variation** – case in most practical models of choice

  - Use of **inappropriate model form** – linear functions to represent non-linear effects
    - Compensatory models to represent non-compensatory behavior
    - Omission of habit and inertia
  - Can address through specification change and more model complexity – total cost may not be easy to estimate and may induce other errors which are costly or impossible to eliminate

  - Some error is always present in models

Errors in Modeling and Forecasting

- **Transfer errors**: Model developed in one context (time and/or place) is applied in a different one
  - Spatial can be partially corrected
  - Time is harder as you look further into the future

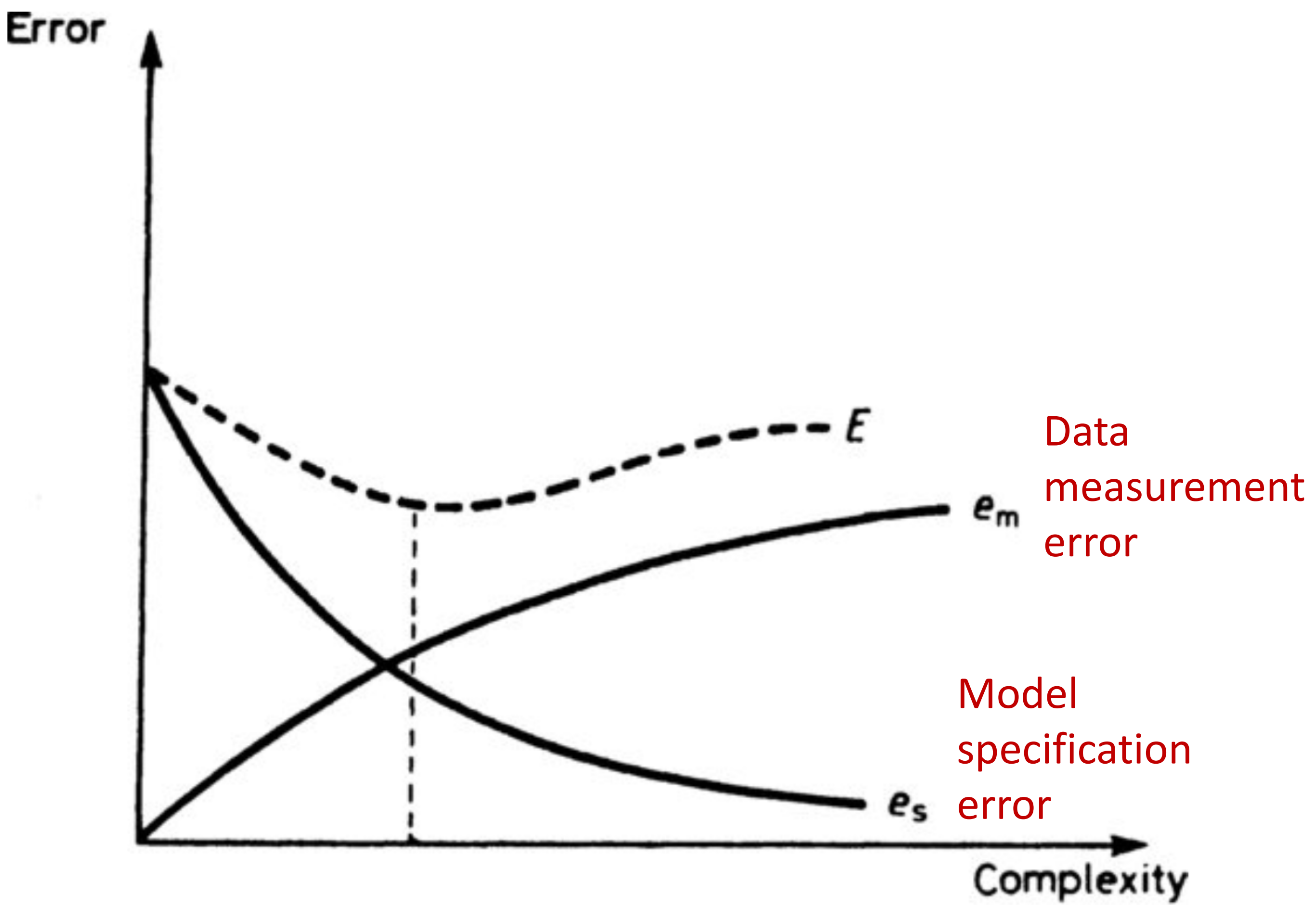
- **Aggregation error**: Arise from need to forecast for groups using individual data
  - **Data aggregation**: Aggregation of departure location and time - at best averages across individuals
  - **Aggregation of alternatives**: bus can mean regular bus, express bus, zone stop bus, BRT that are seldom treated as separate modes
  - **Model aggregation**: Need to aggregate model outputs to perform analysis - flows on roads need to be measured per hour, etc. rather than when an individual vehicle leaves home

Model Complexity & Data Accuracy Trade-off

Often we have a cost tradeoff between **model complexity** and **data accuracy** given a fixed budget

Analytic model methods recommend:

- Focus on variables with large errors
- Focus on the most relevant data
- Try to avoid correlated variables
- More variables means more accurate forecast but also more data measurement error



Variable Choice Issues for Forecasting

- What variables will maintain their effect in the future?
- Can be simple but uncertain – e.g., **future fuel prices**
- Can be complex – e.g., age, gender, income, employment type, marriage status, and number of children.
- **Question**: What would a model designed in 1955 say about gender effects in 2005?

Cross-Sectional Data Challenges

Most analysis relies on cross-sectional data

- Problem of habit and time lags
- Particular cross-sectional conditions will correspond to particular history of changes in certain variables influencing choice – inertia or path dependence
- Mode and location choice driven by gas price, life events, etc.
- Same current characteristics for two individuals but different choices because of how they arrived at that stage

Need for Stated Preference Data

- It is rare that we **observe** the decision-making process
- What if we want to know demand for a good or service that **does not exist** yet (at least locally)?
  - **Example**: Would you use a vertical takeoff and landing vehicle (flying car)?

Summary Comparison Revealed Preference (RP) and Stated Preference (SP) Observations			
Factor	Revealed Preference	Stated Preference	comment
form of choice behaviour observed	concerns actual 'compromise' choices with real-world constraints included	potentially concerns preferences rather than 'compromise' choices, as hypothetical context can be used to remove real-world constraints	relates to purpose of survey
establishing values for explanatory varilables	engineering values expensive to establish; stated values inexpensive but potentially distorted by faulty perceptions and expost justification	presented values inexpensive and unambiguous	advantage with SP
correlation structure in data for estimation	correlation structure uncontrolled; analyst must accept potentially high correlations among explanatory variables and deal with impacts of these correlations on estimation	correlation structure controlled; analyst can dictate correlations among explanatory variables and with low correlations can avoid impacts of correlations on estimation	why SP used
examination of causal-behavioural connections	indirect in that there is a reliance on correlations between observed behaviour and engineering values in particular	direct in that respondents are asked to react to indicated attribute values	not often an issue
flexibility	limited to real-world contexts	not limited to real-world contexts, but validity increasingly questionable as context becomes less familiar; ability to consider non-existing alternatives	why SP used
efficiency	more limited, but rankings of real-world options can be sought; difficult to return to field at a later date to the extent that additional engineering values are required	not so limited, with ability to seek rankings and numerical ratings rather than just single best-choice preferences; possible to return to field at a later date	not often an issue
transferrability	more limited, as real-world conditions and context are tightly woven into observed real-world behaviour in a given implementation	less limited, as hypothetical context can be specified to be identical across implementations	advantage with SP, more for research
speed of implementation	can be slow depending on availability of engineering values for explanatory variables	relatively fast; opportunity to collect multiple responses from same respondent	advantage with SP
validity	certain	can be questioned; important to handle with careful design	why RP used
certainty about respondent comprehension	certain to the extent that respondent made actual choice in real world situation	uncertain to the extent that respondent does not understand process or required response; but survey design can provide for confirmations	not often an issue

Stated Preference Experiment Components

- **Alternatives**: person makes choice between alternatives
- **Attributes**: alternatives are defined by their attributes
- **Attribute levels**: attributes are described by their levels.

Example – Mode Choice in Toronto

Question 41: Please choose the alternative that you would prefer to use **for your typical commuting trip**.

	Drive yourself	Driven by someone you know	Public Transit	Exclusive Ride Hailing	Shared Ride Hailing	Taxi	Bicycling	Walking
Travel Time (mins) ?	6	6	29	6	7	6	6	14
Travel Cost ($) ?	$0.22	$0.11	$ 3.10	$1.94	$1.30	$7.39	-	-
Waiting Time (mins) ?	-	-	7.5	2	5	2	-	-
Walking Time (mins) ?	-	-	5	-	-	-	-	-
Parking Cost ($) ?	$ 15	-	-	-	-	-	-	-
Other Passengers ?	-	-	-	-	1	-	-	-
Delay Time (mins) ?	-	-	1	-	4	-	-	-
Level of Crowding ?	-	-	Moderately crowded (50% chance of getting a seat)	-	-	-	-	-
Frequency of Delays over 5 mins ?	-	-	Once a month	-	-	-	-	-
Your Choice:	○	●	○	○	○	○	○	○