Preparing Model Data

Data in travel demand modeling serve as inputs for developing and updating models, which help to understand and describe existing transportation conditions. Models also measure and monitor the travel impacts of implemented policies, strategies, and investments. To serve these purposes, models need high-quality data from reliable sources. This chapter focuses on determining data needs, identifying reliable sources of data, and data collection practices including how to collect it, when to start collecting it, and steps for checking data for errors, reasonableness, and validity.

Determining Data Needs

Determining data needs early allows better coordination among the lead agency and stakeholders for preparing and examining the data in future steps. These data support model calculations that allow determination of travel in the studied region. Completing an inventory of data needed for the model helps identify the extent of available data and what needs to be obtained. When determining data needs, consider

- What data sets are needed?
    - Land Use
        - Population, employment, school and other zonal data like boundaries.
        - For ABMs, detailed population characteristics like number of children, age, race, sex and ethnicity to allow simulation of entire model region population.
    - Travel Behavior
        - Surveys such as household, on-board transit, visitor, and freight.
        - Origin-Destination (O-D) data from sources such as GPS, mobile apps, social media, connected vehicles, and payment systems.
    - Transportation Network
        - Traffic information.
        - Center lines and geometry information, roadway characteristics like lanes, facility type/functional class, aerial imagery.
        - Other network attributes like district, county, SIS designation, free flow speeds, and probe derived congested speeds for calibration.

        - Transit attributes like route, travel time, service characteristics and mode.

- What is the required resolution of the data?
    - Examples include whether there is a need for data at the individual, household, point, block, block group, link, segment corridor, district, county or other level.
- In what format is the data needed? Example formats may include:
    - Flat text, comma delimited tables, dbf files
    - GIS formats like shapefiles or geodatabases
    - Database formats that are accessed through relational database applications like MS SQL or Oracle
    - Matrix formats
    - Json
    - Proprietary formats (binary network files, tables)
    - Web formats like HTML, XML
- How much processing will be needed for the data to be useful?
    - Consider the level of QA/AC that will be required. Was data pre-vetted before receipt?
    - Will the data need to be combined with other data to enhance utility? Examples may include joining zone record data with employment information or adding specific attributes to networks like traffic volume information, and project ID.
- Once data needs are determined, are there any gaps in data availability?
    - For example, is comprehensive employment data for all regions for the year of interest available?
    - Is traffic data available for all districts and roadway segments of interest at the required level of detail?

Identifying Reliable Data Sources

Data are the most important component of travel demand models because they set the tone for the success of model development. Poor data quality will ripple through the process and result in model outputs of little to no use. Widely used data sets for

models are considered reliable because of the substantial vetting process used by stakeholders.

The inputs used should be from reputable sources such as the US Census Bureau, Federal Highway Administration, Bureau of Transportation Statistics, Florida DOT, local governments, and employment forecasters. Repeat vendors that have repeatedly been used without issue or are known industry standard leaders are also sources of reliable data.

These data are used in an extensive adjustment process to match observed behavior and tested against information such as traffic counts and probe data to ensure reliability. This process is described in more detail in the chapter on validation and calibration. When choosing data sources, consider:

- Where is the data sourced?
    - Consider organizations/vendors that provide the data and the characteristics to observe for each data source. For example, existing population and household data from the US Census Bureau.
    - Are the data publicly or privately available?
    - There are many public data sources available at no cost for use in model development, but the more detailed data is usually obtained from private data vendors. A list of major data sources are shown in Table 1.
- If sourced from a private vendor, what is the cost of the data?
    - The cost of many of these datasets are negotiated with the vendor and are not usually listed. These data often come with a high cost, so it is important to reach out to vendors early to determine cost to better identify any necessary data purchases early in the process.

Table 1: Summary of Data Types and Sources

Data Types	Data Description	Data Sources	
		Nationwide	Statewide/Regional
LAND USE			
Household Survey	Data on how, where, when, and why people travel, and on the socioeconomic and demographic characteristics of households and individuals	National Household Travel Survey (NHTS) American Community Survey (ACS)	FDOT and Local MPOs
Zonal	Population and household characteristics including number of households, number of people, auto ownership, persons per household	U.S. Census Bureau	FDOT, local MPOs and other jurisdictions
Employment	Employment by detailed industry categories or by aggregated generic categories like office, warehousing	Private Vendors	FDOT, local MPOs and other jurisdictions, ports, private vendors
TRANSPORTATION NETWORK			
Highway Network	Comprehensive set of roadway and other transportation networks for use in the model. Usually available in GIS format, can be easily integrated into modeling software tools.	National Highway Planning Network (NHPN)	FDOT GIS Library Local MPOs and other jurisdictions,
Traffic	Comprehensive, statistical traffic information including factors such as daily counts, vehicle classification, speeds,	FHWA's Office of Highway Policy Information	FDOT Open Data Hub FDOT Traffic Information

Data Types	Data Description	Data Sources	
		Nationwide	Statewide/Regional
	weight, directional factor, truck factor, and design hour factor.	National Performance Management Research Data Set (NPMRDS)	MPOs, Counties, Cities
Transit	Detailed ridership data, financial information, and operational statistics for transit agencies. Transit route timetables, stops, stop times, agency	National Transit Database (NTD)	FDOT Public Transit Office Local Transit Agencies via General Transit Feed Specification (GTFS) and other data
Freight Survey	Data collected from industries that ship commodities	Commodity Flow Survey (CFS) Freight Analysis Framework (FAF 5.0) – uses CFS	FDOT's Freight & Rail Office (FRO), Private providers
TRAVEL BEHAVIOR			
Origin-Destination Data	Data documenting observed trip origins and trip destinations.	Census Transportation Planning Products (CTPP) National Household Travel Survey (NHTS) Private vendors	Local and Regional Household Travel Surveys Private vendors
On-board Transit Survey	Data collected directly from public transportation passengers in regions with significant transit ridership to facilitate planning for future transit investment	National Transit Database (NTD)	Local and Regional Transit Agencies

Data Types	Data Description	Data Sources	
		Nationwide	Statewide/Regional
Visitor Survey	Data capturing the travel patterns and behaviors of tourists	National Travel and Tourism Office (NTTO)	Visit Florida, Local Districts and MPOs

Land Use Data

Land use data consist of key items like population, employment, zonal data, and any other data that concerns the people in the region, where they live, type of dwelling, where they work, and where they play. Specific examples include population, household characteristics, school enrollment, and employment for the model region. The model uses this information to assess how many people are traveling and to where they are going. The most common sources of these data are:

- Population and household data: Pulled from public sources such as the U.S. Census Bureau's decennial Census or the ACS, depending on the base year of the model.
- Employment data: Pulled from the U.S. Census Bureau, such as the Longitudinal Employer-Household Dynamics (LEHD) data or County Business Patterns (CBP) data. However, the data from these sources are too highly aggregated to provide meaningful detail needed for the model. To account for this, FDOT purchases employment data from private vendors (such as Dun & Bradstreet or Data Axle) approximately every five years for the state of Florida. This data provides employer-level information across industries that can be aggregated to any model's zonal structure.

This information is assigned to the model's traffic analysis zone (TAZ) structure. Model TAZs are created by the model developer. It is recommended that the zones nest within Census geography (such as tracts or block groups), though this may not be applicable for all models based on data availability and other practical considerations. The TAZ structure is unique to each model and like other parts of model development, exceptions may need to be made where appropriate.

The zone system is made up of two types of zones: internal and external. Internal zones are within the regional boundaries of the model and use land use data. External zones are outside of those boundaries and use traffic data. External zones are commonly referred to as stations because they consist of points along or just outside of the model boundary along roadways that cross the boundary. They inform the model of how many vehicles are traveling in and out of the modeled region and are represented as traffic volumes along the described points. These volumes are ideally pulled from traffic data for the region by the boundary, or in the absence of such data, from neighboring region traffic data.

Transportation Network Data

The most comprehensive nationwide database of the major highway system in the U.S. is the National Highway Planning Network maintained by the US Department of Transportation. This database contains urban and rural arterials and all NHS routes. In addition, FDOT maintains a database of Florida-specific transportation networks.

NHPN data is available in GIS format, and can be easily integrated into modeling software tools. NHPN focuses mainly on road networks and does not typically include transit networks.

FDOT data is also available in a GIS format and provides roadway, transit, and non-motorized mode networks. When using these data, consider:

- How extensive and detailed is the geographic coverage?
    - For example, is a newly built roadway segment represented in recently procured datasets?
    - Is the network represented down to the level of local streets, if necessary for transit or other localized factors, or is it higher level such as collectors and above?
- Does the data have all required attributes for modeling needs?
    - Determine the necessary area type / facility type / context class attribute information to determine link capacities and if not available, assess the options.
- Does the data require a special license to access? If a license is needed, how is the license procured and are there access and usage restrictions?
    - For example, freight/truck O-D data sources from private vendors have usage restrictions.

Traffic data

Traffic data are used in the model as the ground truth for determining accuracy and reasonableness in the final model outputs. They are provided at the daily level but may also be available for specific periods.

FDOT is a key source for traffic data. Some data, like average annual daily traffic (AADT), can be accessed via GIS-based platforms that allows searches by county, road segment, or specific locations.. All FDOT traffic data are available at no cost.

Many MPOs across Florida also collect and publish localized traffic count data that can be accessed via MPOs websites. Traffic counts by MPOs are more granular than state-level sources and may include time of day count information, which is critical for Preliminary Design and Engineering (PD&E) type analyses. All traffic data provided by MPOs in Florida are available at no cost via request from the individual MPO.

National probe data from the NPMRDS provides speed data in 5 minute, 15 minute, or 1 hour increments along all National Highway System (NHS) roadways. The data are available by Traffic Message Channel (TMC) segments, of which there are over 400,000 throughout the country.

Transit ridership and operational data

Transit ridership serves as ground truth for validation comparisons especially for aggregate comparison metrics like passenger miles traveled and unlinked passenger trips. Fare information is also a key input for mode choice and transit assignment routines.

The National Transit Database (NTD), managed by the Federal Transit Administration (FTA), is a comprehensive source for public transportation across the U.S. Since all fixed-route systems operating in Florida are reported to NTD, transit-related data for Florida and Florida agencies can be filtered for use in the modeling process. NTD data is available at no cost. The FDOT Public Transit Office also publishes ridership reports and data that are compiled from local transit agencies in the state.

Transit network data

Transit network data may be publicly available in the General Transit Feeds Specification (GTFS) dataset. Information on routes, station, service patterns and headways are available in a standardized format that allows for efficient incorporation into model transit network representation

Travel Behavior Data

Behavior data describe the patterns of trip making within the population. Historically, the most common data type is a household travel survey, which is a subset of a broader household survey category. Household surveys collect data on the socioeconomic and demographic characteristics of households and individuals. Household travel surveys go a step further and also include information on how, where, when, and why people travel. These surveys aim to document and characterize trips made by each household, including origin and destination, trip purpose, start and end dates and times, mode of transport used, and trip expenditure.

- **National Household Travel Survey (NHTS)** is currently the largest and most comprehensive survey of household travel in the U.S., conducted by the Federal Highway Administration (FHWA) every 2 years. It collects national data by
    - Trip level e.g., trip purpose, mode of travel, trip distance, trip duration, time of day, and accompanying persons;
    - Household level e.g., demographics, vehicle ownership, income, and geographic location; and
    - Individual level e.g., age, gender, employment status, driving status, and personal travel behaviors.

  NHTS is primarily used to calibrate the non-work related trip activity in models such as social / recreation trips, home based other trips, shopping trips, and school trips. For trip-based models, it is used to derive items like cross-classification rates in Trip Generation, average trip length used to calibrate Trip Distribution, and Auto Occupancy and travel mode in Mode Choice. It is often supplemented by additional local survey data where available, especially in the development of Activity Based Models.
- **American Community Survey (ACS)** ACS, part of the U.S. Census Bureau, is an ongoing survey that provides vital information on a yearly basis about our nation and its people including information on jobs and occupations, educational attainment, veterans, work trips taken, whether people own or rent their homes, how many live in a home, and many other topics. ACS data can be accessed for free via census.gov and through other platforms like IPUMS USA. Unlike the NHTS data, the ACS is limited to commuting trips and provides no details on non-work-related travel. Because ACS is ongoing, it provides a valuable snapshot of population trends in intermediate years between the decennial Census tabulations, and may also be considered a form of land use data.
- Many states conduct **statewide and regional household travel** surveys that are tailored to the needs of their states or regions. FDOT has conducted Florida Household Travel Surveys in the past to collect data on travel behavior across the state of Florida. FDOT has also obtained Florida-specific travel behavior data contained in add-on NHTS tabulations for past efforts. Regional household travel surveys in Florida are conducted by MPOs that publish reports and survey results.
    - For example, in 2017, Miami-Dade TPO published the Southeast Florida Household Travel Survey. This survey covered the tri-county region of Palm Beach, Broward, and Miami-Dade. The data obtained during the survey was key data for the regional planning model.
    - Has your region conducted a household travel survey? Consider data from these surveys. The frequency of regional surveys in Florida varies by region, with some being conducted every 5 to 10 years.

- **Specialized household travel surveys** are conducted by and can be obtained from research institutions, universities, and local government or transit agencies. These surveys are narrow and targeted at specific populations or events such as the aging population or telecommuters. Many of these datasets are one-time studies, while others have a wide variance in the interval of data collection. Access to these datasets are typically restricted, require specific permissions, or have associated costs. Specialized surveys conducted by universities or research institutions in Florida might be published in academic journals or available through the institution's website.

Specialized surveys are valuable for insights not captured in regularly conducted surveys like the NHTS. They are useful in the rapidly changing topics like e-commerce and telecommuting which have broad implications for future transportation demand. Their primary limitations are cost and frequency.

Each survey type uses methodologies to balance data accuracy, respondent burden, and the need for comprehensive travel behavior insights. For example, the 2017 NHTS datasets used a combined telephone and online survey mode, but the 2022 NHTS was conducted predominantly online, with a mail version offered for those who requested it.

Also, some surveys may, in lieu of traditional surveys like travel diaries, opt for multi-day smartphone Global Positioning System (GPS) surveys for data collection, depending largely on the agency objectives, budget available, and survey contractor offerings. Although various methods may offer improved opportunities, they can also introduce distinct biases and limitations that should be carefully considered when using the data in modeling.

Non-Household surveys

Unlike household travel surveys which provide insights into the daily travel patterns of residents, non-household travel surveys focus on other important segments, such as visitors, commuters, and freight.

- **On-board transit surveys** are collected directly from public transportation passengers in regions with significant transit ridership to facilitate planning for future transit investment in the region. National-level on-board transit survey data can be accessed through transit agencies or federal repositories like the NTD. In Florida, transit agencies like the Miami-Dade Transit, Jacksonville Transportation Authority, and the Central Florida Regional Transportation Authority (LYNX), conduct these surveys. Survey results are summarized in reports and published on the agencies' websites, while raw data may be provided upon formal request.

- **Visitor surveys** capture distinct travel patterns and behaviors of tourists, which differ significantly from those of local residents. In areas with high visitor volumes, such as Florida, it becomes essential to extend traditional travel demand models by incorporating specific visitor demand segments to improve the accuracy of travel forecasts. Florida's tourism board, Visit Florida, regularly conducts visitor surveys at the statewide level. Additionally, many regional tourism boards publish summarized visitor survey reports on their websites. However, detailed data or full reports from Visit Florida and certain municipalities, such as Visit Orlando, require paid membership for access.
  Information from visitor surveys have been used in Florida models, including SERPM, CFRPM and TBRPM. However, the data are limited and the scope of these surveys are not as extensive as desired to provide a full picture of visitor travel patterns. As with specialized household surveys, special visitor surveys are expensive and infrequent.

- **Freight surveys** collect data from industries that ship commodities. Freight survey data is collected by the Bureau of Transportation Statistics (BTS) and may be part of larger studies like the Commodity Flow Survey (CFS) — the only publicly available national source of data for highway freight. In Florida, FDOT's Freight & Rail Office (FRO) conducts and oversees freight surveys, as well as other freight-related data collection. Summary reports and findings from these surveys are published by the FRO and can be accessed through the FDOT website. Also, the FRO updates of the Freight Mobility and Trade Plan (FMTP), which includes data from various freight surveys. The primary freight survey product used in Florida is the Freight Analysis Framework (latest version is FAF5). It is derived primarily from the CFS. The data provides freight trip from Origin to Destination incorporating all intermediate unlinked modes and segments. Assigned flows are available in GIS as well as tabular O-D format. It is used to develop freight and truck models at various geographic scales.

Origin-Destination Data

This data is important for model calibration and validation. It is used to assess whether the model is replicating real-world travel patterns as reasonably as possible. While many sources offer this information at a cost, the only publicly available source of journey-to-work O-D data is the Census Transportation Planning Products (CTPP), produced by the AASHTO Census Transportation Solutions (ACTS) Program. American Community Survey (ACS) data are used to derive the CTPP tabulations using annualized data over a 5-year period that are statistically aggregated to improve explanatory power.

In addition, CTPP provides additional cross tabulations specifically for transportation planning and modeling that are unavailable via the decennial Census or the ACS. CTPP data is used in calibration and validation of travel characteristics such as how many persons travel alone and average travel time to work. The CTPP only provides O-D data for trips to and from the workplace. Other sources, such as household travel surveys would be needed to capture all other trip types. FDOT and all regional models use CTPP data for modeling purposes.

Collecting Data for Travel Demand Modeling

Data collection efforts can be costly in both time and monetary expense. The data inventory completed at the beginning of model development will help the model development team decide which data to use and how much to spend on collection or purchase.

If a significant data collection effort cannot be funded, it may be possible to borrow data or models from other sources, such as an area with similar sociodemographic and geographic characteristics, preferably in a region that has recently conducted travel behavior surveys. An assessment of similarities among different regions and urban areas for borrowing model parameters should include the following considerations:

- What are similarities in the lifestyle and employment characteristics of the regions?
    - Population concentrations of retirees, students, or tourists.
    - Employment activities such as warehousing, industries, government, or tourist attractions.
- Are the physical geography and development patterns similar?
    - For example, do the regions both contain elongated coastal development like Southeast Florida or do large internal water bodies or bays define the regions.
- What is the availability of transit technologies/services for each region?
    - For example, is transit service in both regions primarily all-day local bus service or do they both have fixed guideway transit systems.
- What are the congestion patterns between regions?
    - For example, are there similar peak spreading patterns, commercial vehicle trip patterns, or tourist trip patterns?

Model developers should reference the data inventory previously described and use the list to systematically gather all necessary data from relevant sources.

New Data Collection Approaches

In recent years, the proliferation of probes, GPS, online survey resources, mobile apps, among others has led to an increase in available data that may be used for transportation planning. Among the data types available from these tools are speed, count and O-D data. The sample sizes available are large when compared to similar data from surveys and the cost per record is commensurately less. However, while very useful for planning, the data collected via this means lacks the contextual information that is found in traditional surveys such as trip purpose, trip maker income. To overcome these limitations the following approach is typically used:

- Combining population, employment, and other characteristics in survey or census data.
    - Several data fusion techniques may be applied including:
        - Statistical matching
        - Bayesian data fusion
        - Machine learning methods
    - These data fusion techniques are detailed in transportation and statistics journals.

Collection Protocol for Developing a New Travel Demand Model

When developing a new travel demand model, data collection should begin well in advance of the model's creation. Ideally, the process should start at least 12 to 18 months before the model is expected to be finalized. This allows ample time to design surveys, conduct data collection, process and clean the data, and integrate it into the model. The timeline for preparing specific data are as follows:

- Land Use Data :
  Since these datasets are more readily available from national or state transportation agencies, the timeline for collecting these datasets might be short, typically 1 to 3 months, depending on the need for any updates or adjustments to the data. It is important to simultaneously source the transportation network data with the zonal data to define TAZs.

- Transportation Network Data:
  Count data from sources like FDOT and local agencies are usually readily available. Sometimes there is a need to collect supplementary count data to enhance the ability to perform validation comparisons. In such instances, more time will be needed, up to 4 – 6 months, depending on level of data cleanup required.

- Behavioral Data:
    - Household surveys: The primary type of behavioral data. Designing and conducting a household travel survey is one of the most time-consuming aspects of data collection for a new model. Survey development can take between 3 and 6 months, including designing the questionnaire, selecting a survey firm, pilot testing, and obtaining necessary approvals. Survey fielding and data collection may span 3 to 6 months depending on the intended sample size, data collection method, and response rates. Data cleaning, processing, and validation can take 2 to 4 months. The entire process from survey design to data readiness may take up to a year or more.
    - Specialized Surveys: If specialized surveys (such as freight or visitor surveys) are required, sourcing and integrating these datasets should start at least 12 months prior to model development. Negotiating data access agreements with vendors or research institutions may add time to the process.
    - Other Data: The timeline for other private sourced data is typically shorter than for specialized or household surveys because they are usually purchased as a package of data that that have already been refined. In some cases, the raw data may be manipulated for custom tabulations and outputs so the timeline may vary depending on the need.

Collection Protocol for Updating an Existing Model

When updating an existing travel demand model, the data collection timeline can be more flexible, focusing on gathering new data that reflects recent changes or filling gaps in the existing model. If a new household travel survey is needed, the timeline will mirror that of a new model, potentially requiring a year or more. However, in many cases, existing survey data (such as the NHTS) can be updated or supplemented with smaller, targeted surveys, reducing the timeline. Other data sources take less time, typically 1 to 3 months, since updates would have been made to the data by the data-providing agency. Overall, the process of data collection for an existing model takes about 6 to 9 months and is heavily dependent on model complexity.

Checking and Correcting Data (QA/QC)

Data integrity checks involve all data types that modelers work with and involve ensuring that the datasets are complete and do not contain obvious flaws. Common examples of flaws may include:

- Incomplete records such as zones with population values but no households;
- Missing records where values for a known geography that should have values are absent;
- Known outlier values that are outside of reasonable ranges; and
- Values that are the wrong format like numeric data in text attributes and vice-versa.

Other issues may include

- Wrong data, such as information from a different area;
- Incorrect mapping projections that cause alignment problems; and
- Duplicate ID values or 'keys' for attributes that need to be unique.

The approach to resolving these data quality problems will depend on the data type in question as described in the following sections.

Land Use and Socioeconomic Data

Validation checks: The primary aggregate validation checks for socioeconomic data involve summing Traffic Analysis Zone (TAZ) data by geographic districts of interest and comparing to observed control totals for the same geographies. These comparisons may include number of persons, number of workers, number of households, school enrollment, income level and/or auto ownership, and employment. They are useful in comparing against historical data from local, and national sources such as Census and the NHTS. This enables the estimation of parameters like mean vehicle ownership, median incomes, and household size distributions. In addition, TAZ data can be sorted to identify outliers using robust statistical methods.

Reasonableness checks: GIS plots are one method of checking for reasonableness. Several relationships among different variables, such as persons per household, jobs per person, jobs per worker, autos per household, and employment and population densities can be sorted or mapped using GIS. Thematic mapping is a useful tool to easily depict densities and other trends. Base and forecast year data can be compared visually for reasonable growth rates using GIS.

Sensitivity checks: Sensitivity checks for socioeconomic data are mostly carried out after the model has been developed. This task adds or subtracts activities (e.g., households, retail employment) in a TAZ and evaluates its travel impact (e.g., traffic volume). It is

reasonable to expect increased activity to correlate with higher travel volumes, and vice versa, especially near the adjusted TAZ. Although it may not be feasible to apply sensitivity checks to every TAZ, a representative sample covering various development and area types can be selected for these checks.

Transportation Network Data

Geospatial accuracy: Like land use data, and transportation network data, consistency checks can be performed using summaries, sorting, and visual displays for comparison with independently summarized data for the same strata. For example, highway link speeds can be sorted and/or summarized by facility type, speed limit, and area type, and compared to similar summaries from available GIS data. Highway network maps include color-coding or posting of attributes such as area types, facility types, lanes, screen lines, traffic counts, and speeds. Color coding allows quick identification of spuriously coded network values. Transit network maps can be used to display transit routes by mode (e.g., local bus, express bus, and fixed guideway), headway, operating periods, stations, and access connectors.

Connectivity checks: Connectivity checks and checking that one-way links are coded correctly and in the correct direction should be performed. Minimum travel-time paths (e.g., time, distance, cost) can be built "on the fly" using the modeling software to check for connectivity, directionality, logic, and consistency. The following is a summary of guidelines to consider when coding centroid connectors:

- Centroid connectors should represent realistic roadway and transit access;
- Centroid connectors should not cross man-made or natural barriers, such as lakes, swamps, rivers, railroad tracks, and limited access highways;
- Sufficient centroid connectors should be included to avoid loading too many trips onto one roadway network link;
- Centroid connectors should not be connected at intersections or directly to interstate ramps; and
- When two centroid connectors are connected to the same roadway segment, the access points should be separated by a certain distance and the logical direction of trips from each should be away from each other.

Figure 1 provides three examples of the proper way to code centroid connectors in relation to physical geography, street patterns, and subdivision access. TAZs are depicted with different fill colors, centroids are represented by a dot, and centroid connectors are depicted as dashed lines with arrowheads.

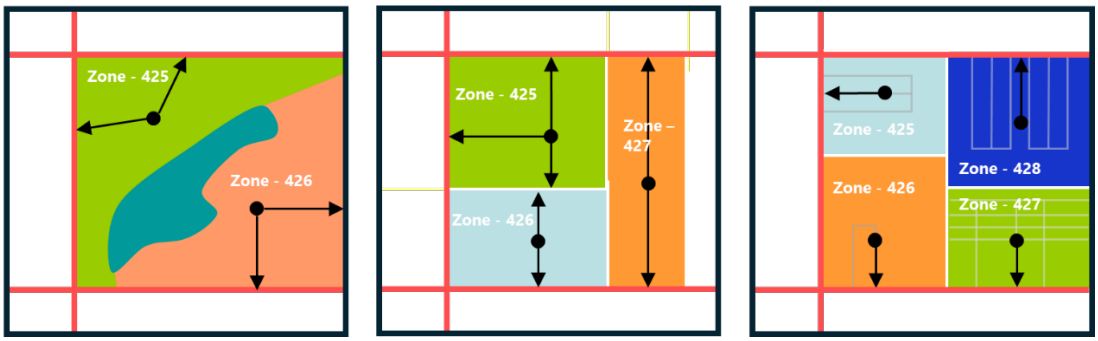

Figure 1 Centroid and Centroid Connector Coding Example

Traffic and Transit

Reasonableness Checks: Traffic and transit volume consistency can be checked by comparing volumes across different times and days to ensure they follow expected patterns. If anomalies are found, the causes of the anomalies (e.g., data collection or equipment errors, temporary disruptions) should be investigated. Inconsistencies and outlier points that cannot be validated should be removed.

Validation and verification checks: Current traffic and transit data should also be cross-referenced against historical records to identify trends or deviations. If deviations are unjustified, compare data with data from other sources and adjust if necessary. In some cases, field verifications may be necessary to validate reported traffic volumes, speeds, and ridership figures.

Accuracy Checks: It is also important to ensure temporal and spatial matching between the traffic/transit data and the corresponding network and zonal data.

Travel Behavior Data

Surveys

Missing data: This occurs when gaps or blank cells appear in the dataset. This can result from respondents' oversight or intentional decision not to answer, equipment failures, or lost files. Identifying the cause of missing data is needed for determining the appropriate correction method. When data are missing at random, the absence can be accounted for by variables with complete information, and several imputation methods can be used to fill in the missing data. However, when missing data is intentional (e.g., a respondent's decision to not complete question of a survey), imputations would be inappropriate. In such cases, the best course of action might be to delete the entire data point for those responses.

Representativeness and statistical inference: For data to be suitable for statistical inference, the sample must reasonably represent the target population. The demographic distribution of survey responses should be compared with known demographic data of the region, such as Census information, to ensure consistency in population characteristics like age distribution, income levels, and vehicle ownership. Because of discrepancies in group response rates, it is not uncommon for some groups to be over-represented or underrepresented. Adjustments can be made by expanding group sizes using weighting factors, to ensure that the survey results accurately represent the target population.

Identifying and handling outliers: Most survey firms use several data checking algorithms to identify and remove unreasonable and problematic responses, such as inattentive respondents, duplicates, responses from speeders. However, the algorithms may not identify all problematic responses, and checking for outliers may ensure better data quality. Box plots, scatterplots, and histograms can reveal the distribution of a single variable or relationship between two variables. Identifying outliers through only visualization may be subjective. Methods, like the interquartile range method or the Z-score method can be used to define the boundary of an outlier or measure the standard deviation of a data point from the mean. Observations that deviated from the general distribution may be outliers and should be removed or re-assessed

Validation and verification checks: Statistical methods, such as calculating mean trip lengths or travel times can be applied to the survey data and compared with historical data or expected ranges. If significant deviations occur, adjustments or exclusions can be made depending on the reason for the deviation (e.g., survey method, outliers). In 2022 NHTS data, for example, validation checks included confirming the number of household vehicles in the Household file matched the number of vehicle records in the Vehicle file.

Respondents with outliers may be recontacted to verify the accuracy of the initial responses. Further verifications may be necessary before exclusion. Outliers may be cross-checked with other variables in the survey. For example, a long travel distance might be legitimate if the mode of transport is air travel, which should be reflected in other variables. Outliers may also be compared with external data sources such as historical travel patterns. It is possible that some outliers accurately reflect the characteristic of a segment of the population (e.g., the excessively high shopping behavior of shopaholics). If, after exhausting all steps in the validation, verification, and recontact process, the necessary corrections could not be made due to insufficient information, the records can be removed.

There are several publicly available resources that provide detailed information on the methodology, data cleaning, and handling of missing data in travel surveys.

- Statistical Policy and Research BTS Guide to Good Statistical Practice 4. Processing Data", which can be accessed on BTS website at https://www.bts.gov/archive/publications/bts_guide_to_statistical_policy/chapter4
- NHTS User Guide provides documentation for each NHTS data-collection cycle. Microsoft Word - 2022 NextGen NHTS User's Guide V2_PubUse.docx (ornl.gov)

Next Steps

Whether developing a new model or updating an existing one, the process of gathering data for land use, transportation networks and behavioral inputs requires careful planning, appropriate timing, and robust methodologies. By following the practices outlined in this chapter, the groundwork is laid for effective model development and refinement.