CALIFORNIA High-Speed Rail Authority

SIGNATURE/APPROVAL SHEET

TO: Bill Casey

FROM: Justin Cassarino

SUBJECT: Approval of the California Ridership Model Documentation

DESCRIPTION OF ENCLOSED DOCUMENT(S): California Ridership Model Documentation

REVIEWER	REVIEWER'S INTIALS/DATE:	COMMENTS
Signer #1 Name (Print): Bill Casey	Signature on file	
Signer #2 Name (Print): Brian Annis	Signature on file	
Signer #3 Name (Print): Thomas Meinl	Signature on file	
Reviewer #1 Name (Print): Bruce Armistead	Signature on file	
Reviewer #2 Name (Print): Christian Hicke	Signature on file	
Reviewer #3 Name (Print): Justin Cassarino	Signature on file	
Author #1 Name (Print): Tanu Bansal	Signature on file	
Author #2 Name (Print): Masroor Hasan	Signature on file	

Report
November 2023

California Rail Ridership Model Documentation

Prepared by:

Steer
800 Wilshire Blvd, Suite 1320,
Los Angeles, CA 90017
USA

+1 (213) 425 0990
www.steergroup.com

Prepared for:

DB E.C.O. North America Inc. on behalf of
California High-Speed Rail Authority

Client ref: Client ref
Our ref: 23454511

Contents

Figures

Tables

Disclaimer

Steer has prepared this material using professional practices and procedures using information available to it at the time, and as such, any new information could alter the validity of the advice, results and conclusions presented.

It has been necessary to base much of this material on information from data collected by third parties. Steer does not guarantee the accuracy or reasonableness of third-party information/data.

The views and estimates contained within this document are influenced by external circumstances that can change quickly and can affect demand for travel and use of any proposed rail services and associated revenues. In particular, the demand for travel may differ from that assumed on the basis of third-party projections contained in this material. Furthermore, the criteria by which customers choose between competing destinations and modes of transportation may change over time and differ from the assumptions that underpin any projections contained in this material. The outcome of such events, circumstances and responses could result in material differences between forecast and actual results.

Steer has prepared this material for DB E.C.O. North America Inc. on behalf of California High-Speed Rail Authority on a for-information-only basis. This material may only be used within the context and scope for which Steer has prepared it and, unless otherwise agreed in writing by Steer, may not be relied upon in part or whole by any party. Any person choosing to rely on any part of this material without the express and written permission of Steer shall be deemed to confirm their agreement to indemnify Steer for all loss or damage resulting therefrom.

Introduction

Steer was commissioned in October 2019 by DB E.C.O. North America Inc. (DB ECO) on behalf of California High-Speed Rail Authority (CHSRA) to support in the development of a rail ridership and revenue modeling framework (the "model," or the California Rail Ridership model ("CRRM")) for California.[1]

Nature of the model and associated outputs

The model has been developed on the basis of our understanding of the key requirements of the model, including items related to the scope of the model, the scope of our agreed work, the required outputs from the model, and how the model will need to be used in the future. Steer's understanding of these key requirements is outlined below:

Scope of the model

- The model was required to cover the entire state of California as well as external travel links to reflect travel to/from neighboring states.
- In terms of the California High-Speed Rail (HSR) System, the model zones were required to cover the entire state, but the primary validation effort in relation to the HSR system was limited to the Phase 1 coverage area.

Outputs from the model

- The developed forecasts are comprehensive in scope but approximate in nature. The modeling output provides estimates and numbers that are appropriate for the study level and according to our agreed scope.[2] Accordingly, all outputs are provided on a non-reliance basis.
- Outputs from the model may be used to assist in:
  - Understanding the existing in-scope demand;
  - Understanding the impacts of future expected growth, including economic and demographic trends;
  - Understanding the willingness to pay and modal preferences of different populations/markets;
  - Understanding interactions with competing modes and where rail and HSR ridership come from;
  - Understanding issues related to access/egress from the core intercity rail network, including station choice;
  - Understanding the benefits to and impacts on different populations/markets;
  - Understanding of the impacts of future developments on origin-destination (OD) patterns and mode choice; and
  - Understanding of the impacts on the wider transportation network (for example, vehicle miles removed from highways).

[1] Where reference is made to the model throughout this document, this is not intended to refer to a single tool but rather to the full suite of tools and modules – including any post-processing diagnostics and presentation of outputs – that constitute the range of analytical outputs developed by Steer as part of this work.

[2] For further details, refer to our scope of work, dated October 14, 2019.

- Alongside core outputs feeding the items outlined above, Steer has developed sensitivity analysis to help explain key risks and quantify their potential impact.

Future use of the model

- Each of the three core parties involved in this work will need to use the model in the future for different purposes:
  - **California High-Speed Rail Authority:** The model will be expected to form part of the inputs to ongoing business planning processes, environmental work, station area planning, funding plans, and, more widely, the external justification for investment.
  - **Caltrans / CalSTA:** The model will aid in understanding ridership on both HSR and complementary rail, transit and regional bus services as part of statewide transportation planning efforts, including the future development of the California State Rail Plan (CSRP).
  - **DB ECO:** DB ECO is under contract to CHSRA as the Early Train Operator (ETO). This model may feed into various analyses of DB ECO, potentially including an understanding of ridership and revenue for a range of alternative service assumptions. Elements of the modeling may also feed into wider workstreams, including operating costs and financial analysis.
- In addition, the model may need to be used in conjunction with MPOs and their own local modeling tools, for example, when seeking to understand the regional impacts of proposed rail projects in their respective jurisdictions.
- The model, therefore, seeks to provide an appropriate balance between a number of competing items, including:
  - Local/service-specific nuance vs appropriate statewide behaviors; and
  - Detailed analysis vs efficient future use.
- Fundamentally, given the importance of the model for such a wide range of different future uses, the model has been developed to be easily useable. As such, overall priority in the development of the model has been given to the efficiency of future application of the model and the robustness of the behavioral responses implied within the model.

The set of forecasts developed and to be developed in the future using this model includes forecasts for different phases of the HSR system as well as forecasts related to the potential statewide rail service changes and enhancements outlined in the CSRP. Prior forecasts have been developed – either by Steer or by other parties – for each of these services. The forecasts developed for these services as part of this work are different from those produced as part of prior work. Differences arise due to a number of factors, including, but not limited to:

- The use of a different model framework (for example, the varying level of details ranging from sketch-planning level to a highly detailed level);
- The use of different input data feeding into the model framework; and
- The adoption of different data sets/sources, assumptions, and parameter values within the model framework.

In some cases, the forecasts produced as part of this work may be materially different from those produced as part of prior work. It is possible that such differences may have led to different decisions being taken by DB ECO on behalf of CHSRA, CHSRA or other parties, in part arising from the prior forecasts. Such differences do not negate or invalidate the forecasts and associated

advice provided by Steer[3] as part of any prior work, which was developed on the basis of the terms, methodology, approach and limitations agreed upon for each prior engagement[4], including:

- Agreed budget and scope limitations (for example, types of models/methods used);
- The information available to Steer at the time;
- Reasonable application of data sets/sources, assumptions and parameter values from that information; and
- Appropriate discussion with DB ECO or other parties as applicable as to the nature and risks associated with the forecasts produced.

Steer has discussed with DB ECO, CHSRA and Caltrans these differences and why the forecasts produced as part of this work differ from those produced as part of prior work. Such discussion, however, is not included as part of this documentation.

This document

The remainder of this document is structured as follows:

- First, provide an overview of the **modeling framework,** including its structure;
- Next, we set out the inherent limitations associated with the development of travel demand models, including specific items to be aware of in relation to this model;
- The model network **supply and assignment** process is then set out;
- 2018 **Base Year demand** is then described, including the data used and its processing;
- This is followed by setting out the data sources and derivation of the **population and employment** data used;
- The **choice modeling** approach, including the use of behavioral surveys, is described;
- Generation and distribution, the final component of the modeling framework, is covered; and
- The model validation demonstrates the robustness of the model in replicating 2018 base year conditions.

[3] Steer does not guarantee the accuracy or reasonableness of any third-party information, including, but not limited to, the validity of prior work undertaken by any other party.

[4] For the specific terms that governed the development of each prior forecast, please refer to the relevant scope of work and terms and conditions agreed in each case.

1 Modeling Framework Overview

This section provides an overview of the model. Further details on each step of the model can be found in subsequent sections.

When developing the model, we reviewed experiences and recommendations from different modeling efforts around the world. We have mainly focused on our internal experience of many of these modeling efforts and on the following sources of information:

- Documents and reviews related to the model that Cambridge Systematics developed for California HSR in 2008 and that has been periodically updated since then;
- Documents related to other existing models in use throughout California[5];
- Guidance on intercity passenger rail ridership and revenue forecasting, developed by Steer for the USDOT[6];
- Research papers on commonly used transportation model structures; and
- The UK's Transport Analysis Guidance (TAG) on various technical areas of model development and the specific recommendations adopted when developing the modeling framework for High Speed 2 (the proposed High-Speed Rail (HSR) system in the UK, linking London, Birmingham, Manchester, and potentially other cities).

Overview of approach

Our approach for the model has been developed based on our understanding of the key requirements, including items related to the scope of work, the required outputs, and how the model will need to be used in the future.

In addition to the items highlighted in the Introduction, we have developed our model framework with the following principles in mind:

- We utilized existing information and models wherever appropriate. In particular, we focused on reviewing parameters, assumptions and data from the California Statewide Travel Demand Model (CSTDM), one of the key strategic models that produce statewide origin-destination trip matrices for assignment to the statewide network, and from the various MPO models.
- MPO models provide inputs to our model but are not run as part of the model. Experience elsewhere suggests that attempting a closer integration between the intercity and MPO models would be time-consuming and of little benefit. This principle is consistent with the primary focus of the model (on intercity travel) and with facilitating future use of the model.

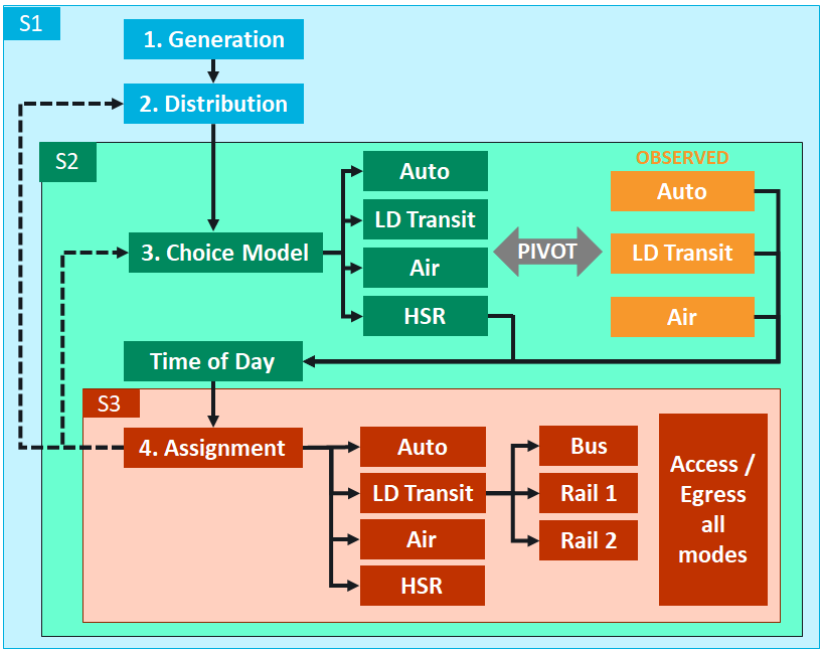
[5] For further details, see the "Existing models and public data" deliverable.

[6] https://www.oig.dot.gov/sites/default/files/files/OIG-HSR-Best-Practice-Ridership-and-Revenue-Report.pdf.

- The model includes autos, High-Speed Rail (HSR), other long-distance transit modes (rail and bus), and air. In practice, all modes contain common network elements, with road-based transit operating on the highway network.
- We have not directly modeled the competitive response of airlines to the introduction of HSR service, or any other rail service, or the future land use changes resulting from rail-related development. Instead, these two impacts are treated as external assumptions and used as inputs in our model. The base assumptions keep both the air supply and land use constant between scenarios, if required, specific sensitivity tests can be conducted should impacts for changes to these base assumptions need to be evaluated.
- The model includes five time-of-day periods, with four covering average weekdays and one covering an average weekend day.
- The model is calibrated using a base year of 2018,[7] with future year forecasts produced for three future years: 2030, 2040 and 2050.

The overall Modeling structure is shown in the following figure.

Figure 1-1: Modeling structure



Where:

1. **Generation:** In this first step of the model, we estimated the number of trips generated and attracted in each zone based on analyses of household surveys and various socio-economic inputs.
2. **Distribution:** In the second step of the model, and based on costs between zones, the zonal production and attraction trip-ends from the previous step are linked to create zone-to-zone trip matrices.

[7] Since this is the last full year for which data is available from a range of sources.

3. **Choice model:** The third step of the model estimates the choices in selecting or changing the mode to complete the trip. The mode choice includes auto, HSR, rail, long-distance bus and air, along with the combined use of surface transit modes (HSR, rail and long-distance bus)
4. **Assignment:** The final step is to determine the routes chosen across all the aforementioned modal options, including separate assignments for the following main travel modes:
   - Auto;
   - Conventional Rail;
   - Long-distance bus (including Thruway service);
   - Air;
   - Combinations of surface transit (rail and/or HSR, plus long distance/connecting intercity bus);
   - High-Speed Rail express;
   - High-Speed Rail limited;
   - High-Speed Rail all stop; and
   - High-Speed Rail bus.

Model coverage

The consideration of the spatial area and network density is a balance between having a detailed enough representation of the study area to capture all of the key impacts of potential rail improvement projects while not being so large that model runtimes can become a problem.

The model covers the entire state of California and a limited number of large metropolitan areas in adjacent states.

Zones are aggregations of the traffic analysis zones (TAZs) defined in the CSTDM, plus some external zones without further geographic breakdown:

- Zones within California: The model includes 1,169 zones within California. Full correspondences to CSTDM zones are included, allowing results at this zone level to be readily translated.
- Zones outside of California: The model includes 17 zones representing individual cities outside of California that are considered to be an important part of the current/future rail system in California.

The external zones are included with network details for connections to California. In this way, demand from these external zones is incorporated within the model and will impact the ridership and revenue forecasts produced. Furthermore, the model is able to undertake ***high-level tests*** in relation to these external zones, for example, estimating the volume of demand on these flows for different infrastructure scenarios for the specific zones representing cities outside California, such as Las Vegas and Reno.

Using observed trip matrices ("pivoting")

Even after a rigorous calibration, it is inevitable that the actual base year trip-making patterns are not fully replicated by the model and its base year input data. To account for this, we use observed matrices developed as pivots to create the final trip matrices by mode that feed the model's assignment step.

As such, the Model has two types of trip matrices:

- **Synthetic trip matrices:** These matrices are the main output of the first three steps of the model and are built for the base and future years. These matrices are synthesized by the model using data in the form of trip rates, gravity models and logit estimations.
- **Observed trip matrices:** These are the trip matrices based on observed data, such as demand volumes and travel patterns for all the modes considered within the model. These matrices represent the current (pre-COVID) situation and, therefore, are built only for the base year.

The observed matrices provide the pivot for the so-called incremental part of the model. Pivoting is applied to county-level trips before running the assignment step. The matrices are built outside the model using information available on travel by each mode.

This approach is appropriate to address the important requirements of the model, including the need to estimate the impacts of large projects (such as the HSR project) and of potentially smaller projects (such as those incorporated within the California State Rail Plan). Without this pivot, it is possible that the forecast levels of rail ridership on certain existing systems could be significantly different from current observed levels, meaning that forecasts of more incremental changes to these existing systems would produce potentially unreasonable results.

Generation

The purpose of this step is to estimate the total number of trips generated and attracted in each zone of the model. Regression models are typically used to estimate trip rates based on socioeconomic information associated with zonal population, households, employment, school and macroeconomic activity. Projections of this data for each zone are needed to prepare forecasts of the trip ends for the future.

Our approach

We use regression models to model trips. Segmentation is key, as it provides the different behaviors that a model like this needs to capture.

The production and attraction trip generation models are built based on the observed travel behavior obtained from the NHTS. The models are then validated based on observed trip tables. Trip production rates are developed using the cross-classification of household socioeconomic characteristics by trip purpose. The attraction model is developed based on regression analysis using household survey data, employment by occupation and population data.

Calibration

We calibrated this step by adjusting the trip rates so the total estimated trip ends match closely, where possible, with the observed number of trips. We undertook this comparison at an aggregate purpose level, using the purposes we have gathered as part of our data collection activities.

Future years

For future years, we assume the trip rates for all purposes remain constant[8] and use forecasts of socioeconomic variables developed for the state.

Where a scenario is being tested that is expected to lead to changes in land use (for example, densification and transit-oriented development (TOD)), the model inputs need to be updated to reflect this. This update process is through a manual input, rather than endogenous to the model.

Distribution

At the end of the generation step, we have purpose-specific trip ends for each zone. The next step is to link the productions and attractions to create (synthetic) trip matrices. This is done through a gravity model that takes the form of a deterrence function that disincentivizes travel as distance, time or costs increase. The distribution is performed by purpose and on a production/attraction basis. Costs are fed from the network assignments and adjusted iteratively until some degree of convergence is achieved between demand and supply.

While the trip ends from the first step of the model represent relatively robust estimates, it is acknowledged that the distribution of those trips is harder to estimate, particularly at a statewide level. The reason for this is that trip distribution, by its nature, is based on many factors – not simply travel costs.

Our approach

Data from NHTS was used to estimate the trip distribution model. A combination of NHTS (2016-2017) and California Add-on data was used to assess the travel behavior of residents of California. There were 26,095 households surveyed in California, and their travel was expanded to the entire population of the state. The NHTS is fairly comprehensive, covering people across demographics and geography; hence, the behavior of people observed is considered to provide a good representation of the population. There are always outliers in the data, and these outliers cannot be captured in the models; the models are built for representing general observed behaviors.

To validate the origin-destination travel patterns from the trip distribution process, observed trips between zones are compared with those modeled. This observed data is described in the relevant base demand appendices.

Calibration

We have calibrated the distribution step at two levels:

- The first level focused on calibrating the deterrence functions to match trip length distributions by purpose using observed data and other regional travel surveys.

[8] Within our base case – changes to this assumption can be tested as part of sensitivity testing of the Model.

- The second level focused on adjusting the location-specific K-factors[9] to modify the synthetic distribution to better represent the attractiveness of different zones and thus match better the observed distribution.

Future years

For future years, we assume that the deterrence functions and K-factors remain fixed and use costs from the future year assignments to run the gravity models.

Choice model

The choice model represents the impact of travel costs[10] on travel behavior decisions. The structure of the choice model has been determined through our behavioral research.

The inputs to the choice model are demand from the distribution step and costs from the assignment step.

Our approach

The mode choice element is purely focused on the core issue of mode choice, with the trip frequency and trip destination steps considered within the earlier elements of the model.

Mode choice is between the main modes – auto, HSR, other Intercity rail, long-distance bus, air and combo (a combination of rail, HSR and long-distance bus).

Airport and rail/bus station choice

Where reasonable, origin and destination zones should be able to access the intercity network at multiple airports or rail/bus stations (we will generically refer to these as "stations" hereafter).

Developing a complex station choice model is not a primary objective of this model. Therefore, we utilize the EMME functionality that distributes each origin and destination zone's modal demand among a reasonable set of available stations. We note this here as part of the mode choice (since station choice is an important component of this), but in practice, this part of the model is implemented within the assignment step (discussed subsequently in this section).

Access/egress mode choice

The model distinguishes between "private auto" modes (park and ride, kiss and ride), "shared auto" (taxi, TNCs) and "all transit" modes (subway, LRT, local bus, etc.) when considering access/egress connections to/from airports and stations.

The CSTDM auto network has been replicated in the model in full, as well as all subway and LRT lines to support those modes. Local bus has adopted the CTSDMv2 approach at a link level. From

[9] K-factors are used to account for individual zonal (or zone-group) variation that is not accounted for in the main gravity model.

[10] Costs here refers to both the time and monetary components that constitute a trip (for further details, see section entitled **"**Incorporate time/travel cost in network."

this representation, the accessibility by auto and transit between each zone and each station available to it can be calculated.

The choice model works at both the access/egress mode choice and the main mode choice, leading to each main mode having up to nine access/egress combinations. This is integrated within the mode choice component of the model.

Calibration

We have calibrated the choice model at two levels:

- The first level focused on the mode choice step. We have checked that the model reproduces, to the extent reasonably possible, the observed modal splits for all the main movements observed in the base year.
- The second level focused on checking that the whole choice model responds realistically to changes in travel costs. For this, we undertook a series of tests changing the travel cost components (for example, fares, fuel cost and travel times) and verifying that the overall demand responses (or elasticities) are in line with available benchmarks to the extent reasonably possible.

Even after a rigorous calibration, the actual base year trip-making patterns are not fully replicated by the model and its base year input data. To account for this, and as a final step, we use the observed matrices as pivots to create the final trip matrices by mode that feed the model's assignment step.

Time of day

The model does not consider travelers' detailed choice of trip timing (i.e., the exact time of day when travelers would ideally like to start or end their trips) or the various factors that affect this. Rather, the model divides the day into a limited number of time periods and assumes factors to convert daily trip matrices into corresponding time period matrices for assignment.

Demand matrices are factored from the 24hr matrices into five time periods and differ by purpose. The periods are shown in Table 1.1, with the factors being shown in Table 1.2.

Table 1.1: Time periods

Period	Code	Hours	Number of hours
Weekday AM peak	AM	06:00 – 10:00	4
Weekday midday	MID	10:00 – 15:00	5
Weekday PM peak	PM	15:00 – 19:00	4
Weekday evening	OFF	19:00 – 00:00	5
Average weekend day	WKD	06:00 – 00:00	18

Table 1.2: Time period factors

Purpose	Factors to convert daily P-A out and back trips into O-D by period						
Business		AM	MID	PM	OFF	WKD	Total
	AM	0.081	0.281	0.120	0.009	0.005	0.496

Purpose	Factors to convert daily P-A out and back trips into O-D by period						
	MID	0.005	0.193	0.087	0.006	0.004	0.295
	PM	0.005	0.001	0.034	0.064	0.001	0.105
	OFF	0.001	0.005	0.001	0.011	0.000	0.018
	WKD	0.002	0.000	0.002	0.000	0.082	0.086
	Total	0.094	0.480	0.244	0.090	0.092	1.000
Commute		AM	MID	PM	OFF	WKD	Total
	AM	0.013	0.105	0.432	0.041	0.002	0.593
	MID	0.001	0.012	0.040	0.052	0.001	0.106
	PM	0.004	0.001	0.006	0.031	0.002	0.044
	OFF	0.002	0.079	0.096	0.008	0.000	0.185
	WKD	0.005	0.002	0.001	0.000	0.064	0.072
	Total	0.025	0.199	0.575	0.132	0.069	1.000
Leisure		AM	MID	PM	OFF	WKD	Total
	AM	0.056	0.098	0.055	0.005	0.003	0.217
	MID	0.020	0.093	0.093	0.022	0.003	0.231
	PM	0.018	0.010	0.072	0.111	0.001	0.212
	OFF	0.008	0.007	0.004	0.040	0.000	0.059
	WKD	0.004	0.002	0.001	0.000	0.274	0.281
	Total	0.106	0.210	0.225	0.178	0.281	1.000
Other		AM	MID	PM	OFF	WKD	Total
	AM	0.085	0.186	0.090	0.006	0.001	0.368
	MID	0.012	0.183	0.078	0.007	0.001	0.281
	PM	0.005	0.005	0.074	0.046	0.002	0.132
	OFF	0.005	0.006	0.006	0.030	0.000	0.047
	WKD	0.004	0.005	0.001	0.000	0.162	0.172
	Total	0.111	0.385	0.249	0.089	0.166	1.000
Non-res		AM	MID	PM	OFF	WKD	Total
	AM	0.088	0.244	0.088	0.022	0.004	0.446
	MID	0.003	0.204	0.072	0.013	0.001	0.293
	PM	0.002	0.000	0.060	0.060	0.000	0.122
	OFF	0.007	0.003	0.004	0.013	0.001	0.028
	WKD	0.007	0.001	0.008	0.007	0.088	0.111
	Total	0.107	0.452	0.232	0.115	0.094	1.000

The conversion factors used to split demand into each time period is assumed to remain constant for future years.

Assignment

The assignment step constitutes the primary detailed representation of the supply side in the overall model. The assignment step takes trip matrices from the previous steps, in OD format by time period and mode, and assigns them to a representation of the transportation network. The assignment step outputs travel volumes at a link level for each mode and time period and provides travel times and costs for each OD movement for input to the model.

Our approach

The main attributes of each of the assignment models are:

- Highway assignment model:
  - Non-dynamic highway assignment[11] using fixed travel times from the CSTDMv2 network, with metro area times adjusted using data from MPO models.
- Transit assignment model:
  - Multimodal assignment including intercity buses, conventional rail services and HSR.
  - Station choice included as part of transit assignment.[12]
  - No representation of crowding in the assignment.[13]
- Air assignment model:
  - Simple assignment model[14] using fixed travel times, costs and air service assumptions.

These components are used to implement the model's route choice assumptions and carry out basic book-keeping functions such as tallying up ridership by line segment, as well as boardings and alightings at stations.

By the modeling approach described above, an end-to-end trip consists of:

- An access leg between the origin zone and an intercity network node;
- A mainline or trunk leg consisting of one or more intercity modes from the origin station to the destination station; and
- An egress leg between an intercity station and the destination zone.

Calibration

Each assignment model has been calibrated separately to ensure that the model suitably replicates observed travel conditions to the extent reasonably possible. Additional focus has been

[11] Also termed "loading."

[12] The modeling process is limited to stations deemed to be reasonable for any given zone.

[13] An assessment could potentially be made as part of a post-processing step if required.

[14] Also termed "loading."

on the rail assignment, including reasonableness of service allocations and proportion of people transferring modes.

Summary

Our approach uses an **incremental four-step model** with**:**

- A generation step of trip rates estimated from data from NHTS and adjusted to match the observed trip ends;
- A distribution model estimated also using NHTS data;
- A choice model that includes auto, HSR, long-distance transit and air in the mode choice set. Observed matrices are used as pivots to estimate final matrices by mode; and
- An assignment model for five time periods and assignment routines for highway, HSR, long-distance transit and air.

The following table summarizes the inputs, outputs, segmentation, and calibration/validation undertaken within each step of the model.

Table 1.3: Modeling framework

	External Inputs[15]	Outputs for next step	Segmentation	Calibration / Validation
Generation	• Socioeconomics (state and MPO models) • Trip rates (NHTS & CSTDM) • Observed trip ends	• Daily trip ends by purpose	• Trip purpose • Zone (Vector) • Daily	• Matching observed trip ends
Distribution	• NHTS data • Costs from assignment (feedback) • Observed matrix	• Daily trip matrices (all modes)	• Purpose • OD/PA • Daily	• Trip length distribution by purpose
Choice model	• Behavioral research (Steer) • Costs from assignment (feedback)	• Daily trip matrices (by mode)	• Purpose • OD/PA • Daily • Mode	• Calibration of estimated matrices to observed • Reasonableness of behavioral responses
Time of Day	• Estimated factors (various)	• Up to five time periods (by mode)	• Purpose (by time periods) • Mode • Time period	• None – input is fixed

[15] The primary inputs anticipated to be used are listed here. This does not mean that other sources will not also be used as part of the Model development.

	External Inputs[15]	Outputs for next step	Segmentation	Calibration / Validation
Assignment	• Networks/level of service • Access/egress to stations/airports /highway from MPO • Behavioral parameters	• Skims of distance, time, cost • Ridership by service	• User class • Mode • Time period	• Volumes • Travel time • Trip patterns

2 Model Limitations

Travel demand modeling is inherently subject to significant uncertainty. When developing forecasts for any proposed service – whether rail or any other mode of transportation – there is no “single” answer. The manner in which people make trip and mode choices today and the way these trends may evolve in the future are governed by a myriad of factors, some of which are difficult to identify and understand even at the level of a specific individual, let alone the entire population.

For example, individuals may make mode decisions based on factors including perceived comfort or safety, which could change depending on their trip purpose (for example, traveling for business or leisure) or who they are traveling with (for example, alone or with their family). While one can attempt to estimate the overall impacts of such preferences through mode constants – potentially varied by journey purpose or size of travel group – this will only ever approximate behaviors even at the individual level.

Thus, when seeking to expand this to considerations of the population of an entire state, and indeed travelers from further afield, as any modeling suite will have to it do, it becomes necessary to use statistically derived averages that are inevitably limiting and may fail to pick up the full nuance of these movements. That is not to say that the model itself is “wrong” or that the output forecasts aren’t useful; rather, it highlights the importance of understanding the purpose for which the model was developed, its strengths and weaker areas, and the data, assumptions and parameter values adopted when developing any forecasts. These realities will be set out, as appropriate and as needed, in any caveats that explain the model’s limitations, how it should and should not be used and within what range of scenarios, thereby enabling intelligent interpretation. All parties should, ideally, agree to avoid the blind acceptance or promulgation of headline output numbers without the associated description, explanation and (agreed) interpretation being an integral part of the official output.

In relation to the model discussed in this document, the section entitled “Introduction” is the understanding of the purpose of the model (in the form of the key requirements regarding its use and outputs). Below, additional limitations under the following sub-sections are discussed:

- Model data, assumptions and parameters;
- Model inputs and outputs;
- Model focus; and
- Model use.

Model data, assumptions and parameters

In relation to the assumptions adopted within the model, it is important to highlight that any model will be able to produce potentially very different forecasts depending on the input data, assumptions and parameters utilized.

In some cases, this may be attributable to specific assumptions being outside the reasonable ability of the model to produce a plausible and stable behavioral response. For example, if one were to assume that gas prices increase in future years to $20 per gallon (in real terms) in isolation – i.e., without any corresponding impacts on the costs of traveling by other modes – the model is likely to forecast a huge shift in demand away from auto and towards other modes (including rail). Such a shift may be entirely appropriate. However, it will only be possible to seek to calibrate the model to handle changes in assumptions within the bounds of what has been observed historically or what could be reasonably perceived by individuals as a plausible future scenario. Beyond such bounds, while the model will theoretically be able to provide output forecasts, it is not reasonable to consider such forecasts as anything other than highly speculative and subject to significant uncertainty (and lacking the required behavioral stability).

In other cases, the individual parameter values may appear reasonable in isolation, but their cumulative effect on the model may result in output forecasts that are considered to be unreasonable. For example, this may include a range of assumptions related to a future rail service – ease of access, interoperability of ticketing platforms, seamlessness of transfers, quality of rolling stock, reliability of service – which in isolation appear reasonable, but in combination result in forecast ridership levels which go beyond the bounds of that observed on more mature systems exhibiting many of these attributes. Another example may relate to assumptions that all other proposed complementary projects in the state, whether committed or aspirational, are delivered to time, cost, budget and scope/functionality by the time the HSR project is operational.

It is critical, therefore, that any forecasts using this model – be they those presented in this report or others to be produced as part of later work – are viewed explicitly in the context of the data, assumptions, and parameters that underpin them.

Model inputs and outputs

The model outputs can only be as good as the model inputs. Any limitations in the model inputs necessarily lead to limitations in the model outputs, irrespective of the quality of the processing algorithms and overall model structure and flows. These limitations will include both the quality of any data that Steer receives from third parties – for example, growth forecasts provided by Caltrans and the Department of Finance – and any limitations due to Steer not receiving requested information, leading us to necessarily make additional assumptions or utilize potentially less robust alternative data sources. Steer has, as part of our professional duty of care and diligence, reviewed and checked all third-party inputs for basic provenance, coherence, consistency and credibility. Steer does not, however, guarantee the accuracy or reasonableness of any third-party information or data.

Model focus

The focus of this model is intercity rail travel across the entire state of California. This is a wide coverage area and brings with it a range of inherent limitations, including the level of confidence that can be reasonably achieved regarding different levels of outputs from the model.

Specific consideration of future rail services

At the time of developing the model, a number of scenarios were known to be a focus of the analysis, including, in particular, a few HSR scenarios and the State Rail Plan scenario. As such, the behavioral research undertaken and the model developed have been explicitly tailored to focus on collecting inputs and developing a model framework that aligns with these scenarios to cover the entire state and reflect appropriate regional differences.

However, it is intended that the model will, in the future, potentially be used for the assessment of any intercity rail investment across the entire state. While this scope is known – and therefore, the geographic extent of the model has been defined to encompass this scope – it is not possible to collect behavioral research data specific to each of these potential future options (especially since many of them cannot be known at this stage and there a limit as to the number of options that can be incorporated as part of any behavioral research). The behavioral research was, therefore, necessarily more general in nature, seeking to obtain behavioral parameters that can be applied across the entire state (albeit segmented where appropriate by key market/segment to capture observable differences in behaviors).

This approach is considered appropriate given the nature of the model and its intended future use. However, it does limit its ability to always provide a high level of confidence when assessing some specific future projects to something less than the level that might be achievable should a specific scope be commissioned to look at any given future project.

Ability of the model to calibrate across all modes and services throughout the State

The model needed to represent each of the core intercity modes and services across the State, and we sought to calibrate it as far as reasonably possible to the behaviors of passengers on these modes and services. Given this wide focus, that calibration is almost certainly less accurate than could be achieved if the calibration were focused on only a single mode or service (for example, the Pacific Surfliner service only) or a smaller geographic region or corridor.

This issue is not unique to the development of the model; any statewide model is likely to encounter similar issues.

The model framework approach has sought to mitigate this to the extent reasonably possible through the application of an incremental model approach, including a pivot to estimated existing demand. However, even with this approach, there will nonetheless be some limitations in the level of accuracy achieved for any specific service.

Aggregation of access/egress modes

Given the statewide focus of this model, it is not possible to consider local transportation options at the level of detail that can be achieved through, for example, an MPO model. Indeed, the level of effort that goes into the development of an individual MPO model is often equal to or greater

than the level of effort committed for this work. There are 18 MPOs across California, most of which have their own local models. It would, therefore, not be practical within any reasonable timeframe or budget to consider modeling that level of detail within this model. (Additionally, any model attempting to do so would be enormously complex and completely impractical to use given the long run times this would cause.)

Further, the access and egress situations typically are not impacted by changes to the longer-distance travel options (the core focus of the model). As such, there is typically no need to re-run the access and egress calculations when alternative intercity rail service plans are tested.

As such, the approach is to utilize outputs from MPO models as much as possible and appropriate to aid our understanding of local movements and, in particular, to feed the access/egress assumptions with regard to intercity travel. Within this, it has been necessary to aggregate the local transportation options available into broad modes. For example, it is not possible to consider each individual local bus route within each MPO models. Instead, the model incorporates overall times and costs for access/egress by bus but does not know precisely which bus (or combination of buses) an individual might take to access the core intercity network.

This is the standard practice and is considered to be appropriate and sufficient for understanding intercity movements and enables the model to provide outputs estimating the volume of demand accessing the core network by different modes. However, the level of detail in these outputs will necessarily be increasingly lower the more focused one gets in terms of geography. As such, while these outputs can be used to adequately provide a general indication of access/egress by different modes, more detailed local analysis would be needed if a detailed understanding of these movements is required.[16] Should more detailed local changes want to be considered, it would be possible to make adjustments in individual MPO models and then feed these into the model, or else simply make higher-level adjustments to the aggregate access/egress times and costs to approximate the desired change.

Model use

Finally, there was an explicit tension in our scope of work between the development of a model that is sufficiently detailed to answer the range of questions that may be posed and the development of a model that is sufficiently efficient (in terms of run time) to be of practical use. The document has sought to highlight this tension throughout and elements to the approach have been explicitly proposed to seek to find a balance in this regard.

Outlined below are some specific areas to be aware of with regards to this tension[17]:

- **Station choice:** Modeling of station choice is incorporated as part of the access/egress modeling. The modeling process is limited to stations deemed to be reasonable for any given zone.

[16] Any more detailed local analysis deemed to be required in future is not part of Steer's current scope of work.

[17] Some of the items listed overlap with items highlighted earlier in this section; they are repeated here for completeness.

- **Zoning:** The use CSTDM as a starting point for the zoning system but group its zones into zones for the model in a way that is less aggregate in and around the urban areas and more aggregate elsewhere.
- **Model validation**: The model zones cover the entire state, but the primary validation effort in relation to the HSR system has been limited to the Phase 1 coverage area. Further, the model has been calibrated to seek to match existing rail ridership to the extent reasonably possible for each of the in-scope services. Validation of model outputs against rail and intercity bus services/connections, and air demand data (i.e., not simply focusing on the total numbers, but also how this is distributed) have been sought. Due to the large scale of the model and the intercity focus of the model platform, the model against local transit counts such as city bus or transit connections have not been validated.
- **External zones**: External zones include demand from neighboring states. While choice modeling is undertaken for these external zones, the demand and networks are focused on principal cities and their relationship to California. Tests of network changes to these zones may not capture all the expected impacts (such as more local stations for an enhanced rail service).
- **Modeled time periods:** The model develops forecasts for 5 time periods. However, the model does not consider travelers' detailed choice of trip timing (i.e., the exact time of day when travelers would ideally like to start or end their trips) or the various factors that affect this (it is not a "schedule-based model"). As such, should the impacts of a new timetable want to be tested, the following process would be followed:
  - Convert the timetable information into the required model input[18] for each of the time periods. This could be through the use of a simple averaging process or potentially through a more detailed assessment of average wait time (for example, to approximate the benefits of a more evenly spread schedule).
  - Run the model, which will output the forecast ridership in each time period.
  - If required, potentially seek to distribute the projected ridership at a more detailed level through a post-processing step (i.e., the model itself would only output results at a time period level; any addition required detail would need to be based on reasonable assumptions applied outside of the model).

The next several sections will describe in more detail some of the inputs to the model and the modeling steps.

[18] The precise input is yet to be determined and will, at least in part, depend on the final formulation from the behavioral research undertaken.

3 Supply and Assignment

Network

Development of networks for the California Rail Ridership model (CRRM) was done with an objective of representing the multi-modal travel options for the State of California. The California Statewide Travel Demand model v2 (CSTDMv2) has been used as the initial basis for the supply network for this study, supplemented with available General Transit Feed Specification (GTFS) files for transit data and data from available Metropolitan Transportation Organization (MPO) models.

Develop physical network elements

The supply network for the model captures the full range of available and relevant modes across the state of California, including:

- Rail and Thruway bus;
- Scheduled air services;
- Scheduled Intercity bus services;
- Subway and Light Rail Transit (LRT) services, along with a representation of local bus services, for access/egress; and
- Highway and principal roads network.

The CSTDMv2 model network was used as the foundation of the CRRM network. This CSTDMv2 network was available for years 2020 and 2040 in Cube Software. The network was translated from Cube to EMME and converted from the year 2020 to 2018 network conditions for use in the base year and 2040 for the future year. Since the CRRM model has a focus on rail, the network was enhanced with detailed rail and access information from various sources, as shown below for the base and future year networks:

- GTFS are available from each transit agency for their existing routes. The files contain information on stations, routes and schedules. The data, which includes GIS information, can be imported into EMME networks or used to update the schedule information on the current transit routes. These files were used for adding or updating many multi-modal services in the network.
- Where GTFS information was not available, printed and online information was collated and employed.
- The CSTDMv2 approach to local bus transit services was adopted, using auto times and measures of local service levels obtained from FTA data.
- High-speed rail service information to develop scenarios for future year HSR alternatives was provided by the client.

Headways on non-auto modes are initially entered as the true headway for the respective period. For skimming and assignment, these are converted into effective headways using a wait time penalty curve. This curve reflects the behavior and perception of non-auto users of random arrivals for low headways (such as for typical subway systems) transitioning to timed arrivals for scheduled services (such as Intercity rail or air travel).

Incorporate time/travel cost in network

Auto times are initially taken from the CSTDMv2 network for the base year network but adjusted using MPO model data to reflect average speeds across the MPO model areas, as well as changes therein for future years.

These congested link times are assumed to be fixed within each network year. This approach has significant benefits in terms of model run times. The implicit assumption of this approach is that road travel times will not materially change as rail demand changes. This is considered a reasonable assumption given the relative expected mode shares across the state. This also has the advantage of mitigating any issues related to the use of speed flow curves to derive road travel times when auto travel demand may not be fully modeled and included.

The derivation of the end-to-end route time and cost is therefore a combination of these elements, as follows:

- For auto trips, the derived auto network will provide end to end auto times and distance.
- For trips where the main mode is air, HSR, rail etc., various combinations are possible, comprising:
  - Local access/egress of auto, taxi/TNC or a combination of walk, subway, LRT and local bus transit, and
  - The trunk mode itself.

The travel time in the network is represented by the In-vehicle time (IVT) – the time spent in the auto, train, bus, or plane on the main trip mode. Travel time for a trip also includes the walk time from/to or between stations, waiting time for transit, auto access times to transit or park and ride lots. While actual travel times can be derived from the network via a skimming process, these travel times are perceived different by different riders or different modes, e.g., in-vehicle time is typically perceived more favorably to time spent walking or waiting for transit. The values of these factors are obtained from behavioral research which was carried out as a part of this project.

Traveling typically incurs an “out of pocket” monetary cost as well, with the principal elements being the fare (for air, rail, bus, transit or taxi), gas and parking (at the trip attraction). For auto trips, the “hidden” costs of vehicle ownership and maintenance may not always be perceived by travelers as part of the costs of an individual trip but may be considered when undertaking more extensive trips (in particular, over longer-distances).

For a given origin-destination, the perceived time and monetary cost elements are combined through utility coefficients to calculate the total travel utility (the values for IVT and monetary cost capturing the implied value of time).

The following table provides details of a trip chronologically by each mode and what core components are typically utilized when considering supply data.[19]

Table 3.1: Components of a trip: time and cost

Element	Auto	Air	HSR / rail / intercity bus / transit
Time			
Access from origin		Access time to airport/station/stop	
At departure airport/station/stop		Time from arriving at airport to plane departing (including check-in, security, boarding and wait time)	Time from arriving at station/stop to train/bus departing
Main mode	Drive time by auto	Gate-to-gate flight time	HSR/rail/bus/transit journey time
At arrival airport/station		Time from plane arriving to leaving the airport (including deplaning, baggage claim and walking to egress mode)	Time from rail/bus/transit arriving to leaving the station/stop
Egress to destination		Egress time from airport/station/stop	
Cost			
Access from origin		Dependent on mode used to access airport/station/stop: - Access by auto: Gas, tolls & other perceived operating costs - Access by taxi (incl. Uber/Lyft): Fare + tip (where applicable) - Access by transit: Statewide average transit boarding fare to reflect the related cost as an average value	
At departure airport/station/stop		Auto parking (where applicable)	
Main mode	Gas & other perceived operating costs	Airfare	HSR/rail/bus/transit fare
Egress to destination		Dependent on mode used to egress airport/station/stop: - Same components as access from origin	

Network validations

The CSTDMv2 was used as the initial data source for the development of the CRRM network, supplemented by data from the MPO models, as follows.

MPO Models Data

The following figure shows a map of the California MPOs.

[19] Other components, such as the comfort, ability to work on-board and expected reliability of each service, can also influence the perceived 'cost' of a trip. These components are considered as part of the behavioral research.

Figure 3-1: Map of California Metropolitan Planning Organizations (MPOs)

Source: Caltrans

The following table summarizes the MPO models used to inform CCRM development and the availability of future scenarios and data.

Table 3.2: Summary of MPO model data used

MPO	Model Base Year	Model Forecast Years Provided	Model Data Provided
Association of Monterey Bay Area Governments (AMBAG)	2015	2020, 2035, 2040	No
Butte County Association of Governments (BCAG)	2014	2040 (and 2020 + 2035 interim model scenarios)	Yes
Fresno Council of Governments (FresnoCOG)	2014	2018, 2031, 2037, 2042	Yes
Kern Council of Governments (KCOG)	2015	2018, 2037, 2042	Yes
Kings County Association of Governments (KCAG)	2015	2018, 2030	Yes
Merced County Association of Governments (MCAG) San Joaquin Council of Governments (SJCOG) Stanislaus Council of Governments (StanCOG)	2015	2042	Yes
Madera County Transportation Commission (Madera CTC)	2015	2020, 2030, 2042	Yes
Metropolitan Transportation Commission (MTC)	2015	2020, 2035, 2040	Yes
Sacramento Area Council of Governments (SACOG)	2016	2020, 2027, 2035, 2040	Yes
San Diego Association of Governments (SANDAG)	2016	2020, 2030, 2040, 2050	Yes
San Luis Obispo Council of Governments (SLOCOG)	2015	2020, 2035, 2045	Yes
Santa Barbara County Association of Governments (SBCAG)	2010	2020, 2035, 2040	Yes
Shasta County Regional Transportation Planning Agency (SCRTPA)	2010	2015, 2020, 2025, 2030, 2035, 2040	Yes
Southern California Association of Governments (SCAG)	2012	2018, 2030, 2040	Yes
Tahoe Regional Planning Agency (TRPA)	2018	None	Yes
Tulare County Association of Governments (TCAG)	2015	2042	Yes

Validation of the following network elements was carried out and modifications were made where there were significant differences:

- **Zoning system**: The zoning systems for the MPO models do not necessarily match up with the zoning system for CSTDMv2 and hence CRRM. To compare zoning, correspondence tables were developed in GIS by conversion of the shapefile to the same coordinate system as CSTDMv2 and then overlaying of the two shapefiles and computation of the intersection to create a correspondence table between the two zoning systems.
- **Networks:** Spot checks were conducted on the link attributes to check general consistency with CSTDMv2.
- **Demand matrices and socioeconomic data:** A comparison of demand and socioeconomic metrics (such as population, jobs, households) from MPO models with aggregated MPO totals for similar metrics from CSTDM was undertaken. Many MPO models segment demand data differently than CSTDM. Also, as mentioned above, the zoning systems are generally not consistent between the individual MPO models and CSTDM. Therefore, the comparative checks were conducted at a high-level of aggregation. For example, comparative checks between the MPO model and CSTDM were conducted generally at the county-to-county level, rather than at the individual zone to zone level.
- **Skims:** Skims provide zone to zone journey times, speeds and costs that are computed at the assignment stage of travel demand modes. Key origin – destination pairs were selected and compared directly against appropriate skim values from CSTDM**.**

Transit schedules

The following transit schedule data was collected to aid model validation:

- **Rail / Thruway bus / subway / LRT schedules:** information on timetabled services and regular headway services (such as subway and LRT) were collected via GTFS data, supplemented by on-line data of the respective operators and agencies.
- **Air schedules**: similarly, data was obtained from the Official Airline Guide for current or recent services, supplemented by data from airline websites. In addition to the schedules, the supply network also needs to reflect the time spent traversing the airport – essentially the time between arriving at the airport and the flight departure or landing and departure from the airport. The outbound time will include check-in and security, with the inbound including any baggage claim (given the focus is on intra-state trips, there should be no time spent on passport/border control). For these, reasonable assumptions have been made.
- **Intercity bus schedules:** no GTFS data exists for these privately operated services and reliance was placed on data from the websites of the respective operators and agencies (including, but not limited to, FlixBus, Greyhound, Megabus and Bolt Bus).

Rail fares

The main source of data is Amtrak and Thruway service information. This includes all Amtrak rail services operating in California (this includes the California section of longer distance, out-of-state services) and is divided into route segments. The data includes:

- Total demand in each route segment
- Revenue per route segment
- Passenger mile per route segment

Note that the Thruway data represents an entire trip and includes the bus *and* rail segment data. There is no price differential between Amtrak and Thruway services.

This data was analyzed, and an average per-mile fare rate determined by regional market pair.

Long distance bus fares

The average cost per passenger mile was calculated from 2020 trip fares and journey length from Greyhound, Megabus and FlixBus service information. Average fares were derived using the minimum fares, based on 3-week advance travel, and maximum fares, based on next day travel.

Air fares

The main source of flight fares is the DB1B database. This is developed by the Office of Airline Information of the Bureau of Transportation Statistics (BTS) and the database contains air volumes and fares for airport origin and destination across the US. It is based on a 10% sample of airline tickets.

From this data, the average fare and distance for 242 airport ODs for 2018 was analyzed. However, these ODs include route ODs with limited demand and the fare information could be heavily affected by the limited 10% sample size and so the top 40 ODs by demand were also analyzed. Both a demand weighted average of all ODs and the top 40 ODs gave comparable cost rates, and this was adopted for the CRRM.

Auto tolls and operating costs

The extent of tolled highway facilities is limited in California (with a mix of some express and managed lanes, Bay Area bridges, and fully tolled highways in Orange County and San Diego County). Given the complexity of incorporating express and managed lanes, and their limited extent and impact, only the Bay Area bridges, and fully tolled highways are reflected in the network. For these, a high-level approach was adopted, converting average toll values to travel time using a segment agnostic value of time and the derived time penalty equivalent added to the link time.

Auto fuel prices were derived from historic data on pump prices and typical fleet fuel efficiency rates. The Energy Information Administration (EIA) reports historic, current, and forecast values for both fuel prices and fleet efficiency assumptions, which include important effects such as the impact from the shift towards electric vehicles. Further, the EIA provides California-specific data in some publications, allowing us to ensure relevance to the local market. The auto operating cost only reflects out-of-pocket expenditures to operate and maintain the vehicle but does not reflect

depreciation or financing cost for a private vehicle. Therefore, the cost is lower than typical per-mile reimbursement rates or publications describing the total cost of driving.

Parking costs in urban areas were obtained from the MPO models; outside of these areas, a simple, and reasonable, approach was to assume no parking costs. Airport parking costs were obtained from on-line data (for both on-site and off-site operators).

Summary

Table 3.3 provides a summary of the resulting fare rates used in the model.

Table 3.3: Summary of fare rates by main mode

Mode	Auto	Taxi/TNC	Air	Long-Distance Bus	Rail / Thruway Bus	HSR (Future Years)	Other HSR Services (Future Years)
$/mile	0.23	2.98	0.16–1.81 (Varies by Region)	0.14	0.15–0.24 (Varies by Region)	As defined in future year assumptions	As defined in future year assumptions
Fixed Boarding Fee	---	---	---	---	---	As defined in future year assumptions	As defined in future year assumptions

In addition, an average boarding fee of $2.38 for transit access and egress (auxiliary transit modes s and t – defined in

Table 3.4 below) is applied to these access and egress trips.

Skimming and assignment approach

The structure of the EMME network and zoning has been designed with the purpose of meeting the overall modeling framework needs, notably:

- Segregation of main mode and access/egress modes,
- Use of intermediate zones at main mode rail stations, airports and bus (long distance and Thruway) stations to facilitate choice modeling of access location,
- Use of fixed auto times (taken from the CSTDMv2 congested auto assignment) treated as auxiliary transit within the EMME network, enabling transit only algorithms to be used throughout the skimming and assignment process.

Modes

The following table sets out the EMME network modes employed, indicating the main modes and modes treated as auxiliary transit.

Table 3.4: EMME network modes

Modes	Description	Main mode	Auxiliary Transit
a	auto	✓	✓
l	long distance bus	✓	
f	flight	✓	
r	rail	✓	
t	local transit		✓
h	HSR default	✓	
i	HSR limited	✓	
j	HSR express	✓	
x	Other HSR services	✓	
k	HSR bus	✓	
b	ferry		
z	thruway bus	✓	
s	subway/LRT		
w	walk		✓

The set of six main modes and combinations thereof for skimming and assignment purposes comprise the following:

Table 3.5: Main modes for skimming and assignment

Modes	Description	Skimming	Assignment
a	auto	✓	✓
l, z	long-distance bus, thruway bus	✓	✓
f	flight	✓	✓
r	rail	✓	✓
h, i, j, x	HSR	✓	✓
Any combination of: l, r, h, i, j, k, z (with h/i/j/x and/or r mandatory)	Any combination of long-distance bus, rail, HSR, HSR bus, and thruway bus where HSR and/or rail is present in the combination	✓	✓

Skimming

The skimming process provides 46 sets of data (per time-period), derived through:

- Three access and three egress modes (transit, auto and taxi/TNC), nine combinations in total, for each of the five (non-auto) main modes[20]
- Auto all the way as a main mode (no access/egress)

The EMME algorithms allow any modal combination to be skimmed, so the process cycles through the various access/egress and main mode combinations to derive the full set required.

There are a number of monetary costs which are captured via the skimming process, including:

- Auto operating costs
- Airport auto parking costs
- Airfare, rail, HSR, Thruway bus, long distance bus fare costs (on a per mile basis plus boarding fee where applicable for HSR)

These are calculated using skimmed distances, the respective region pair for the trip and routing information (such as the airport used to assign the associated parking cost). Note that monetary costs are not included in the network routing process and are simply a post skimming calculation (the skimming and assignment is purely a time-based algorithm).

Access location choice

As noted, the EMME model includes intermediate zones for the range of main mode access points (thus excluding most transit stations, such as BART, where the transit service falls under the transit access/egress umbrella). To facilitate access location choice where more than one station/airport is possible, the EMME triple-index algorithm is used. This combines all the reasonable permutations of access location choice with the onward travel to the ultimate destination. For choice modeling purposes, the best choice is used; for assignment, the two-leg trip chain process utilizes a logit choice model approach to spread the demand across the reasonable choices of access station.

Overall process

The overall process is summarized below:

1. Skimming and assignment is done using the Extended Transit Assignment functionality in the EMME software, invoking "*Flow distribution between lines*," option "*Frequency and Transit time to destination*". For skimming, a unitary matrix is assigned and the respective times/costs saved; demand matrices from the choice model are assigned in the same way.
2. Skim for access times: transit (using modes *w*, *t*, *b*, *s*) and auto (using auxiliary transit mode *a*). Transit skims are based on generic perception parameters, namely:
   - Walk mode *w* time weighted by 2,
   - Local transit mode *t* (unweighted) and

[20] In practice, auto and taxi/TNC is derived from a single skimming process, with the time common to both and the distance skimmed simultaneously to be used to derive the cost for taxi/TNC using a rate per mile.

– IVT and waiting time for modes *b* and *s* (where waiting time is half the headway, weighted by 2).

The auto skims are the (unweighted) time and distance (to derive taxi/TNC cost). Note that the auto skim also provides the auto as main mode information too.

3. Skim for (non-auto) main mode and range of egress modes: the main modes in turn, each using transit (using modes w, t, b, s) and auto (using mode a) to derive the egress mode[21] skims, using the same parameters and approach as skimming for the access mode.
4. Compute the in-scope access choice set, deriving the logsum combined value[22] for each OD for each main mode access/egress combination (46 per segment and time period) and pass to the demand model.
5. Once the demand model has completed its calculations, the resultant demand by HSR and rail is assigned to the network using the same approach as for skimming.

Worked example.

The following sets out an example trip from Bakersfield to San Francisco by rail, with auto access and transit egress, using the preceding data and processes.

Table 3.6: Example trip by rail: Bakersfield (auto access) to San Francisco (transit egress)

Element	Description	Value
Time		
Access from origin	Access time to station	3 mins (1.2 miles)
Transfer time	Transfer from the access mode to the main mode	5 mins
At departure airport/station/stop	Time from arriving at station to train departing	50 mins headway penalty
Main mode	Journey time	375 mins (287.9 miles)
At arrival airport/station	Transfer time	5 mins
Egress to destination	Egress time from station	62 mins
	Total	500 mins
Cost		
Access from origin	Dependent on mode used to access airport/station/stop: - Access by auto: Gas, tolls & other perceived operating costs - Access by taxi (incl. Uber/Lyft): Fare + tip (where applicable) - Access by transit given the complexity of capturing numerous local fare systems, this	$0.28

[21] Recognizing that this will produce true ODs using the same access and egress modes, but these are not used.

[22] Computing Accessibility Measures for Two-Leg Trip Chains, September 7, 2020, INRO

	cost is currently only included as a statewide flat fee	
At departure airport/station/stop	Auto parking (where applicable)	$0
Main mode	HSR/rail/bus/transit fare	$51.24
Egress to destination	Dependent on mode used to egress airport/station/stop: - Same components as access from origin	$2.38
	Total	$53.89
Daily utility value for choice model		
		-4.9933

Note: Data from Base 2018 model run number 868, AM period, OD 500532-500283, auto-rail-transit, segment 7 (Leisure, Employed, Middle income)

Pivot process

Assignment utilizes the post pivot OD based demand matrices, split by main mode and access/egress combination.

The pivot process is used to match forecast and observed Base demand, with this adjustment being applied to future year forecasts. Two interdependent processes are applied as follows:

- **Base Year:** Synthetic trip matrices at the zone level are scaled to the county-to-county level to replicate observed modal demand.
- **Future Year:** Base year Scalable Quality Value (SQV) factors are used to scale future year forecasts to reflect the differences that existed in the base year. This is done directly for existing modes, with HSR related demand using a scaling factor for total demand. All modes are then scaled to match the total trip volume from the sum of Productions and non-resident trips.

The SQV factor is a statistical measure of the goodness of fit, similar to the GEH statistic used in demand modeling for many years. However, the SQV factor is symmetrical with values between 0 and 1. It has been used in the CRRM as a proxy hybrid of absolute differences and proportional differences, both of which have challenges in their independent application where there are small values and/or material changes in forecast demand.

4 Base Year Demand

This section describes how the base year demand (or trip tables) for each of the relevant intercity mode (auto, rail, air and intercity bus) was developed.

Auto

Objective and use of auto trip table

When considering intercity travel across California there is no single data source that provides a high degree of confidence regarding the volume and patterns of long-distance travel throughout the state. However, a few key sources exist, as follows:

- **Traffic counts on California highways**: They are considered to be a reliable source for traffic volumes over specific highway segments, however they provide no information regarding the origin-destination (OD) patterns of trips and also may include trips that are out-of-scope for our purposes (such as local trips and trips by non-passenger vehicles).
- **Census Transportation Planning Products Program (CTPP)**: Provides a detailed record of journey-to-work trips, but does not include non-commuter trips, and therefore will likely provide an under-estimate of trips, especially for longer distances (where the share of commuter trips is likely very small).
- **2017 National Household Travel Survey – California Add-On (NHTS)**: A survey of trips across California for an assigned travel day,[23] which in theory includes all in-scope trips but will include relatively low sample sizes for longer-distance trips and will also exclude any non-California residents. Further, it will include a large volume of short-distance trips that are not in-scope for our intercity trip table.

Given this lack of clear data, a key part of the development of the CRRM is to develop a robust estimate of current intercity trip patterns by auto, while acknowledging that there will always be inherent uncertainties within the estimate.

It is important to note that this full zone to zone (based on the CRRM zone system) trip table **will not be used directly within the model framework**. Rather, this trip table will be used to provide **control totals at a county-to-county level** within the generation and distribution steps of the model framework.

County level is the lowest level of geography at which we consider we can have reasonable confidence in the auto trip table estimates. The relative zone level estimates within each county from the model framework will therefore be retained. This approach has the benefit of utilizing

[23] Assigned travel dates ranged from April 19, 2016, through April 25, 2017.

our estimated trip tables at the level of geographic disaggregation that can be considered reasonable, while also retaining consistency with wider elements of the model framework (such as the basis on which future socioeconomic changes are estimated to impact travel patterns).

As such, the trip table discussed in this document is not the final base year matrix that will be used within the model framework directly, and hence this section should not be considered as a validation note. Nonetheless, the estimated auto trip table discussed here forms an important component of the development of the final base year matrix.

Potential data sources

Traffic counts on California highways

AADT (Average Annual Daily Traffic) has been collected for 2018 (our model base year). This data comes from two separate sources:

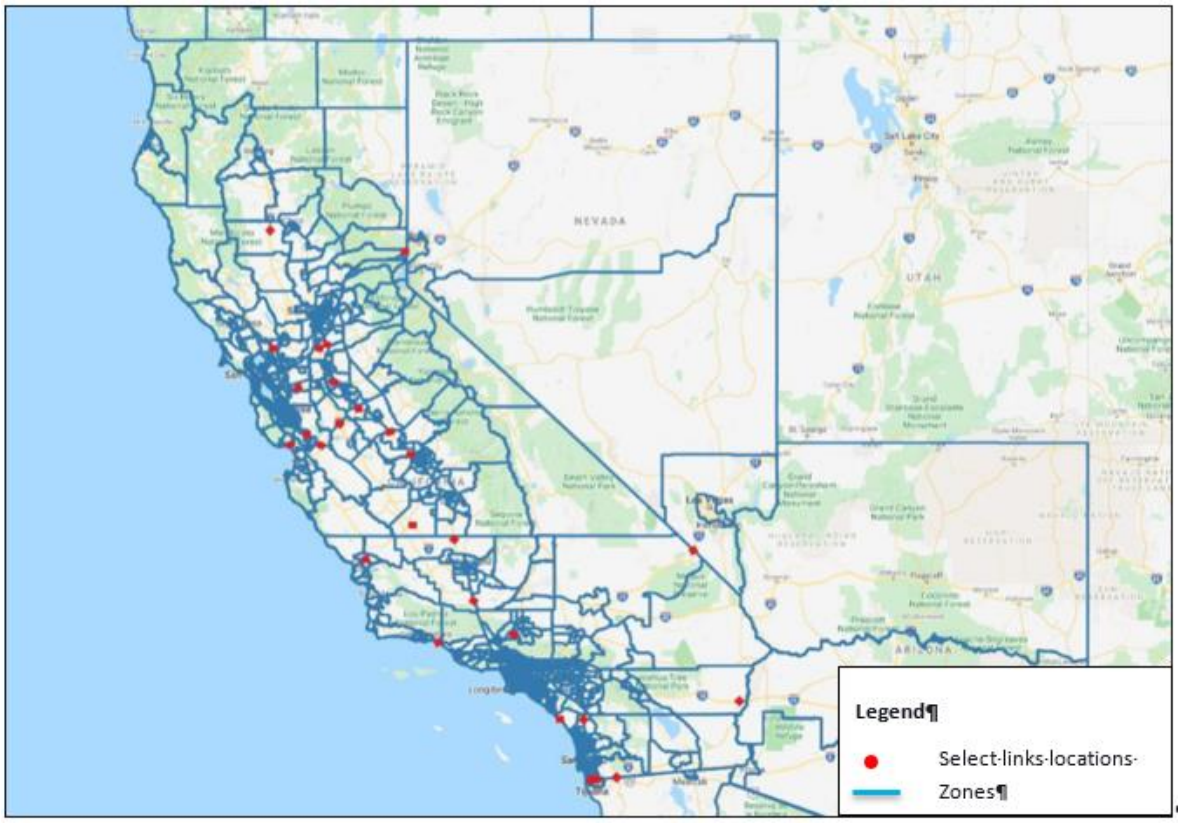
- Caltrans Census Traffic Count Data: This reports AADT counts for selected sites across the state.
- Caltrans Performance Measurement System (PeMS): This provides over ten years of data for historical analysis. It also provides hourly count data and truck shares. The data differentiates between weekdays and weekend days[24].

A map showing the locations of traffic count data collected is provided in Figure 4-1. These are the same locations where we obtained select link OD data for from StreetLight[25] (hence the reference to “select links” in the figure legend).

[24] Although it should be noted that the breakdown between average weekday and weekend used within the estimated trip table comes from StreetLight data given that long-distance trips do not necessarily follow the same temporal distribution as other trips (see section “Development of the trip table”).

[25] Data purchased from StreetLight Data, Inc. 4 Embarcadero Center Suite 3800 San Francisco, CA 94105

Figure 4-1: CRRM zone system and traffic count / select link locations.



Source: Steer

The traffic counts are considered to be a reliable source of volumes over specific select links, however they provide no information regarding the origin-destination (OD) patterns of trips. The counts may include trips that are out-of-scope for the purposes (such as local trips and trips by non-passenger vehicles), although the process used to develop the trip tables has sought to account for these shortcomings (see section “Development of the trip table”).

Given that traffic counts are considered the most reliable source for volumes over specific select links, the trip table has been scaled to these count volumes using an OD Matrix Estimation (ODME) procedure where the raw StreetLight OD trips were assigned onto the highway network. (See section “Development of the trip table” for further details.)

Census Transportation Planning Products Program (CTPP)

The Census Transportation Planning Products Program (CTPP) is a state Department of Transportation (DOT) funded, cooperative program that produces special tabulations of American Community Survey (ACS) data. It includes census data on demographic characteristics, home and

work locations, and journey to work travel flows including mode of travel to work. In particular, we have used 5-year average ACS CTPP journey-to-work data for our analysis.[26]

By itself, the number of commuting trips by auto is not directly comparable to the total auto trips used to estimate for the in-scope trip table (since it will exclude any non-commuter trips and therefore will provide an under-estimate of trips, especially for longer distances (where the share of commuter trips is likely very small)). However, the data is considered to provide a useful comparison, especially for shorter-distance trips such as those in the Central Valley, where it can be reasonably considered as a lower bound estimate for auto trips between given county pairs.

This data is also considered to be of use in terms of confirming the reasonableness of the relativity of different flows – in particular for short- and medium-distance trips. As such, it is used to compare the ranking of different flows to the trip table (see section "Comparison to public sources").

2017 National Household Travel Survey – California Add-On (NHTS)

The California Add-On survey supplements the 2017 National Household Travel Survey (NHTS) with additional household samples and detailed travel behavior for an assigned travel day.[27]

In theory, this source should include all in-scope trips, but it has relatively low sample sizes for longer-distance trips and will also exclude any non-California residents. Further, it will include a large volume of short-distance trips that are not in-scope for our intercity trip table.

We have used this source for two purposes:

- First, vehicle occupancy factors were used from the NHTS by trip distance to convert the vehicle trips data received from StreetLight into person trips.
- Second, as with the CTPP data, while the absolute volume of trips from NHTS are not considered for direct use, it was used to compare the ranking of different flows to the trip table (see section "Comparison to public sources").

[26] ACS 5-yrs **https://www.census.gov/data/developers/data-sets/acs-5year.html**

[27] NHTS 2017 CA Add-On **https://www.nrel.gov/transportation/secure-transportation-data/tsdc-nhts-california.html**, assigned travel dates ranged from April 19, 2016 through April 25, 2017

The vehicle occupancy factors used to convert the StreetLight vehicle trip data into per trip data are shown below.

Table 4.1: NHTS vehicle occupancy

Trip length	Vehicle occupancy
0 – 50 miles	1.7
50 – 75 miles	2.1
75 – 100 miles	2.2
>100 miles	2.4
Average	**1.8**

Source: Steer analysis of NHTS CA add-on 2017

Development of the trip table

StreetLight data

Origin-destination (OD) auto trip data from smartphone location-based services (LBS) are available from commercial vendors, with varying levels of detail and output types. The primary source of OD auto trips is from LBS cell phone data from StreetLight, an established provider of such data within the transportation industry. These vendors (including StreetLight) obtain raw tracking data from a sample of cellular devices and given a zone system definition, process the raw data into population-wide estimates of zone-to-zone vehicle movements over a particular time period. This process ensures that the data is anonymized, and it is not possible to trace any specific user.

In addition to overall vehicle trip tables for the defined zone system, we also collected "select link" trip tables (i.e. the OD matrix of trips that use a specific section of a transportation facility, such as a section of a given road segment); and trip table breakdowns by inferred trip type/purpose (e.g., a person's journey-to-work trips might be inferred by observing the locations where his/her device spends most nights [home] and travels to most weekday mornings [work]).

As noted above, the StreetLight data provides vehicle trips (not person trips).

The StreetLight data was obtained for the entire year of 2019 for both long-distance and regional shorter-distance tours[28] and for 27 select link locations:

- ***Long-distance tours*** are defined as journeys between OD zone pairs whose zone centroids are at least 75 miles apart as the crow flies. Trips are included in the same long-distance tour if there are less than 90 minutes between consecutive trip stops – otherwise they would be included as two distinct tours.
- ***Shorter-distance tours*** are defined as journeys between OD zone pairs whose centroids are less than 75 miles apart as the crow flies. Trips that are part of long-distance tours were excluded while constructing shorter-distance (also called regional) tours. Shorter-distance tours cannot start and end in the same zone. Trips are included in the same regional tour if

[28] We typically refer to the StreetLight data as tours, rather than trips, because it can include trips that have been combined together when there is only a short stop between them. The threshold used for this are discussed subsequently in this document.

there is less than 15 minutes between consecutive trip stops – otherwise they would be included as two distinct tours.

- ***Select link OD matrix:*** In addition to the long and regional OD matrices, 27 select link OD matrices of the long-distance trips were obtained at 27 count locations. For each location, these matrices represent trips that pass through a specific location and therefore can be used to identify flows that can be scaled to match the specific count data at that given location. The proportion of long-distance trips (out of all trips going through a location) was also provided and used to estimate the proportion of long-distance trips at each count location. Finally, the proportion of intra-zonal trips (trips originating and terminating in the same zone) was also provided for the shorter distance trips and used to estimate the proportion of local (out of scope) trips.

Process overview

Steer used the location-based data collected by StreetLight for the development of the auto trip table and adjusted the trip table using OD matrix estimation (ODME) techniques constrained in accordance with observed (adjusted) traffic counts. Counts taken on major highways in rural areas and removed local and non-passenger vehicle traffic were used. Steer used an OD Matrix Estimation technique that requires iteratively assigning the initial (seed) matrix on the network and then scaling the volumes until the modeled volumes match the adjusted traffic counts.

The input matrix is the customed Location-Based Services (LBS) cellphone data from StreetLight. It is scaled to adjusted traffic counts using a set of 27 different network locations and assigning the OD trips onto the highway network used in the model. The output of this process is a scaled trip table that replicates the adjusted counts when it is assigned onto the network.

- AADT was adjusted to only include passenger vehicles and non-local trips; effectively excluding local traffic as well as trucks, commercial vehicles and other non-passenger vehicles from the AADT targets.
- The long and medium distance LBS data input was assigned onto the network and compared it with the adjusted AADT at select links and screenlines. The network uses congested speeds.
- The Origin-Destination Matrix Estimation (ODME) process allowed the input demand matrix to be updated (scaled) using the adjusted count data in the network.
- After ODME, we compared the resulting matrix with other data sources such as the previous BPM V3 trip tables. All comparisons were made at the daily level.

The network developed specifically for this study was used in this exercise. A travel distance skim was completed and compared to Google Maps drive distance to check the quality and connectivity of the network. The traffic assignment and ODME process were carried out within the EMME software. The network congested speeds were used for the estimation process.

Limitations and caveats

There are known limitations to the usage of LBS data to inform travel patterns for a large variety of trip types over a large geographic area like the state of California. These include:

- The vendors' processes to end trips is often based on typical characteristics of urban rather than intercity travel.

- Their processes to expand the sample data to the entire population frequently rely on imperfect input data.
- The ODME process might dampen demand in very congested corridors due to the count representing throughput, whereas the matrices represent demand.
- The ODME process is not guaranteed to produce a unique OD trip table given the ground truth, i.e., the adjusted traffic counts; and
- Because of the strict anonymity requirements, it is generally not possible to obtain trip maker details (e.g., demographic characteristics) directly from the cell phone data. Rather, vendors typically derive these based on socioeconomic characteristics of the Census tract where the device is resident.

The following challenges were identified to address:

- Steer requested custom processes to end trips:
  - For long-distance trips, typically over 75 miles, 90-minute time breaking criteria was used to end a trip (if a device was seen stopping for less than 90-minutes at a given location, the trip would continue and be chained until the device is seen within the same zone for more than 90 minutes).
  - For shorter-distance trips we used a 15-minute time-breaking criteria.
- The Steer process to expand the sample data to the entire population relies on observed traffic counts along the corridor. Counts taken on major highway in rural areas and removed local and non-passenger auto traffic were used. Steer used an OD Matrix Estimation technique that requires iteratively assigning the initial (seed) matrix on the network and then scaling the volumes until the modeled volumes match the adjusted traffic counts.
- To mitigate the risk of over-relying on this input as well as the ODME process potentially generating multiple trip tables for the same traffic counts, volumes at the county, MPO and regional pair levels with the CA NHTS (including add-ons) and the CTPP journey to work data were compared. The volumes observed give us comfort that the ratios and the resulting trip table used are appropriate.
- Also compared were the volumes with established trip rates by distance, and the resulting trip rates were aligned with our expectations.
- Steer compared the socio-economic data provided by StreetLight with demographic information from the US Bureau of Economic Analysis.

These mitigation actions serve to provide a greater level of confidence in the data, but nonetheless a degree of uncertainty in the volume and distribution of trips will always exist given the nature of the problem.

Approach details: Trip table preparation

The following key inputs are used:

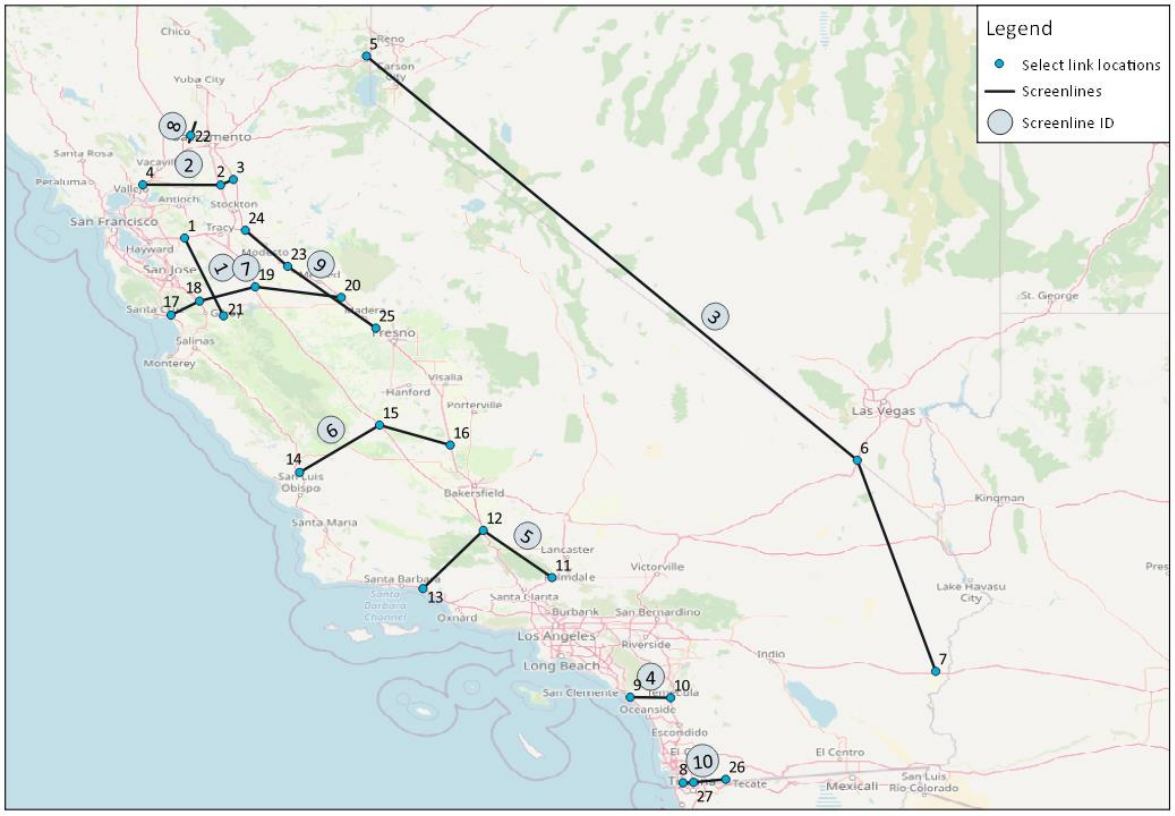
- **Input matrices from StreetLight**:
  - Origin-Destination analysis of personal, long-distance tours (>75 miles).
  - Origin-Destination analysis of personal, regional tours (<=75 miles).
  - Origin-Destination with middle filter (i.e., select link) analysis of personal, long-distance tours and proportion of local traffic which is derived by comparing the select link volumes to the total count value less heavy truck traffic.

- **Network:** CRRM network. The network developed specifically for this study was used. A travel distance skim was completed and compared to Google Map drive distance to check network and connectivity.
- **Zone system:** 1,186 zones, including 1,169 internal zones covering all of California and 17 zones external to California covering neighboring states.
- **Select links:** 27 count locations on major highways and border crossings were selected including some prominent highway locations, and three select links added at the border crossings at San Ysidro / Tijuana, Tecate and Otay Mesa.
- **AASHTO vehicle classification:** Only class 2 vehicles (passenger autos) and a proportion of class 3 vehicle (pickups and vans) are included in the auto trip table effectively excluding trucks, commercial pickups and motorcycles.
- **Software:** The EMME software was used for traffic assignment (best path or all-or-nothing path building procedure) and ODME procedures to adjust matrices to match adjusted count volumes.

Select links and screenlines.

The 27 count locations were grouped into 10 screenlines. Figure 4-2 shows the 27 select link locations and the 10 screenlines used to factor the LBS trip table.

Figure 4-2: Screenline locations map



Source: Steer

Adjusted AADT for ODME

The AADT needs to be adjusted to reflect the proportion of local traffic and the proportion of trucks and other non-passenger vehicles such as commercial pickups and vans – both considered out-of-scope, by definition, for the auto table that we are developing. Table 4.2 shows the series of adjustments made to the observed traffic counts at the 27 select links locations to obtain adjusted targets for the OD Matrix adjustment process.

Table 4.2: Select link counts and adjusted targets.

Select Link ID	Screen line ID	Name	AADT 2018	Long to short trips ratio [R1]	Non-passenger vehicle proportion [R2]	Proportion of short distance trips that are local (intra zonal) [R3]	Adjusted AADT for long distance (D1)	Adjusted AADT for short distance (D2)	Adjusted AADT (combined D1 + D2)
1	1	SF I580	198,960	0.15	27%	30%	21,858	87,212	109,070
2	2	Sacramento I5	61,755	0.38	37%	23%	14,884	18,730	33,614
3	2	Sacramento I99	73,817	0.25	27%	41%	13,516	23,988	37,504
4	2	Sacramento I80	212,000	0.22	21%	39%	36,916	79,716	116,632
5	3	Reno Ext	34,828	0.48	27%	46%	12,244	7,188	19,432
6	3	Vegas I15 Ext	46,116	0.89	28%	69%	29,354	1,122	30,476
7	3	Phoenix I10 Ext	28,000	0.83	27%	81%	17,022	656	17,678
8	10	Tijuana	139,000	0.17	27%	28%	17,308	60,754	78,062
9	4	San Diego I5	143,421	0.44	9%	60%	57,350	29,156	86,506
10	4	San Diego I15	139,601	0.19	27%	27%	19,426	60,546	79,972
11	5	LA Lancaster SR14	91,272	0.09	27%	17%	6,016	50,434	56,450
12	5	LA Bakersfield I5	88,523	0.85	27%	26%	55,110	7,218	62,328
13	5	LA Santa Barbara 101	56,284	0.41	23%	53%	17,676	12,000	29,676
14	6	Route 101	10,800	0.22	27%	70%	1,740	1,832	3,572
15	6	I5	38,042	0.95	36%	54%	23,116	558	23,674
16	6	Tulare I99	62,147	0.56	29%	77%	24,648	4,540	29,188
17	7	Santa Cruz Route 1	74,104	0.07	27%	33%	3,800	34,042	37,842
18	7	San Jose 101	126,673	0.23	27%	18%	21,340	58,626	79,966
19	7	I5	39,130	0.86	42%	43%	19,642	1,836	21,478
20	7	Merced 99	46,330	0.62	30%	66%	20,156	4,176	24,332
21	1	Bay Area 152	24,651	0.57	27%	18%	10,316	6,386	16,702
22	8	Red Bluff I5	29,200	0.79	27%	77%	16,896	1,016	17,912
23	9	Merced Modesto 99	75,833	0.42	27%	55%	23,328	14,614	37,942
24	9	Modesto Stockton 99	126,483	0.22	27%	45%	20,380	39,986	60,366
25	9	Fresno Madera 99	80,304	0.43	30%	20%	24,230	25,694	49,924
26	10	Tecate (SR 188)	7,600	0.16	27%	27%	890	3,414	4,304
27	10	Otay Mesa (CA 905)	46,000	0.11	12%	31%	4,452	24,730	29,182
		Total	**2,100,873**	**0.35**	**26%**	**36%**	**533,614**	**660,170**	**1,193,784**

Source: Steer analysis of CA PeMS, CA AADT data and truck shares, and StreetLight local, long and short distance trips. Caltrans AASHTO vehicle classification, NHTS California Add-On, and StreetLight long to short distance trips and intra-zonal trips.

The long to short trips ratio [R1] was obtained by comparing the number of trips within each StreetLight select links matrix for long distance trips with the total number of devices seen at each select link.

The proportion of non-passenger vehicles [R2] includes truck and other non-passenger vehicles such as commercial pickups and vans. It was obtained from Caltrans using the detailed 15-class ASSHTO classification scheme.[29] Only Class 2 vehicles (passenger autos) and a fraction of Class 3 vehicles (other 2-axles, 4-tires, single unit vehicles such as pickups and vans) were included in the adjusted AADT. The proportion of non-commercial pickups and vans was estimated from the NHTS California Add-On survey where personal pickups and vans account for 11% of personal (non-commercial) passenger vehicle trips.

The proportion of local (intra-zonal) traffic [R3] was obtained from StreetLight by comparing the short distance volumes with the number of trips starting and ending within each zone.

Separate targets were set for long and short distance trips to ensure that the proportion of short and long-distance trips is respected in the output matrix.

Table 4.3 shows the same adjusted targets but aggregated at the screenline level.

[29] AASHTO 15 Classification: Class 1 – Motorcycles; Class 2 - Passenger Cars; Class 3 - Other Two-Axle, Four-Tire, Single-Unit Vehicles; Class 4 – Buses; Class 5 - Two-Axle, Six-Tire, Single-Unit Trucks; Class 6 - Three-Axle, Single-Unit Trucks; Class 7 to Class 14: Four-or-More Axle; Class 15 – Unclassified.

Table 4.3: Screenline counts and adjusted targets.

Screenline ID	Screenline name	AADT 2018	Long to short trips ratio [R1]	Non-passenger vehicle proportion [R2]	Proportion of short distance trips that are local (intra zonal) [R3]	Adjusted AADT for long distance (D1)	Adjusted AADT for short distance (D2)	Adjusted AADT (combined D1 + D2)
1	East of Bay Area (E-W)	223,612	0.20	27%	29%	32,174	93,598	125,772
2	South of Sacramento (N-S)	347,572	0.25	25%	37%	65,316	122,434	187,750
3	California East Border (E-W)	108,944	0.74	27%	56%	58,620	8,966	67,586
4	North of San Diego (N-S)	283,020	0.32	18%	41%	76,776	89,702	166,478
5	North of Los Angeles (N-S)	236,078	0.45	26%	27%	78,802	69,652	148,454
6	North of Bakersfield (N-S)	110,988	0.66	31%	74%	49,504	6,930	56,434
7	South of Bay Area (N-S)	286,238	0.34	29%	28%	64,938	98,680	163,618
8	North of Sacramento (N-S)	29,200	0.79	27%	77%	16,896	1,016	17,912
9	Highway 99 (N-S)?	282,620	0.33	28%	41%	67,938	80,294	148,232
10	California South Border (N-S)	192,600	0.16	23%	29%	22,650	88,898	111,548
Total		**2,100,872**	**0.35**	**26%**	**36%**	**533,614**	**660,170**	**1,193,784**

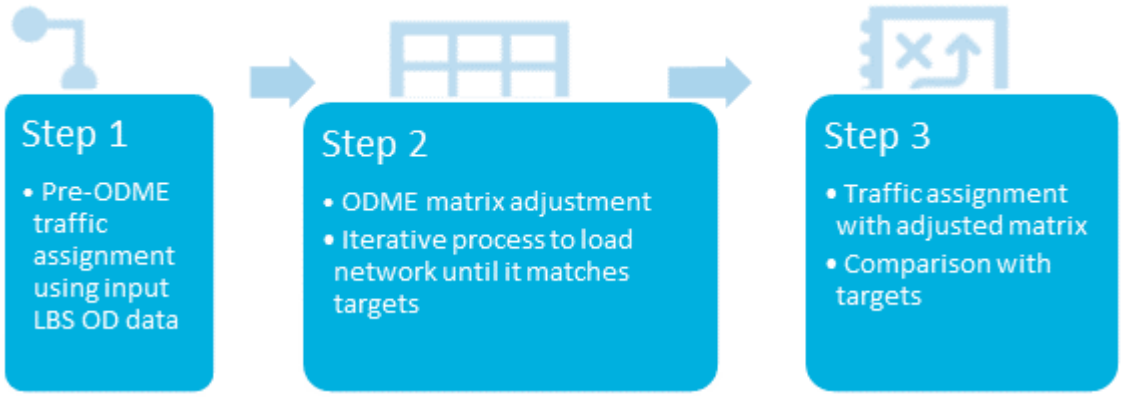
Source: Steer analysis of CA PeMS, CA AADT data and truck shares, Caltrans AASHTO vehicle classification, NHTS California Add-On, and StreetLight long to short distance trips and intra-zonal trips.

Overall, approximately 57% of the raw AADT value is estimated to be attributable to trips that are in-scope for the trip table (with the remainder attributable to out-of-scope local trips and non-passenger vehicles).

Approach and key results

The matrix estimation process is shown in Figure 4-3. It includes assigning the input (raw) StreetLight LBS data on the network (step 1) and scaling the flows until the assignment matches the adjusted traffic counts (step 2). The resulting matrix is then assigned to the network using an all-or-nothing (shortest time path) assignment process using free-flow speeds for comparison with the targets adjusted AADT (step 3).

Figure 4-3: Auto base demand matrix estimation process



Source: Steer

Origin-Destination matrix adjustments were conducted using the EMME software to scale the input StreetLight LBS matrix to actual adjusted traffic flows. Post-ODME traffic assignments outputs are shown below for the 27 select links and associated 10 screenlines.

Traffic assignment post-ODME

Table 4.4 compares the target traffic counts with the post-ODME scaled StreetLight volumes on the network at the screenline levels. Overall, the scaled matrix accounts for 99% of the adjusted AADT volumes.

Table 4.4: Traffic assignment post ODME by screenline – all traffic

Select Link ID	Target (all)	Pre-ODME ALL	Post-ODME all	Post-ODME vs. Target (all)
1	125,772	8,961	125,879	100%
2	187,750	9,592	187,697	100%
3	67,586	3,710	65,788	97%
4	166,478	12,547	166,377	100%
5	148,454	7,054	148,399	100%
6	56,434	2,942	54,805	97%
7	163,618	10,460	163,358	100%
8	17,912	711	17,170	96%
9	148,232	9,639	148,077	100%
10	111,548	1,086	104,215	93%
Total	**1,193,784**	**66,702**	**1,181,765**	**99%**

Source: Steer

Table 4.5 compares the target traffic counts with the post-ODME scaled StreetLight volumes with short- and long-distance targets presented separately. Long- and short-distance trips account for 99% and 99% of the target long- and short- distance traffic, respectively.

Table 4.5: Traffic Assignment post ODME by screenline – long and short distance trips separately

Select Link ID	Target (Long Distance)	Target (Short Distance)	Pre-ODME long	Pre-ODME short	Post-ODME long	Post-ODME short	Post-ODME vs. Target (long)	Post-ODME vs Target (short)
1	32,174	93,598	1,891	7,070	32,178	93,701	100%	100%
2	65,316	122,434	2,891	6,701	65,250	122,447	100%	100%
3	58,620	8,966	3,264	446	58,600	7,188	100%	80%
4	76,776	89,702	4,031	8,516	76,675	89,702	100%	100%
5	78,802	69,652	3,145	3,909	78,657	69,742	100%	100%
6	49,504	6,930	2,614	328	49,561	5,244	100%	76%
7	64,938	98,680	3,532	6,928	64,735	98,623	100%	100%
8	16,896	1,016	663	48	16,884	286	100%	28%
9	67,938	80,294	3,156	6,483	67,746	80,331	100%	100%
10	22,650	88,898	78	1,008	18,731	85,484	83%	96%
Total	**533,614**	**660,170**	**25,265**	**41,437**	**529,017**	**652,748**	**99%**	**99%**

Source: Steer

Regional shorter distance trips account for nearly 60% of all traffic (just over 650k trips). While there are some individual screenlines where larger percentage differences exist for either the long or short distance movements (most notably screenline 8), each of these are for smaller counts (for example, the short distance target for screenline 8 is 1,106 which is more than 6 times lower than any other target). Overall, there is considered to be a strong alignment between the trip table and the adjusted traffic counts.

Sensitivity test on cut-off time

StreetLight was asked to run sensitivity tests to examine the change of trips on the key Origin-Destination markets by various cut-off times before a journey is considered to have terminated and a new journey begun. The test was conducted on 1/16th of the entire sample data in one month of 2019. The subset of sample data was randomly selected to seek to be representative of the annual average. As an example, roughly 80% of the new tours obtained by increasing the tour break criteria to 180 minutes are air trips (have average end-to-end speeds exceeding 90mph). It is highly unlikely for any auto trip to have end-to-end average speed greater than 90mph and hence these would be removed as outliers. The results supported our assumption at a high-level that 90 minutes is a reasonable cut-off point. In addition, scaled the raw data was scaled to targets that are not related to speeds (only to counts). So, any under-counting of the raw data would have been compensated by the scaling process of the ODME.

Limitations and caveats

- There is no full-proof way to develop a long-distance auto trip table. There are several available options. But ODME was chosen to be our preferred option since it allows us to match against the best observed data on record – i.e., actual traffic counts. Nonetheless, there inevitably remains inherent uncertainties within any estimates of traffic developed.
- The proportion of local traffic data from StreetLight cannot be verified using external sources, and, with the proportion of non-passenger vehicles, it is one of the driving inputs during the OD adjustment process. To mitigate the risk of over-relying on this input long and short distance volumes at the county, MPO and regional pair levels with CTPP journey were compared to work data and with the NHTS. The volumes observed provide comfort that the ratio used are appropriate.
- Also compared were the volumes with established trip rates by distance, and the resulting trip rates were aligned with expectations.

Further Auto Base Demand Matrix Adjustment

4.1 As a part of the calibration exercise, the auto-based demand matrix in the Central Valley region was adjusted. Keeping the overall trip total constant within the Central Valley, the distribution of intra and inter county trips were matched closer to LBS data (from Streetlight) trip patterns (as shown in Table 4.6 and Table 4.7) . As a result of the adjustment, the overall auto trips in the region remained unchanged, while the auto VMTs increase by a minimal 2% daily.

Table 4.6: Streetlight Trip Patterns - Central Valley

County	Fresno	Kern	Kings	Madera	Merced	Tulare	Total
Fresno	39.3%	0.1%	0.5%	1.1%	0.2%	0.9%	**42.1%**
Kern	0.1%	29.8%	0.0%	0.0%	0.0%	0.3%	**30.3%**
Kings	0.5%	0.0%	3.5%	0.0%	0.0%	0.4%	**4.5%**
Madera	1.1%	0.0%	0.0%	2.1%	0.1%	0.0%	**3.3%**
Merced	0.2%	0.0%	0.0%	0.1%	4.9%	0.0%	**5.2%**
Tulare	0.9%	0.3%	0.4%	0.0%	0.0%	13.0%	**14.7%**
Total	**42.1%**	**30.3%**	**4.5%**	**3.3%**	**5.2%**	**14.7%**	**100.0%**

Table 4.7: Auto Base Demand Trip Patterns - Central Valley after Adjustment

County	Fresno	Kern	Kings	Madera	Merced	Tulare	Total
Fresno	39.2%	0.4%	0.6%	0.2%	0.1%	1.2%	**41.6%**
Kern	0.4%	29.7%	0.2%	0.0%	0.0%	0.5%	**30.9%**
Kings	0.6%	0.2%	3.5%	0.1%	0.0%	0.4%	**4.9%**
Madera	0.2%	0.0%	0.1%	2.0%	0.0%	0.1%	**2.5%**
Merced	0.1%	0.0%	0.0%	0.0%	4.8%	0.0%	**5.0%**
Tulare	1.2%	0.5%	0.4%	0.1%	0.0%	12.8%	**15.1%**
Total	**41.7%**	**30.9%**	**4.9%**	**2.6%**	**5.0%**	**15.1%**	**100.0%**

Comparison to public sources

As highlighted previously in this document, when considering intercity travel across California, there is no single data source that provides a high degree of confidence regarding the volume and patterns of long-distance travel throughout the state.

However, as previously noted, a few key sources exist, as follows:

- **Traffic counts on California highways**: These have been used directly in the development of the adjusted StreetLight data, and so cannot subsequently be used seeking to compare to other sources.
- **Census Transportation Planning Products Program (CTPP)**: This is a useful source for shorter-distance trips but includes a large number of very short out-of-scope trips and will also significantly under-estimate long-distance trips (as it excludes non-commuter trips).
- **2017 National Household Travel Survey – California Add-On (NHTS)**: A useful source for travel across California but will include a large number of very short out-of-scope trips and will include relatively low sample sizes for longer-distance trips and will also exclude any non-California residents.

Given these limitations, this section focuses on the relative size of different flows across the StreetLight trip table, the CTPP data and the NHTS data, as opposed to the absolute volumes included within any source. This section also focus on county-to-county flows, as opposed to zone-to-zone flows, since a large number of zone-to-zone entries in each dataset will be very small in absolute value, meaning that even small absolute differences in values across datasets could result in quite large changes in terms of relative ranking.

The remainder of this section sets out these county-to-county flow rank comparisons.

Ranking comparisons

There are 58 counties in California, resulting in 3,364 individual county-to-county flows. However, not all sources have data for all flows. The NHTS data includes the lowest number of entries, with non-zero values for only 1,403 (approximately 40%) of the total county-to-county flows. This is not to say that nobody makes trips between these counties (the non-zero values in other sources strongly suggests people do) but is rather a reflection of the nature of the source – a survey with a limited sample as opposed to a comprehensive picture of all trips.

As a result, comparison of rankings only on the 1,402[30] flows which have non-zero values in each of our three sources are the focus. Collectively, these flows account for at least 97.6% of demand in each of the datasets, and so this restriction has minimal impact on the comparisons presented.

The following comparisons are made:

- Top 20 flows: Comparison of the top 20 ranked flows for any County-County flow across the state.
- All flows: Comparison of all non-zero County-County flow across the state.
- Shorter-distance flows: Comparison of all non-zero flows entirely within either the SCAG, MTC or Central Valley[31] areas. The purpose of this restriction is to focus more on shorter-distance flows where the relativities within the public data should, in theory, be more reliable.

For each comparison, two charts are provided:

- **The first comparing the relative ranking between the CTPP and NHTS datasets**: The purpose is first to show the degree to which the public sources align themselves (to demonstrate further the point that there is no clear "ground truth," but also to put the comparison with the StreetLight data into context). This is shown through having the CTPP ranking on the x-axis and the NHTS ranking on the y-axis. If the rankings are fully aligned, then this would output a straight line at a 45-degree angle; any deviation from this indicates the degree to which the datasets do not align.
- **The second comparing the relative ranking between the CTPP and StreetLight datasets**: The purpose is to show how closely the StreetLight data rankings align with the public sources.[32] This is shown through having the CTPP ranking on the x-axis and the StreetLight ranking on the y-axis. As with the first chart, if the rankings are fully aligned then this would output a straight line at a 45-degree angle; any deviation from this indicates the degree to which the datasets do not align.

[30] There is one flow (Del Norte to Orange) which has a non-zero values within the NHTS data but a zero value in the CTPP data.

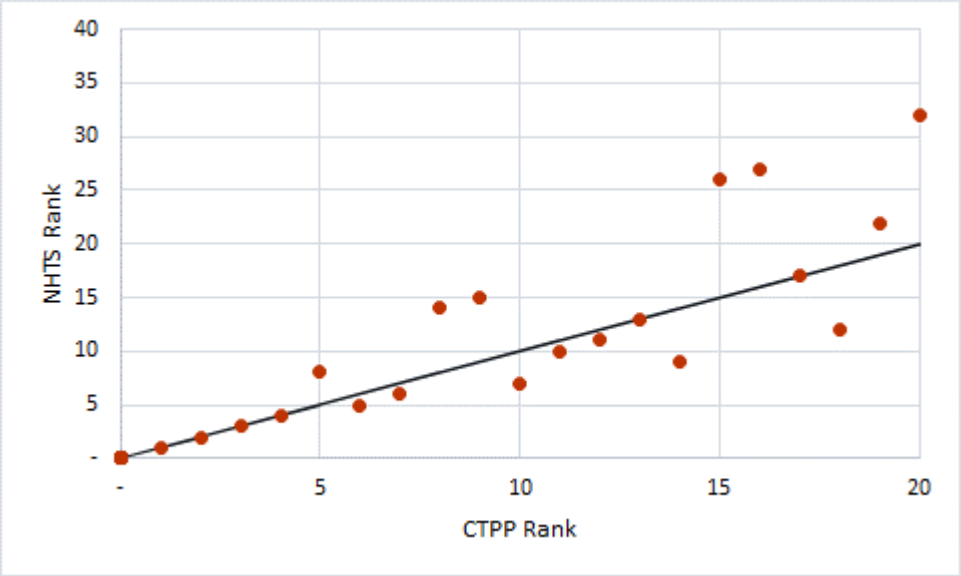
[31] Central Valley defined here as the following counties only (to focus on shorter-distance movements): Fresno, Kern, Kings, Madera, Merced, San Joaquin, Stanislaus and Tulare.

[32] Either the CTPP or the StreetLight data could have been used for this comparison – the CTPP data was chosen since, overall, it provides a more complete dataset (i.e., with fewer non-zero values) than the NHTS dataset.

Top 20 flows

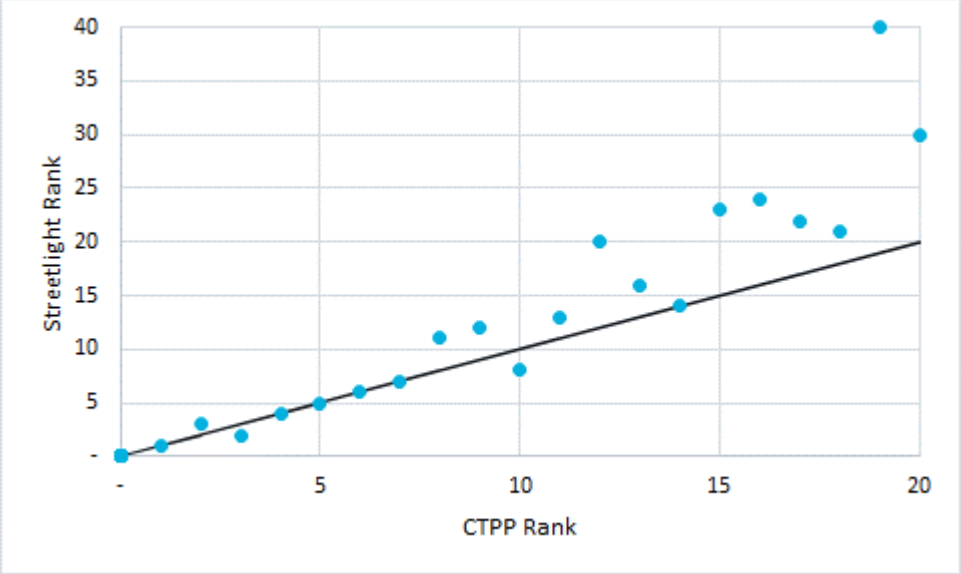
Figure 4-4 shows a comparison of the top 20 ranked flows within the CTPP data to the same flows within the NHTS data. Figure 4-5 then shows a comparison of the top 20 ranked flows within the CTPP data to the same flows within the StreetLight data.

Figure 4-4: CTPP vs NHTS Rankings – Top 20 CTPP flows



Source: Steer analysis of CTPP and NHTS data

Figure 4-5: CTPP vs StreetLight Rankings – Top 20 CTPP flows



Source: Steer analysis of CTPP and NHTS data

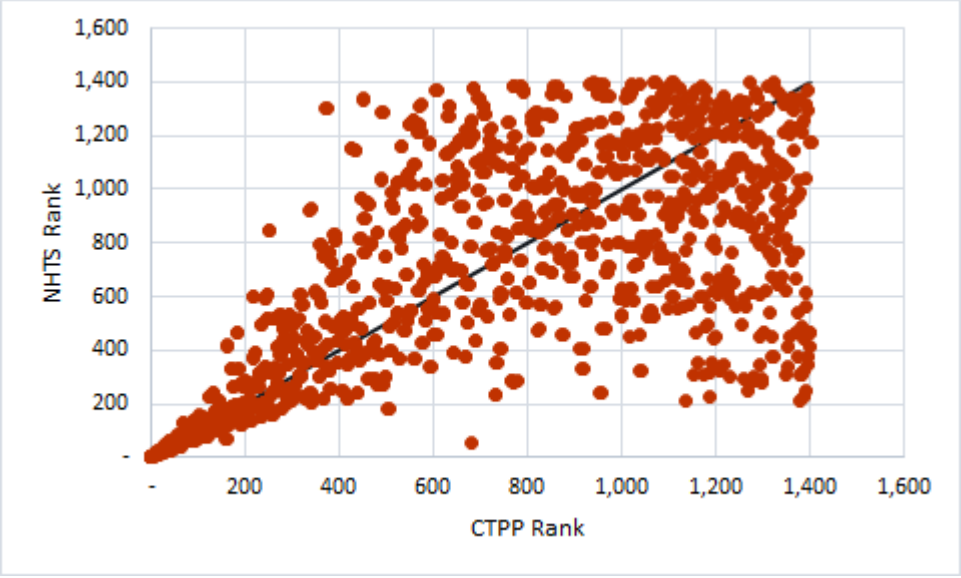
The match between the StreetLight and CTPP data is considered to be as good as the match between the NHTS and CTPP data:

- Each of the top 10 ranked flows within the CTPP data are within the top 12 of the StreetLight data; by comparison, each of the top 10 ranked flows within the CTPP data are within the top 15 of the NHTS data:
  - The root-mean-square deviation (RMSD) for the top 10 flows within the StreetLight data is 1.5.
  - By comparison, the RMSD for the top 10 flows within the NHTS data is 3.0.
  - A lower RMSD indicates a closer match. As such, this indicates that the StreetLight data matches the CTPP data more closely than the NHTS data does, indicating that the uncertainty regarding the relative flow volumes within the StreetLight data is considered to be no more than the inherent uncertainty with any of the potential sources of "ground truth."
- Each of the top 20 ranked flows within the CTPP data are within the top 40 of the StreetLight data (with all but one within the top 30); by comparison each of the top 20 ranked flows within the CTPP data are within the top 32 of the NHTS data:
  - The RMSD for the top 20 flows within the StreetLight data is 8.2 (or 4.8 if the one major outlier is excluded)
  - By comparison, the RMSD for the top 20 flows within the NHTS data is 5.6.

All flows

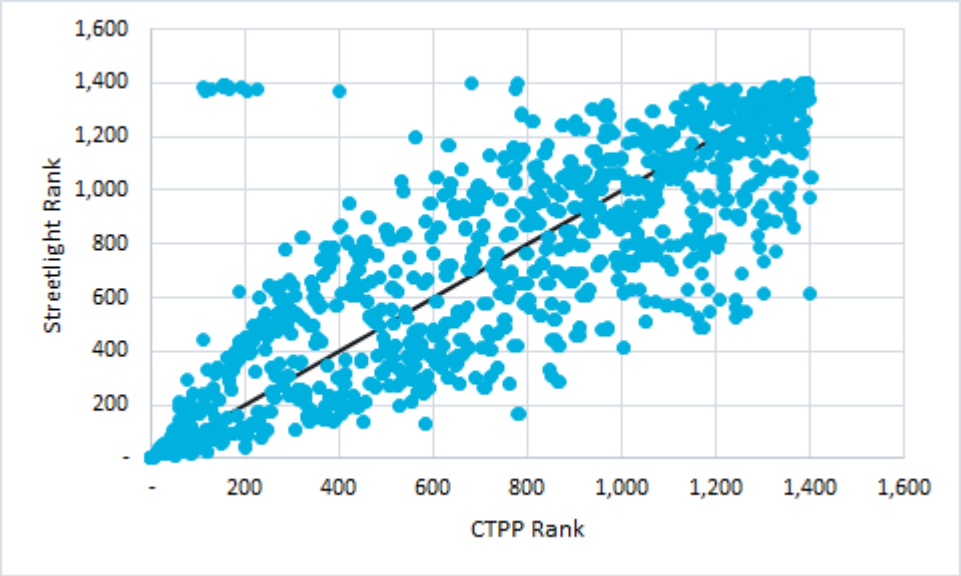
Figure 4-6 shows a comparison of all non-zero flows within the CTPP data to the same flows within the NHTS data. Figure 4-7 then shows a comparison of all non-zero flows within the CTPP data to the same flows within the StreetLight data.

Figure 4-6: CTPP vs NHTS Rankings – All non-zero CTPP flows



Source: Steer analysis of CTPP and NHTS data

Figure 4-7: CTPP vs StreetLight Rankings – All non-zero CTPP flows



Source: Steer analysis of CTPP and NHTS data

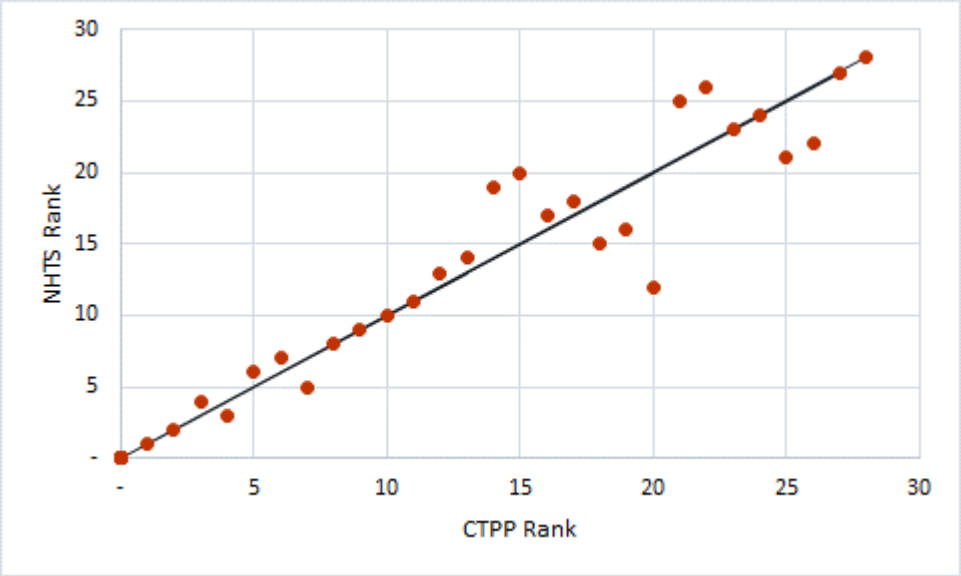
Again, while the deviations in rankings increase significantly (as is expected), the match between the StreetLight and CTPP data is considered to be as good as the match between the NHTS and CTPP data:

- The RMSD for all non-zero flows within the StreetLight data is 260.
- By comparison, the RMSD for all non-zero flows within the NHTS data is 332.

Shorter-distance flows

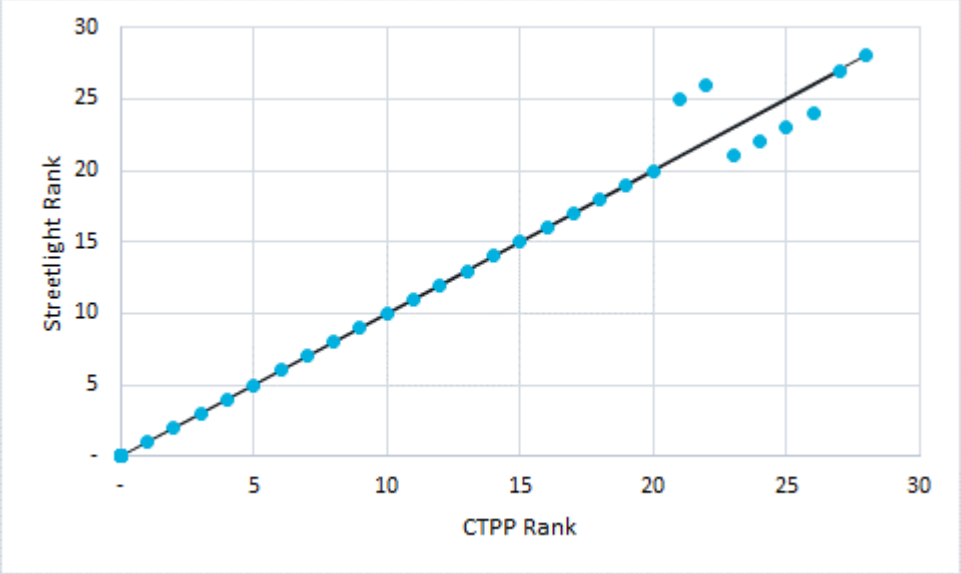
Figure 4-8 shows a comparison of all non-zero flows in the SCAG area within the CTPP data to the same flows within the NHTS data. Figure 4-9 then shows a comparison of all non-zero flows in the SCAG area within the CTPP data to the same flows within the StreetLight data.

Figure 4-8: CTPP vs NHTS Rankings – All non-zero SCAG area CTPP flows



Source: Steer analysis of CTPP and NHTS data

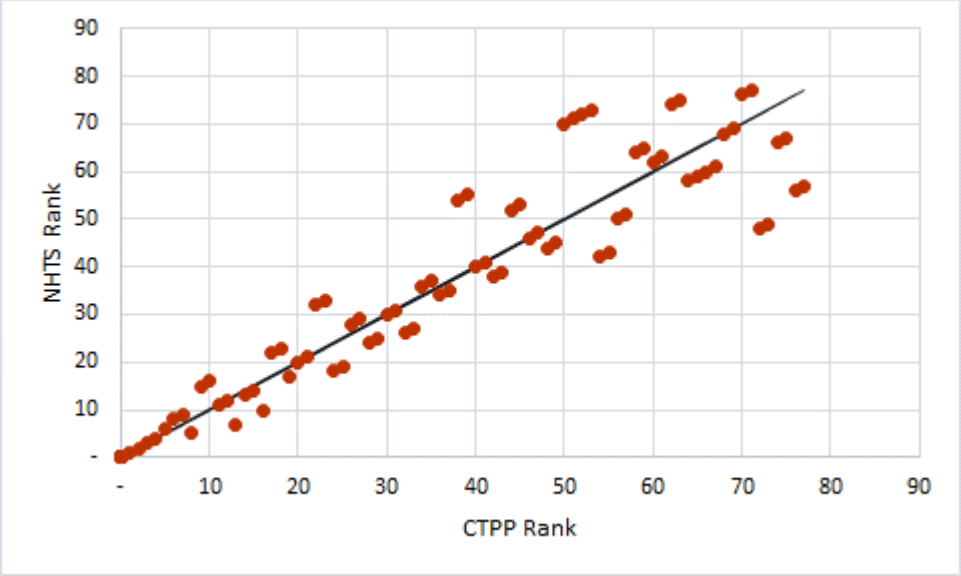
Figure 4-9: CTPP vs StreetLight Rankings – All non-zero SCAG area CTPP flows



Source: Steer analysis of CTPP and NHTS data

Figure 4-10 shows a comparison of all non-zero flows in the MTC area within the CTPP data to the same flows within the NHTS data. Figure 4-11 then shows a comparison of all non-zero flows in the MTC area within the CTPP data to the same flows within the StreetLight data.

Figure 4-10: CTPP vs NHTS Rankings – All non-zero MTC area CTPP flows



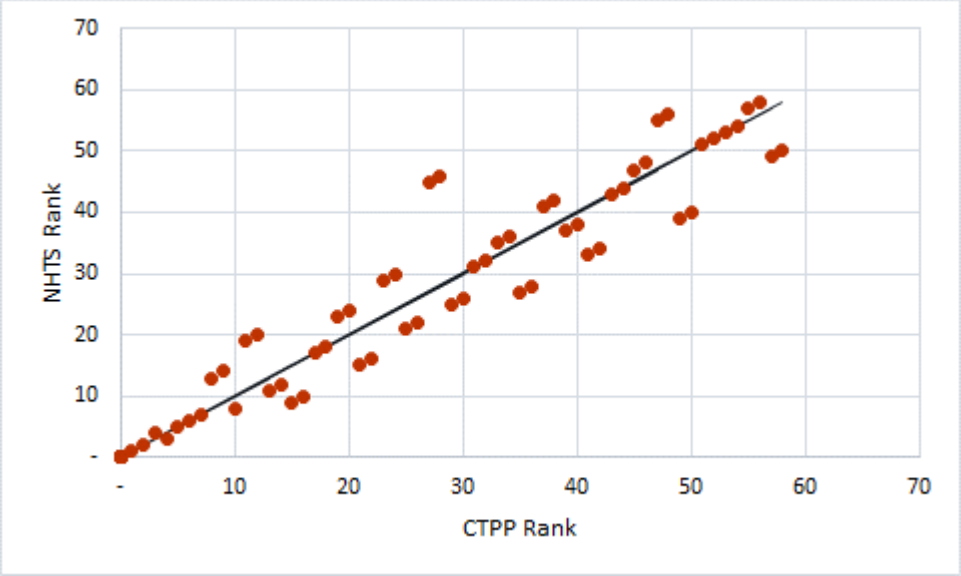
Source: Steer analysis of CTPP and NHTS data

Figure 4-11: CTPP vs StreetLight Rankings – All non-zero MTC area CTPP flows

Source: Steer analysis of CTPP and NHTS data

Figure 4-12 shows a comparison of all non-zero flows in the Central Valley area within the CTPP data to the same flows within the NHTS data. Figure 4-13 then shows a comparison of all non-zero flows in the Central Valley area within the CTPP data to the same flows within the StreetLight data.

Figure 4-12: CTPP vs NHTS Rankings – All non-zero Central Valley area CTPP flows



Source: Steer analysis of CTPP and NHTS data

Figure 4-13: CTPP vs StreetLight Rankings – All non-zero Central Valley area CTPP flows

Source: Steer analysis of CTPP and NHTS data

The match between the StreetLight and CTPP data is considered to be as good as the match between the NHTS and CTPP data:

- The RMSDs for all non-zero flows within the StreetLight data are:
  - 1.3 for the SCAG area
  - 11.1 for the MTC area
  - 6.6 for the Central area
- By comparison, the RMSDs for all non-zero flows within the NHTS data are:
  - 2.7 for the SCAG area
  - 8.8 for the MTC area
  - 5.8 for the Central area

Comparison to modeled sources.

This section provides a comparison between the trip table and other modeled sources available. This comparison focusses mainly on the existing CA HSR Business Plan model V3 (BPM V3) trip tables. For the avoidance of doubt, this comparison is provided not to confirm the reasonableness of the trip table, but rather to provide an understanding of the differences between this previously used model and our estimated trip table.

Sources for comparison

We compared the OD flows of the scaled StreetLight LBS data with the following data sources:

Table 4.8: Sources for comparison

Source	Description	Main Use
Business Plan model V3 (BPM V3) Auto and Air Trip Volume V2V 2029	This is the latest trip table used in the latest CA HSR Business Plan (obtained from the CA HSR Authority). It is based on extensive surveys conducted in 2015 and is mainly focused on longer-distance trips. As such it doesn't have a good representation in shorter distance trips such as those in the Central Valley.	Comparison for trips over 100 miles apart
CSTDM 2015 Daily Auto Trip Volumes	The California State Travel Demand model auto demand volumes come from individual MPO models and household travel surveys. It is therefore more representative of shorter distance trips.	Comparison for trips less than 100 miles apart.
RMAT 2010 Trip Volume Average Weekday	The State Rail Plan RMAT trip table is based on the BPM V3BPM V3 California HSR model trip table and the NHTS. It was developed in conjunction with BPM V3. The data are not an independent source of OD data but is used in the development of forecasts within the California State Rail Plan and so included as a comparator.	Additional comparator mainly for information purposes

Source	Description	Main Use
Teralytics LBS data 2019[33]	Teralytics is, similarly to StreetLight, a third-party vendor providing OD matrix from LBS data. The "trip ends" filter is "30 minutes staying in the same Census tract," the data are validated against CTPP and NHTS surveys, so the data are not an independent source of OD data.	Additional comparator mainly for information purposes
FAA DB1B Air OD data[34]	The Federal Aviation Administration (FAA) Airline Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers collected by the Office of Airline Information of the Bureau of Transportation Statistics. Data includes origin, destination and other itinerary details of passengers transported.	This database was used to determine air traffic patterns and passenger flows within California and compare it with auto flow in markets that have commercial air service.
U.S. Bureau of Economic Analysis (BEA) 2018 Population and Employment[35]	National data tables show the number of full- and part-time wage or salary workers, and the number who are self-employed by type of industry, state and county.	We used this data to compare population and employment centers with trip rates and trip production and attraction centers to assess the reasonableness of the relative distribution of trips.

Source: Steer

Summary across all sources

Table 4.9 shows the daily auto person trips by distance for each source of auto data.

Table 4.9: Daily auto person trips by distance and by data source

Trip Length Category	Scaled StreetLight	BPM V3 2019	CSTDM 2015	RMAT 2010	Teralytics 2019
0 - 100 miles	5,152,279	836,900	16,120,954	109,024,282	72,173,254
> 100 miles	448,720	712,226	89,798	881,126	123,240
Total	5,601,000	1,549,126	16,210,752	109,905,408	72,296,494

Source: Steer analysis of StreetLight, BPM V3, CSTDM, RMAT, and Teralytics, trip tables

The differences across sources are very large, with the highest estimate of trips (RMAT) being almost 100 times higher than the lowest estimate of trips (BPM V3).

[33] Teralytics website **https://www.teralytics.net/**

[34] BTS FAA DB1B **https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=125**

[35] U.S. Bureau of Economic Analysis (BEA) **https://apps.bea.gov/itable/index.cfm**

This significant difference is not surprising. Each source has been developed for a fundamentally difference purpose and as such – especially for shorter distance trips – includes significantly different definitions of what is and is not in-scope.

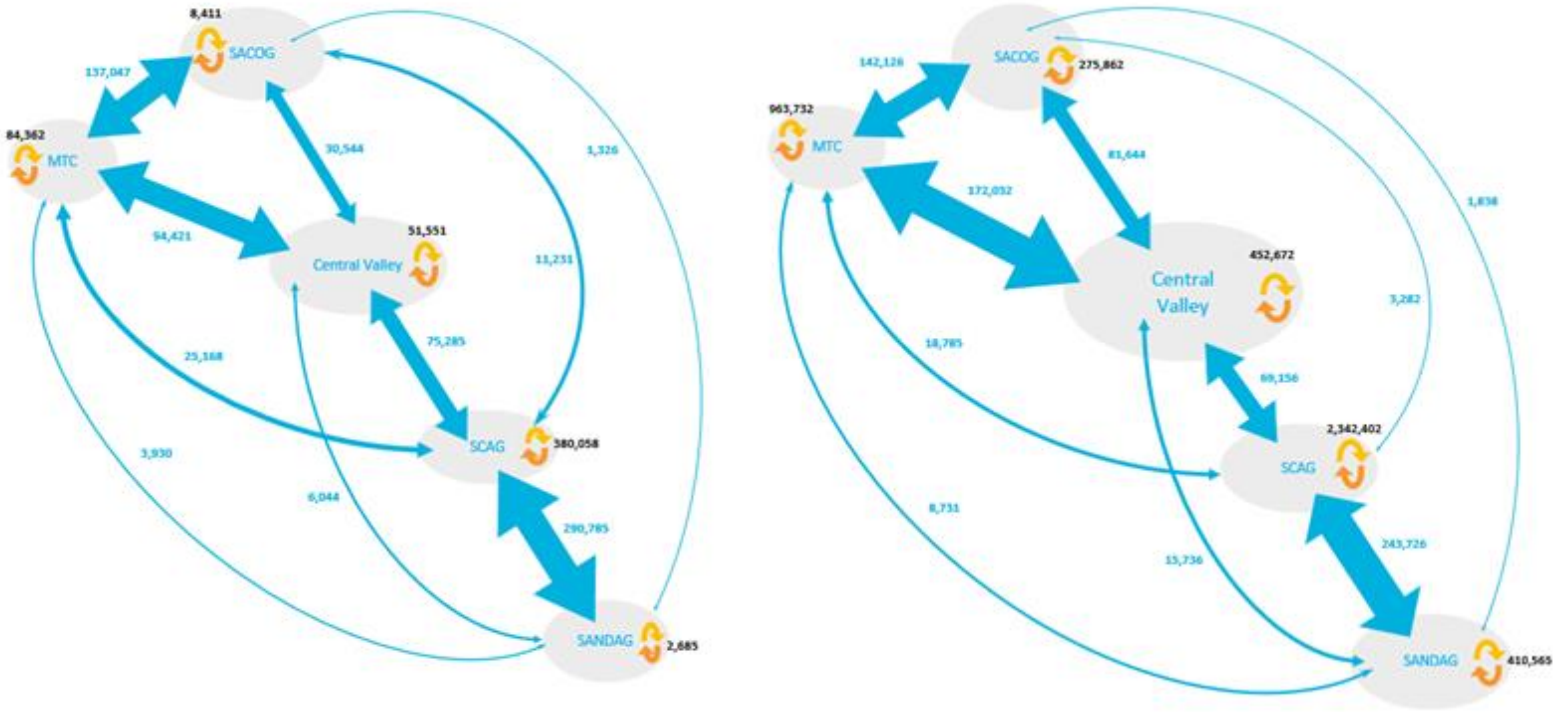
As such, this comparison should be considered for informational purposes only. For the remainder of this section, we focus on the comparison between the scaled StreetLight data (used in our estimated auto trip table) and the BPM V3 data (used in the existing CHSRA Business Plan model).

Comparison to BPM V3: By BPM reported flows.

When displaying summary results for flows throughout California, the outputs from BPM V3 focused on movements between broad regions of primary interest for potential shift to HSR.

Figure 4-14 shows equivalent flow maps for BPM V3 (at left) and StreetLight (at right) respectively.

Figure 4-14: BPM V3 regional-level traffic 2019 (left) vs StreetLight regional-level traffic 2019 (right)



Source: Steer analysis of BPM V3 and StreetLight data

As highlighted in Table 4.9, in total there are significantly more trips within the StreetLight data than within the BPM V3 data.[36] The vast majority of this difference is accounted for by trips entirely within a single region (for example, entirely within the SCAG region):

- There are approximately 4.4m trips internal to each region within the StreetLight data compared to just 0.5m trips internal to each region within the BPM V3 data. This difference can be largely attributed to the difference in scope of each data: The StreetLight data has been developed to seek to understand intercity trips throughout the entire state whereas the BPM V3 data is focused more on longer distance trips specifically along the CAHSR alignment. As such, this is certainly not a like-for-like comparison.
- When considering only inter-regional trips,[37] there are approximately 760k within the StreetLight data compared to 680k within the BPM V3 data (a difference of approximately 12%). As such, while the StreetLight data remains higher, the difference is significantly reduced.

The broad relativity of flows between the two datasets is similar, with the SCAG-SANDAG flow being the largest in each case, followed by MTC-SACOG and MTC-Central Valley (although the relative order of these last two is reversed).

The most material difference between the two sources is that the StreetLight data includes more trips to/from the Central Valley region, most notably from the MTC region. This result is not surprising since the Central Valley was much less of a focus when the BPM V3 matrices were developed.

Between the SCAG and MTC MPOs (the flow that will drive most of the long-distance trips anticipated on CAHSR), the auto trips estimated using StreetLight are lower than BPM V3 (18,785 versus 25,168). In isolation, this would appear to suggest a smaller long-distance market within the StreetLight data. However, the FAA DB1B database[38] reports over 29,000 daily air trips between SCAG and MTC compared to approximately 13,500 within the BPM V3 air matrix. As such, when combining auto and air, the overall demand estimated using combined StreetLight and DB1B is considerably higher than within the current BPM V3 model.

[36] Table 4.9 shows there are 5.6m trips in the StreetLight data and 1.5m trips in the BPM V3 data. Figure 4.14 includes most, but not all, of these trips; the relative total trips in this figure are 5.2m trips in the StreetLight figure and 1.2m trips in the BPM V3 figure (the remainder relate to flows to/from other areas not included in the figure).

[37] Based on the regional definitions used in the BPM V3 outputs.

[38] Used in the development of our base air trip tables.

Auto trip table summary

When considering intercity travel across California there is no single data source that provides a high degree of confidence regarding the volume and patterns of long-distance travel throughout the state.

However, a few key sources exist:

- Traffic counts on California highways
- Census Transportation Planning Products Program (CTPP); and
- 2017 National Household Travel Survey – California Add-On (NHTS).

Traffic counts are used to scale our raw StreetLight data. Our approach is to assign the customed Location-Based Services (LBS) cell phone data from StreetLight on the highway network and to scale it to adjusted traffic counts using OD Matrix Estimation (ODME) techniques. The output of this process is a scaled auto trip table for relevant longer trips that replicates the adjusted counts when it is assigned on the California highway network. The trip table specifically excludes non-passenger vehicles and local traffic that would not be in scope for the modeling.

CTPP and NHTS data are used to compare the relative rankings of flows to confirm the reasonableness of our trip table. The match between the StreetLight and CTPP ranking is considered to be as good as the match between the NHTS and CTPP data. As such, the uncertainty regarding the relative flow volumes within the StreetLight data is considered to be no more than the inherent uncertainty with any of the potential data sources.

Given the inherent limitations of each public source of data, the trip table developed through use of the StreetLight data, adjusted to match highway traffic counts, is considered a reliable and appropriate estimate of existing trips for use within the CRRM.

When comparing to wider modeled sources, there are significant differences between estimates of auto trips. These significant differences are not surprising given the fundamentally different purposes that each source has been developed for and the significant differences in definitions of in-scope trips. As such, comparisons to these other sources should be considered as being for information purposes only.

Limitations and caveats

This report has largely focused on high-level comparisons between sources – for example, with regards to trends and volumes for long- and shorter-distance movements. These high-level comparisons the StreetLight data appears to provide a reasonable basis on which to develop our base travel demand matrices.

However, at a detailed level – for example for individual zone to zone flows – the differences between sources can be significantly larger. This is inevitable since there is no definitive source for travel volumes/trends throughout California; rather, all sources are estimates only with inherent limitations.

Rail

Rail service in California

Amtrak service

Amtrak is the major intercity rail operator in California. There are three intra-California routes overseen by Caltrans and the Joint Powers Authorities. These routes are:

- **Capitol Corridor** serving Auburn–Sacramento–Emeryville (with Thruway service to San Francisco)–Oakland–San Jose.
- **Pacific Surfliner** serving San Luis Obispo–Santa Barbara–Los Angeles–San Diego; and
- **San Joaquins** serving San Francisco Bay Area or Sacramento–Stockton–Fresno–Bakersfield.

Amtrak also operates the following long-distance routes with stops in California:

- **California Zephyr** serving Chicago–Denver–Emeryville (San Francisco).
- **Coast Starlight** serving Seattle–Portland–Sacramento–Emeryville (San Francisco)–Los Angeles.
- **Southwest Chief** serving Chicago–Kansas City–Albuquerque–Flagstaff–Los Angeles.
- **Sunset Limited** serving New Orleans–San Antonio–Tucson–Phoenix–Palm Springs–Los Angeles; and
- **Texas Eagle** serving Chicago–St. Louis–Dallas–San Antonio–Palm Springs–Los Angeles.

Other rail services

Other rail services in the state range from regional commuter rail services to urban heavy rail systems and include:

- **ACE (Altamont Corridor Express):** commuter rail serving Stockton–San Jose.
- **BART (Bay Area Rapid Transit):** heavy rail serving the Bay Area.
- **Caltrain:** commuter rail serving San Francisco–San Jose–Gilroy.
- **Coaster:** commuter rail operated by NCTD serving San Diego–Oceanside.
- **LA Metro:** heavy rail serving the Los Angeles metro area.
- **Metrolink:** commuter rail serving the Los Angeles metro area with outer termini in Ventura, Lancaster, San Bernardino, Perris, Riverside, and Oceanside.
- **Muni Metro (San Francisco):** light rail serving San Francisco.
- **Sacramento RT Light Rail:** light rail serving Sacramento.
- **San Diego Trolley:** light rail serving San Diego County.
- **SMART (Sonoma-Marin Area Rail Transit):** commuter rail serving Santa Rosa–San Rafael.
- **Sprinter:** hybrid commuter / light rail operated by NCTD serving Escondido – Oceanside; and
- **VTA (Santa Clara Valley Transportation Authority) Light Rail:** light rail serving Santa Clara County, including San Jose and its suburbs.

In-scope rail services and stations

To account for all rail trips of intercity significance, in-scope services and stations to include all commuter/intercity rail services and stations, and a select set of Amtrak Thruway stations have been defined. This results in the following categorization of existing and future rail services:

- Existing services in scope (all stations):
  - State-supported services (Capitol Corridor, San Joaquins, Pacific Surfliner)

  - Amtrak long-distance trains (Coast Starlight, California Zephyr, Southwest Chief, Sunset Limited)
  - ACE
  - Caltrain
  - Coaster
  - Metrolink
  - SMART
- Future services in scope (all stations):
  - HSR Phases 1 and 2 (San Francisco to Anaheim, extensions to Sacramento and San Diego)
  - ACE extension (Stockton to Sacramento)
  - Antelope Valley Rail (Los Angeles to Palmdale)
  - Caltrain / Capitol Corridor extension (Gilroy to Salinas)
  - Coachella Valley Rail (Amtrak service from Los Angeles to Indio)
  - Cross Valley Rail (regional rail from Lemoore to Visalia via Hanford)
  - Dumbarton Rail (Newark/Fremont to Palo Alto)
  - Far north service (Sacramento to Oroville (via Marysville), Chico and Redding)
  - High Desert Corridor Rail (Victorville to Palmdale)
  - SMART extension (Santa Rosa to Cloverdale)
  - Valley Link (Fremont to Stockton)
  - Brightline West (high-speed rail from Las Vegas to Rancho Cucamonga)
- Services not in scope:
  - BART
  - LA Metro
  - Muni Metro
  - Sacramento RT Light Rail
  - San Diego Trolley
  - Sprinter
  - VTA Light Rail
- Services partially in scope:
  - Amtrak Thruway buses (in-scope stations below represent major population centers and tourist attractions that do not otherwise have rail service):
    - Eureka
    - Redding
    - Monterey
    - Santa Cruz
    - Yosemite National Park (single zone)
    - Lake Tahoe (single zone)
    - Paso Robles
    - Las Vegas/Victorville
    - Perris
    - Indio

Operator ridership data

A detailed base year (2018) ridership data from all in-scope operators was requested. The data received in response to this request ranged from station-to-station volumes to boarding and

alighting counts, to on-board survey data. The data obtained from each operator is described below:

- Amtrak (rail-only): unlinked station-to-station trip counts (does not include transfers) by day of week for FY 2018; selected Capitol Corridor on-board survey data—includes access and egress modes.
- Amtrak (Thruway-rail): FY 2018 estimates of volume by linked Thruway-rail itinerary.
- ACE: calendar year 2019 boardings and alightings by station/train/day.
- Caltrain: calendar year 2019 average boardings and alightings by train.
- Coaster: FY 2018 quarterly average weekday trips by station.
- Metrolink: 2018 on-board survey data—includes origin and destination stations, access and egress modes.
- SMART: 2018 on-board survey data—includes origin and destination stations.
- Sprinter: FY 2018 average weekday/Saturday/Sunday boardings and alightings by station.

Methodology

The following steps were followed to construct the matrix of existing rail demand to be used in the CRRM:

Step 1: Estimate initial station-to-station trip volumes from ridership data. Due to the differences in ridership data, the exact method varied by operator:

- For operators where station-to-station trip volumes were directly provided (Amtrak rail, Amtrak Thruway), we simply filtered the data to only include in-scope trips.
- For operators where on-board survey data with origin and destination information was provided (Metrolink, SMART), we aggregated the weighted totals of respondents reporting each origin-destination pair to obtain estimates of station-to-station volumes.
- For operators where only boarding/alighting information was provided (ACE, Coaster, Sprinter), we used station boardings/alightings as row and column control totals and used iterative proportional fitting to obtain reasonable estimates of station-to-station volumes.

Station-to-station trips were estimated separately for weekdays and weekends.[39]

Step 2: Subtract Thruway-linked rail legs from Amtrak rail-only trips. Since Amtrak Thruway trips always include a transfer to rail,[40] and the corresponding rail legs were also included in the Amtrak rail data, the rail portions of in-scope Thruway-rail trips were identified, aggregated by route, origin, and destination, and subtracted the resulting volumes from the corresponding Amtrak rail origin-destination volumes to avoid double-counting. Throughout the rest of this process, Thruway-rail trips are treated separately from rail-only trips.

[39] Only weekday trips were estimated for ACE, as no weekend service is provided.

[40] For the period of data provided – the rules of Thruway bus use have subsequently changed.

Step 3: Allocate transfers between rail services. For pairs of services that share stations or utilize closely located stations,[41] we used on-board survey information and professional judgement to estimate transfer rates between services and link legs from each service to build linked trips[42]. This process used the following three-step approach. This analysis was performed separately for weekday and weekend trips:

1. Identify valid transfer points and the pairs of services relevant to each.
2. For each possible transfer direction at each identified transfer point, estimate the share of boardings on the destination service that can be attributed to transfers from the origin service. For cases where on-board survey data is available for one or both of the operators, the transfer rate was calculated from the data.[43] For cases where no on-board survey data was available for either operator, it was assumed a transfer rate based on the types of services, direction of transfer, and transfer rates at similar transfer points for which we have on-board survey data for one or more of the operators.
3. Then allocated the corresponding destination service boardings to stations on the origin service according to the general distribution of boardings on that route. Exceptions were made for parallel routes; passengers transferring from northbound Coaster to the northbound Pacific Surfliner were assumed to continue northbound, for example.

Step 4: Assign rail and Thruway-rail trips to origin and destination CRRM zones. Once the final station-to-station trip volumes were estimated, both rail and Thruway-rail trips to the CRRM zone the station was located in were assigned. In practice, most rail trips would likely not have their true ODs in the same zone as the station, but as the observed rail matrix is only employed in the CRRM at the County level, it was considered that this was sufficiently robust to capture County-County level rail flows.

Step 5: Assign rail and Thruway-rail trips to time periods. Once rail and Thruway-rail trips were assigned to true origin and destination zones, weekday trips to the following four time periods were allocated:

- Weekday AM Peak (6–10 AM)
- Weekday Midday (10 AM–3 PM)
- Weekday PM Peak (3–7 PM)
- Weekday Off-Peak (7 PM–12 AM)

Due to differences in available data, the exact method varied by operator:

- For operators where boardings by train by station were provided (ACE, Caltrain), schedule data was obtained and matched each stop of each train to the scheduled departure time from that stop. Next, weekday trips were allocated for each origin-destination pair to time periods according to the temporal distribution of departure times from the boarding station.

[41] This includes intersecting Amtrak routes, as the raw Amtrak data represented unlinked trips.

[42] A maximum of one transfer per trip was assumed.

[43] Transfer rates typically ranged from 0.5% to 5.0%.

- For operators where on-board survey data with origin-destination and time of day information was provided (Metrolink, SMART), the weighted totals of respondents reporting trips occurring in each time period were aggregated for each origin-destination pair to obtain OD-specific time period shares. Then weekday trips were allocated for each origin-destination pair to time periods according to these shares.
- For operators where ridership could not be directly linked to time periods (Amtrak rail, Amtrak Thruway, Coaster, Sprinter), it was assumed that trips are distributed similarly to capacity. First, schedule data was obtained and used to determine for each station pair the share of departures that falls within each time period. The resulting period-specific factors were then multiplied by the corresponding volumes to obtain the volumes for each period.

A weekend (all day) time period was also included, with trip volumes coming directly from the weekend trip table output in Step 4.

Outputs

The process described above resulted in 129,800 average daily weekday in-scope rail trips and 47,100 average daily weekend in-scope rail trips, Overall, this represents 106,200 average annual daily in-scope trips. A map of in-scope weekday flows is presented in Figure 4-15 below. Wider lines represent larger flows.

Figure 4-15: Average weekday in-scope rail flow

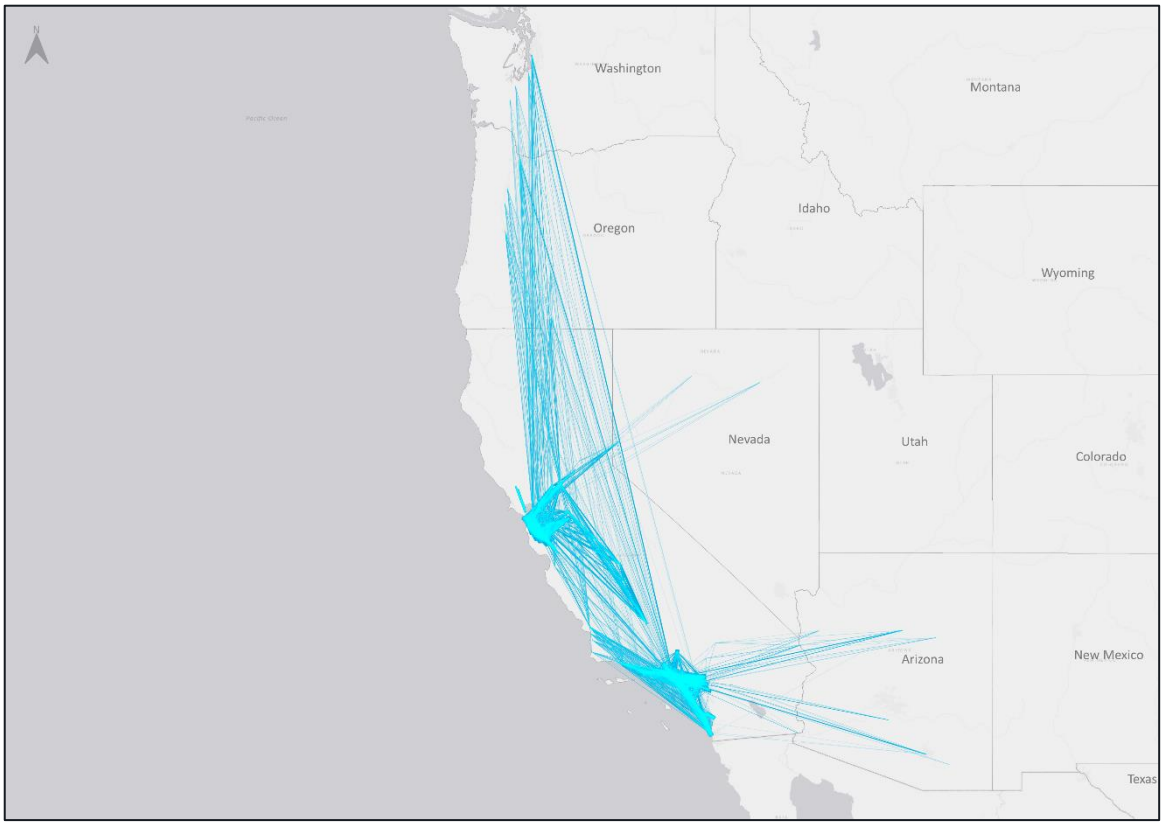


Table 4.10 and Table 4.11 present the top rail stations by in-scope daily passenger volume for commuter rail (includes ACE, Caltrain, Coaster, Metrolink and SMART) and intercity rail (Amtrak)

respectively. Note that some stations are served by both commuter and intercity service. Nine of the top ten commuter rail stations are served by Caltrain.

Table 4.10: Top rail stations by daily in-scope commuter rail passengers

Rank	Station	Average Weekday Passengers (each direction)	Average Weekend Passengers (each direction)	Average Daily Passengers (each direction)
1	San Francisco (Caltrain)	14,875	2,759	11,413
2	Los Angeles Union Station (Metrolink)	12,124	6,366	10,479
3	Palo Alto (Caltrain)	6,768	1,255	5,193
4	San Jose Diridon (ACE/Caltrain/Amtrak)	6,598	1,344	5,097
5	Mountain View (Caltrain)	4,555	845	3,495
6	Redwood City (Caltrain)	4,031	748	3,093
7	Sunnyvale (Caltrain)	3,471	644	2,663
8	Hillsdale (Caltrain)	2,626	487	2,015
9	Millbrae (Caltrain)	2,608	484	2,001
10	22nd Street (Caltrain)	2,330	432	1,788

Table 4.11: Top rail stations by daily in-scope intercity rail passengers

Rank	Station	Average Weekday Passengers (each direction)	Average Weekend Passengers (each direction)	Average Daily Passengers (each direction)
1	Los Angeles Union Station	1,801	1,913	1,833
2	Sacramento	1,546	1,121	1,424
3	San Diego	862	1,020	907
4	Emeryville	838	530	750
5	Bakersfield	555	643	580
6	Oakland Jack London Square	592	383	532
7	Davis	596	330	520
8	Fresno	486	593	516
9	Solana Beach	511	516	512
10	Irvine	471	615	512

Table 4.12 and Table 4.13 present the top commuter rail and intercity station pairs by passenger volume, respectively. Each of the top six commuter rail station pairs, and seven of the top ten commuter rail station pairs, includes the San Francisco Caltrain Station. Nine of the top ten intercity rail station pairs include either Los Angeles Union Station or Sacramento.

Table 4.12: Top in-scope station pairs by average daily commuter rail passenger volume

Rank	Origin	Destination	Average Weekday Passengers (each direction)	Average Weekend Passengers (each direction)	Average Daily Passengers (each direction)
1	San Francisco Caltrain	San Jose Diridon	2,016	374	1,547
2	Palo Alto	San Francisco Caltrain	1,936	359	1,485
3	Mountain View	San Francisco Caltrain	1,868	346	1,433
4	San Francisco Caltrain	Sunnyvale	1,492	277	1,145
5	Redwood City	San Francisco Caltrain	1,264	234	970
6	Hillsdale	San Francisco Caltrain	1,190	221	913
7	Palo Alto	San Jose Diridon	900	167	691
8	Fullerton	Los Angeles Union Station	844	277	682
9	Industry	Los Angeles Union Station	751	0	537
10	San Francisco Caltrain	San Mateo	641	119	492

Table 4.13: Top in-scope station pairs by average daily intercity rail passenger volume

Rank	Origin	Destination	Average Weekday Passengers (each direction)	Average Weekend Passengers (each direction)	Average Daily Passengers (each direction)
1	Los Angeles Union Station	San Diego Santa Fe Depot	384	404	389
2	Emeryville	Sacramento	306	183	271
3	Los Angeles Union Station	San Diego Old Town	210	214	211
4	Los Angeles Union Station	Solana Beach	210	181	202
5	Richmond	Sacramento	195	110	171
6	Oakland Jack London Square	Sacramento	172	114	156
7	Los Angeles Union Station	Oceanside	149	146	148
8	Martinez	Sacramento	150	114	140
9	Irvine	Los Angeles Union Station	121	150	130
10	Bakersfield	Fresno	125	141	130

Air

This section sets out the methodology used to develop matrices of existing commercial air demand and presents a summary of the airport-to-airport trip tables.

In-scope airports

Air demand in the CRRM is focused on a set of commercial airports that covers nearly all passenger activity. The following table identifies 23 airports that comprised over 99% of 2018 enplanements in California and were included in our analysis. All other airports in California were excluded due to their low passenger volumes and consequently low potential impact on any current or future rail project.

Table 4.14: Top California airports by enplanements, 2018

Code	Airport	City	Enplanements
LAX	Los Angeles International Airport	Los Angeles	42,624,050
SFO	San Francisco International Airport	San Francisco	27,790,717
SAN	San Diego International Airport	San Diego	12,174,224
TIJ	Tijuana International Airport[44]	Tijuana/San Diego	7,823,744
SJC	San Jose International Airport	San Jose	7,032,851
OAK	Oakland International Airport	Oakland	6,686,603
SMF	Sacramento International Airport	Sacramento	5,907,629
SNA	John Wayne International Airport	Santa Ana	5,201,642
BUR	Bob Hope Airport	Burbank	2,680,240
ONT	Los Angeles/Ontario International Airport	Ontario	2,498,993
LGB	Long Beach Airport	Long Beach	1,908,635
PSP	Palm Springs International Airport	Palm Springs	1,163,883
FAT	Fresno Yosemite International Airport	Fresno	853,538
SBA	Santa Barbara Airport	Santa Barbara	403,745
SBP	San Luis Obispo Airport	San Luis Obispo	235,570
STS	Sonoma County Airport	Santa Rosa	217,480
MRY	Monterey Airport	Monterey	186,806
BFL	Meadows Field	Bakersfield	105,104
SCK	Stockton Metropolitan Airport	Stockton	98,908
ACV	Arcata Airport	Arcata/Eureka	69,575
RDD	Redding Municipal Airport	Redding	42,775
MMH	Mammoth Yosemite Airport	Mammoth Lakes	23,522
SMX	Santa Maria Public Airport	Santa Maria	23,008

Source: FAA Air Carrier Activity Information System Passenger Boarding Data

The following four non-California airports were also considered in-scope, due to their significant passenger volumes and proximity to California. Trips between these airports and airports in Table

[44] Note that while Tijuana International Airport was considered in-scope since it serves some travelers from the San Diego area, there is currently no scheduled service between Tijuana International Airport and any airport in California.

4.14 were considered in-scope, while trips exclusively between non-California airports (for example, LAS-PHX) were not.

Table 4.15: In-scope airports outside California by enplanements, 2018

Code	Airport	City	Enplanements
LAS	McCarran International Airport	Las Vegas, NV	23,795,012
PHX	Phoenix Sky Harbor International Airport	Phoenix, AZ	21,622,580
RNO	Reno/Tahoe International Airport	Reno, NV	2,048,916
MFR	Rogue Valley International Airport	Medford, OR	492,217

Source: FAA Air Carrier Activity Information System Passenger Boarding Data

Air traffic data

Two primary sources of detailed air traffic data exist, both of which are provided through the Office of Airline Information of the Bureau of Transportation Statistics (BTS). Data was collected from these two sources for all in-scope airports identified above. Data for calendar year 2018 was used, as it was the most recent full year available when this process began.

Airline origin and destination survey (DB1B)

The DB1B database contains air volumes and fares by itinerary and is based on a 10% sample of airline tickets. Due to its itinerary-level basis, DB1B data can be aggregated over itineraries sharing the same origin and destination airport to obtain estimates of true OD volumes. International DB1B data does exist, but was not used in this effort, due to its limited usefulness and the significant effort required to obtain it. Some small commercial carriers are not required to report DB1B data, as reporting requirements for DB1B are less stringent than those for T-100 (described below).

Air carrier statistics (T-100)

T-100 contains flight segment-level data, including scheduled and actual operations, scheduled and operated seats, passengers, and airtimes for all commercial air carriers that operate flights within the United States. This data is useful for determining how many passengers traveled on flight segments between specific airports but provides no information on true passenger origins and destinations. T-100 is also essential for determining level-of-service characteristics such as travel time and frequency, and in estimating volumes for carriers who do not meet the DB1B reporting threshold.

Additional data

StreetLight LBS data

Location-Based Services (LBS) cell-phone origin-destination trip data was procured from StreetLight and used to inform allocation of airport trips to CRRM model zones, as discussed below. The specific data used for this process was a matrix of relative short-distance zone-to-zone trip volumes throughout the model area.

Ground access survey data

Data from ground access surveys at LAX, SFO and OAK were used to adjust the distribution of access and egress trips to/from these three airports at the county level, as described below.

Official Airline Guide (OAG) schedule data

Airline schedule data from OAG was used to allocate trips on an airport-to-airport basis to time periods as described below.

Methodology

The following steps were followed to construct the matrix of existing air demand to be used in the CRRM:

Step 1: Aggregate DB1B passenger volumes by origin-destination pair. For each in-scope airport pair, DB1B passenger volumes were aggregated over all itineraries. Since DB1B represents a 10% sample of total tickets, the resulting values were multiplied by 10.

Step 2: Adjust total trips to account for small carriers. Since reporting thresholds differ between DB1B and T-100, an adjustment was made to account for trips that are included in T-100 but not DB1B. First, it was necessary to determine which small carriers operate in in-scope markets but do not report DB1B data. Once these carriers were identified, the corresponding in-scope passenger volumes from T-100 data were determined, scaled down based on professional judgment to remove connecting passengers[45], and added them to the DB1B volumes computed in Step 1.[46] This step yields final airport-to-airport trip volumes.

Step 3: Assign airport trips to origin and destination CRRM zones. Once the final airport-to-airport trip volumes were obtained, these trips were assigned to origin and destination CRRM zones based on the LBS data.

First, the estimated share of access/egress trips to/from each airport that begin/end in each other zone by assuming airport access/egress trips is distributed similarly to all trips to/from the zone containing the airport. Next, distributed trips for each airport pair to all corresponding sets of origin and destination zones. During this process, it was assumed that origin and destination shares are independent (i.e., the share of trips allocated to an individual destination zone only depends on the destination airport and does not depend on the origin airport or zone). All origin-destination combinations were enumerated, and then for each combination, the OD volume was calculated as the airport pair volume multiplied by the share of origin airport access trips allocated

[45] In this adjustment, we assumed 25% of total traffic is local when either the origin or destination airport is LAX, SFO or LAS (major hubs where we would expect the majority of travelers from small airports to connect to other flights), and that 50% of total traffic is local when either the origin or destination airport is BUR, OAK, SJC, SNA or SMF (large non-hub airports where we would still expect a significant share of travelers from small airports to connect to other flights).

[46] Note that a similar adjustment would have been needed to account for air trips between Tijuana and California airports, since only domestic DB1B data was available. However, there is currently no such scheduled commercial service, rendering this adjustment unnecessary.

to the true origin zone multiplied by the share of destination airport egress trips allocated to the true destination zone.

Once this allocation was performed, any trips that did not meet the following criteria were removed and scaled up the remaining trips on an airport-to-airport basis to account for trips that were removed:

- The sum of access plus egress distance cannot be more than 50% of airport-to-airport distance.
- The distance from the origin zone centroid to the destination zone centroid must be at least 75% of airport-to-airport distance.

Once trips were allocated to CRRM zones, county-level shares of access trips to and egress trips from LAX, SFO and OAK were scaled to match county-level shares from recent ground access surveys while maintaining the distribution of origins/destinations within each county.

Step 4: Assign trips to time periods. Once air trips were assigned to true origin and destination zones, were allocated to the following five time periods:

- Weekday AM Peak (6–10 AM)
- Weekday Midday (10 AM–3 PM)
- Weekday PM Peak (3–7 PM)
- Weekday Off-Peak (7 PM–6 AM)
- Weekend (all day)

This allocation assumed that trips are distributed similarly to capacity. First, OAG schedule data was used to determine for each airport pair the share of seats that falls within each time period.[47] The resulting period-specific factors were then multiplied by the corresponding volumes to obtain the volumes for each period. While assuming a constant load factor across periods, this process was set up with the ability to vary the load factor by period.

[47] For airport pairs requiring a connection, the shares of seats in each period were calculated based on all in-scope seats departing the origin airport.

Outputs

The process described above resulted in approximately 89,000 daily in-scope air trips. Of these, nearly 50,000 are between in-scope California airports. The adjustment for small carriers in Step 2 accounted for approximately 100 of these daily trips. A map of these in-scope flows is presented in Figure 4-16 below. Wider lines represent larger flows.

Figure 4-16: Average daily in-scope air flows

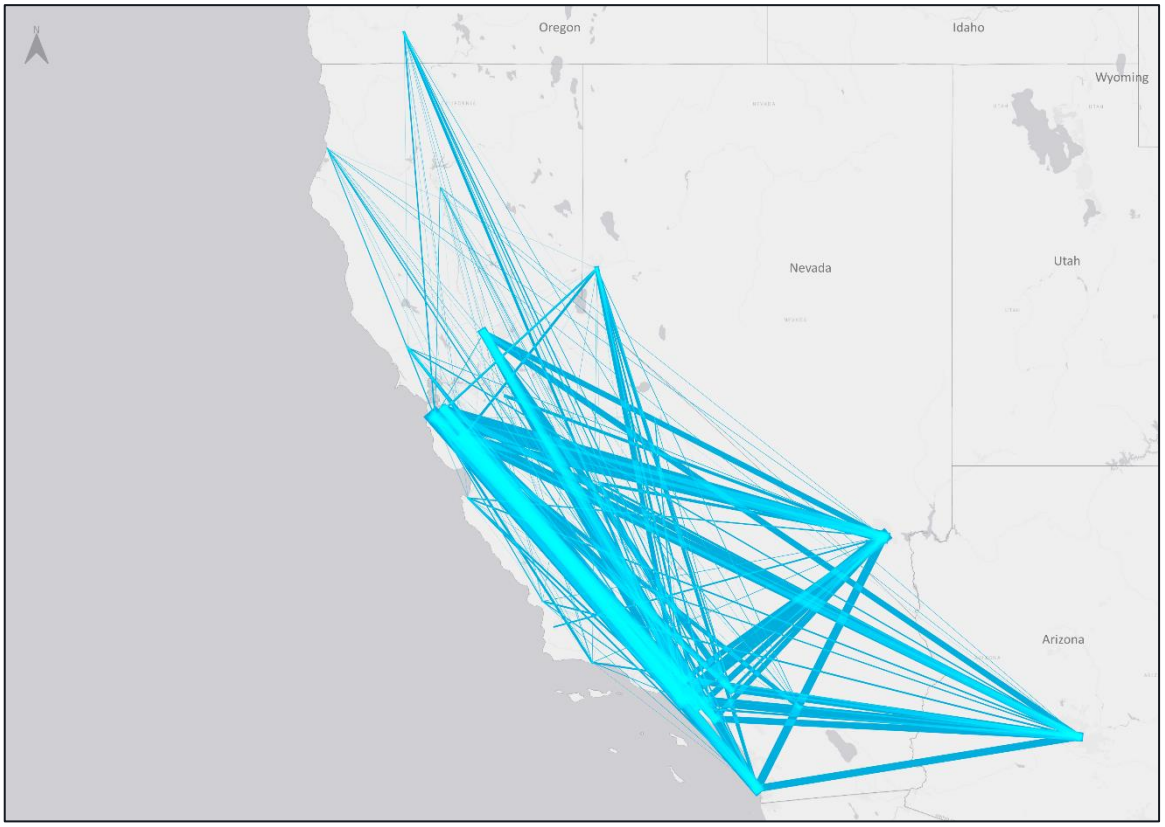


Table 4.16 presents the top airports by in-scope daily passenger volume. The top airports are generally consistent with their relative enplanements in Table 4.14 and Table 4.15, except for PHX, which ranks sixth in daily in-scope passengers, despite having the fourth most enplanements of the in-scope airports. This is due to a smaller share of PHX trips being in-scope (i.e., destined for California airports) compared with the other large in-scope airports.

Table 4.16: Top airports by daily in-scope passengers

Rank	Airport	Daily Passengers (each direction)
1	LAX	12,036
2	SFO	11,190
3	LAS	10,727
4	SAN	8,438
5	OAK	7,469
6	PHX	7,241
7	SJC	7,065
8	SMF	6,193
9	SNA	4,878
10	BUR	4,391

Table 4.17 presents the top airport pairs by passenger volume. Each of the top ten airport pairs includes at least one of the five highest volume airports.

Table 4.17: Top in-scope airport pairs by average daily passenger volume

Rank	Origin	Destination	Daily Passengers (each direction)
1	LAX	SFO	3,734
2	LAS	LAX	2,418
3	SAN	SFO	1,935
4	LAS	SFO	1,872
5	SAN	SJC	1,381
6	LAX	PHX	1,345
7	SAN	SMF	1,332
8	LAX	OAK	1,299
9	LAS	OAK	1,242
10	LAX	SJC	1,229

Intercity bus

This section sets out the methodology used to develop matrices of existing intercity bus demand for the California Rail Ridership model (CRRM) and presents a summary of the city-to-city bus trip tables.

Supply side analysis

Existing intercity bus operators are typically private firms who do not make public any information on ridership. As such, it is necessary for us to estimate intercity bus use through analysis of the supply side data that is publicly available, and reasonable estimates of typical loadings.

The tasks described in this section define the process to summarize the supply of intercity bus service across California; this includes determination of how many buses and seats serve each origin and destination every day. The supply side analysis forms the basis for the demand side analysis; based on a combination of load factor and city-pair size.

Supply side data

We have collected data from aggregator websites such as busbud.com or checkmybus.com, individual operator websites such as Greyhound, and by reviewing public operator timetables.

Main operator data included Greyhound, Megabus and FlixBus. Smaller operators were included (such as LAX FlyAway) when considered important intercity bus connections within California.

Data collected for each operator included:

- **Service departure times by day of week (weekday, Saturday, and Sunday) for every stop along the route.** Since most services do not operate on a specific headway, all departure times were recorded.
- **Stop locations.** For each stop/service, location characteristics (i.e., downtown, suburbs, out of town etc.) were flagged. This helped to identify where stop locations differ between operators within the same city.
- **Average fare information for each city-to-city pair.** The range of fares is listed on the aggregator and operator websites. Due to the ongoing COVID-19 situation, advance fare information was collected for Monday, September 7th, 2020 (Greyhound/FlixBus) and Monday, July 1st, 2020 (FlixBus) to avoid collecting data which may include discounted fares to attract ridership.[48]
- **Available seats by operator.** The available seats for each operator determine bus capacity:
  - Greyhound buses typically have 50-55 seats. A value of 50 was used for this analysis.
  - Megabus uses higher capacity 72-seat coaches rather than standard vehicles.
  - FlixBus buses have 52 seats.
  - For other buses, 50 seats were assumed.

An average fare for each route (city-to-city) was calculated based on advance fare operator data (based on the dates described above). The operator data was averaged to get a unique fare for each city pair.

[48] Data was collected in March 2020.

Below are further notes specific to each bus operator:

Greyhound

The Greyhound route network in California serves more than 70 cities and towns along 14 distinct routes. Service is available between most combinations of cities although many city-to-city points involve a transfer. To estimate demand, a maximum of one transfer was considered. Trips involving two or more transfers are not considered viable trips for Intercity bus.

Fare data was collected, and demand data calculated, for most city-to-city combinations along each of the 14 lines. This information was analyzed, and a similar profile was applied to the remaining city pairs (where data was not directly collected) based on a per-mile fare. An estimation was then made for the fares and demand on routes involving a transfer by using the per-mile fare estimation. Checks were made on various routes involving a transfer to ensure the estimation technique was accurate, before applying to remaining city to city pairs.

Information collected for each of the 14 Greyhound lines allowed an estimate to be made of the maximum total seats available for those routes. For example, Los Angeles to San Francisco has a maximum of 400 seats available on a weekday, therefore Los Angeles to Oakland (which is on the same route) also has a maximum of 400 seats available. This exercise was replicated for all other routes where information was not directly collected.

Megabus, FlixBus, and smaller operators

In California, Megabus operates almost exclusively from either commuter rail stations or transfer stations for local transit buses. Megabus operates on 4 main routes with data available from the aggregator websites:

- Los Angeles–San Jose–San Francisco.
- Los Angeles–Oakland–San Francisco.
- San Francisco–Sacramento; and
- Los Angeles–Riverside–Las Vegas.

FlixBus offers service from about 30 locations across California. However, service is only available between certain city pairs. FlixBus data is available from the aggregator websites as well.

LAX FlyAway is a dedicated airport bus service connecting LAX to various points in Los Angeles County. This service provides a strategic connection between the airport and the rail system at LA Union Station. LAX FlyAway bus service to Union Station was included in our trip tables.

There are a few smaller operators which serve specific long-distance markets within California and neighboring states. They include **CoachRun**, **Las Vegas Express** and **Tufesa**. However, these are not considered major markets. **Lux Bus** is another charter bus service that was not included in our analysis.

Aggregation of operator data

The bus operator data was aggregated to generate a consistent dataset that includes the following:

- The number of daily services available between all city pairs (e.g., Greyhound has 8 daily weekday services from Los Angeles to San Francisco).
- The maximum number of daily seats available between all city pairs, (e.g., Greyhound has 400 daily weekday seats available from Los Angeles to San Francisco).
- Average fare for all city-to-city pairs.
- Distance for all city-to-city pairs.
- City-to-city supply, subject to the route-specific capacity. A route-level capacity analysis for city pairs to be used in an "available seats" analysis was conducted. This step noted the maximum number of seats available between city pairs given the overall route-level capacity. For example, the Greyhound route from Oxnard to San Francisco (250 weekday seats) operates on the Los Angeles to San Francisco service (400 weekday seats) with a high element of "commonality." In practice, however, each of these cannot be fully filled, since if 300 of the seats from Los Angeles to San Francisco are occupied, say, this only leaves a maximum of 100 left for people traveling from Oxnard to San Francisco. This step specified a service limit (that cannot be exceeded within the demand calculations) to ensure that the outputs make sense in this context.

The resulting dataset provides an overview of intercity bus supply across California as the total city-to-city bus capacity/seats available at any given time of day/day of week.

With regards to separate allocation to zones within the CRRM, note the following:

- Different operators may use different stations within the same city. In such instances (where stations are close by, such as a few blocks apart), all stations were allocated to the same city.
- Where bus stops are located in different cities but within the same metro area (e.g., Greyhound stops in Oakland, but Megabus stops in Berkeley) data collected were separated for those separate cities if they appear as separate zones within the aggregated CRRM.

Demand estimation

One key input to the demand estimation is the load factor assumed on the buses. These may ultimately vary by time of day (e.g., higher load factor in peak than off-peak), and by operator (higher load factor for smaller / less frequent services). In the absence of any observed load factor data, a load factor of 70% has been assumed for all routes except for LAX FlyAway, where a value of 50% has been assumed. These values are based on our professional judgement but could be refined through station surveys, should additional data be made available.

A combination of population factors and the maximum number of seats available for each city pair was used in conjunction with the load factors to estimate the proportion of demand on a city-pair basis. The following steps were used in this process:

Step 1: Take the populations of all respective cities and calculate a population factor for each city pair (based on City A and City B population).

Step 2: Adjust the population factor based on the maximum number of seats available for each city pair (so cities which are not served by as many buses are considered).

Step 3. Normalize the factor calculated in Step 2 to obtain an estimate of the maximum seats available between all city pairs. Note that the data collection informed the theoretical maximum

of seats available between any two city pairs; however, in reality all those seats will not be taken up by one city to city pair (but a wide range of city pairs which collectively sum to the route demand at any given point). The normalized factor in Step 3 seeks to account for this.

Step 4: Calculate a reasonable seat capacity based on the above factor (e.g., max seats for city pair x Step 3 factor).

Step 5: Apply a load factor to the Step 4 value to obtain an estimate of demand for all city pairs.

Allocation of trips to CRRM zones

Once the final city-to-city trip volumes were estimated, we assigned these trips to origin and destination CRRM zones based on StreetLight Location-Based Services (LBS) cell-phone origin-destination trip data. The specific data used for this process was a matrix of relative short-distance zone-to-zone trip volumes throughout the model area.

We first estimated the share of access/egress trips to/from each bus stop zone that begin/end in each other zone by assuming bus stop access/egress trips is distributed similarly to all trips to/from the zone containing the station. We then distributed trips for each city pair to all corresponding sets of origin and destination zones. During this process, it was assumed that origin and destination shares are independent (i.e., the share of trips allocated to an individual destination zone only depends on the destination bus stop and does not depend on the origin bus stop or zone). All origin-destination combinations were enumerated, then for each combination, the OD volume was calculated as the city pair volume multiplied by the share of origin bus stop access trips allocated to the true origin zone multiplied by the share of destination bus stop egress trips allocated to the true destination zone.

Once this allocation was performed, we removed any trips that did not meet the following criteria and scaled up the remaining trips on a city-to-city basis to account for trips that were removed:

- The sum of access plus egress distance cannot be more than 50% of stop-to-stop distance.
- The distance from the origin zone centroid to the destination zone centroid must be at least 75% of stop-to-stop distance.

Allocation of trips to time periods

Once intercity bus trips were assigned to true origin and destination zones, we allocated trips to the following five time periods:

- Weekday AM Peak (6–10 AM)
- Weekday Midday (10 AM–3 PM)
- Weekday PM Peak (3–7 PM)
- Weekday Off-Peak (7 PM–6 AM)
- Weekend (all day)

This allocation assumed that trips are distributed similarly to capacity. First, stop-level schedule data was used to determine for each origin-destination pair the share of seats by departure time

from the boarding stop that falls within each time period.[49] The resulting period-specific factors were then multiplied by the corresponding volumes to obtain the volumes for each period. Since the weekend period is intended to represent the average weekend day and not the full average weekend, weekend volumes were divided by two. While we assume a constant load factor across periods, this process was set up with the ability to vary the load factor by period.

Outputs

The process described above resulted in approximately 9,000 daily intercity bus trips. A map of these in-scope flows is presented in Figure 4-17 below. Wider lines represent larger flows.

Figure 4-17: Average daily in-scope intercity bus flows

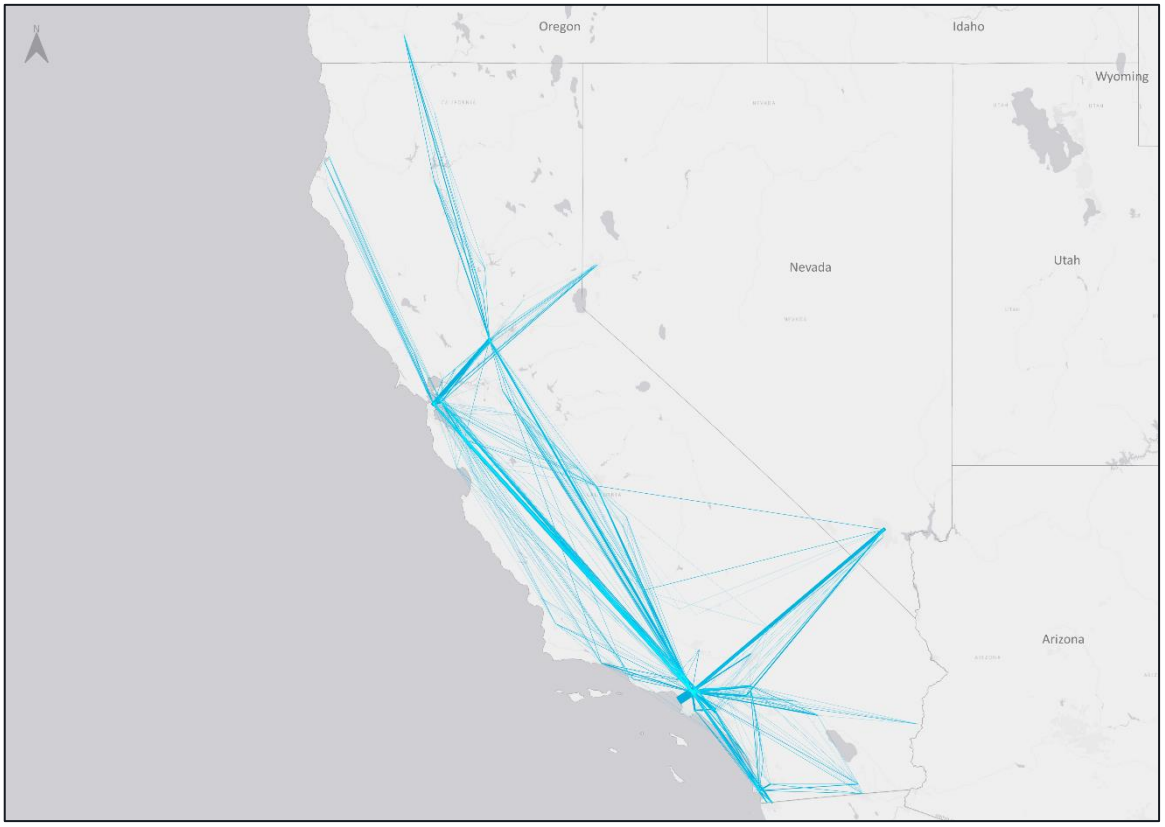


Figure 4.18 presents the top intercity bus destinations by in-scope daily passenger volume. The top destinations are generally located in the largest metro areas which follows logically since demand levels are population driven. Note that LAX FlyAway service contributes 1,150 trips to both the Los Angeles and LAX locations.

[49] For trips spanning multiple time periods, the period that included the scheduled boarding stop departure time was used.

Figure 4.18: Top bus destinations by average daily in-scope passengers

Rank	Location	Daily Passengers (each direction)
1	Los Angeles	2,973
2	LAX	1,150
3	San Francisco	681
4	Sacramento	515
5	Anaheim	344
6	Oakland	318
7	Las Vegas	288
8	San Diego	238
9	San Jose	229
10	San Bernardino	194

Table 4.19 presents a summary of the top ten origin-destination demand pairs by daily passenger volume, with values representing daily demand in each direction.

Table 4.19: Top in-scope origin-destination pairs by average daily passenger volume

Rank	Origin	Destination	Daily Passengers (each direction)
1	Los Angeles	LAX	1,150
2	Sacramento	San Francisco	311
3	Las Vegas	Los Angeles	226
4	Anaheim	Los Angeles	211
5	Los Angeles	San Bernardino	113
6	Oakland	San Francisco	106
7	Los Angeles	Riverside	96
8	Los Angeles	San Francisco	95
9	Burbank	Los Angeles	93
10	Los Angeles	San Diego	91

The largest demand is between Los Angeles and Los Angeles International Airport (LAX). As with the top destinations, most of the top pairs involve the largest cities (e.g., Los Angeles, San Francisco, Oakland, San Diego, etc.). Eight of the top ten pairs include Los Angeles, and six of the top ten are entirely within Southern California.

5 Population and Employment Data

Introduction

One of the key objectives of the California Rail Ridership Model (CRRM) is to capture the heterogenous transportation behavior of different individuals. To accomplish this, different segments of the model rely on having individual's socioeconomic variables as inputs. This section discusses what the requirements of those inputs are, and the approach taken to create those inputs and focuses on intra-California resident trips, the core trip making element within the model.

Non-resident and external trips are also included in the CRRM, using observed demand and application of the choice model only (no generation or distribution); growth in these is set out at the end of this Chapter.

Requirements

The CRRM population inputs need to be segmented by the key variables that are used in different steps of the model. These include the generation of trips and the mode choice for the trips, including access and egress. Table 5.1 shows each of these components and the socioeconomic variables that are being used. Note that employment is included in the trip distribution stage of the modeling process.

Table 5.1: Model components and needed socioeconomic variables.

Model component	Household size	Household income	Employment status
Trip generation	x	x	
Mode choice		x	x

To this end, the final model inputs are the number of individuals by:

1. **Modeling zone.** The 1186 model zones, 1169 of which are in California.
2. **Household size.** Number of individuals in a household, segmented as:
    i. 1 person
    ii. 2 people
    iii. 3 people
    iv. 4 people
    v. 5+ people
3. **Household income**. The household income, segmented as:
    i. Low income, defined as households making less than $50,000;

ii. Middle income, defined as households making between $50,000 and $100,000; and
iii. High income, defined as household making more than $100,000.

4. **Employment status.** The status of each individual, segmented as:

i. Employed
ii. Retired
iii. Homemaker
iv. Student
v. Other

Approach

The population inputs for the CRRM take advantage of the synthetic population that was already developed for the California Statewide Transportation Demand Model (CSTDM). This synthetic population, which was developed using Public Use Microdata Sample (PUMS) inputs from 2015 and scaled to the 2015 American Community Survey (ACS), was updated for use in the CRRM. On the other hand, the CRRM is designed with a base year of 2018. To update the synthetic population, a reweighting process was undergone. This involved creating new weights such that the aggregate population matches updated totals at the county level.

The following steps were taken for creating the CRRM population inputs from the CSTDM synthetic population:

- Create the employment status column from PUMS data;
- Aggregate to county-level totals for household size, household income, and employment level;
- Develop target totals for each column at the county level;
- Reweigh the aggregate segments to meet aggregate totals for each county; and
- Apply the county-level reweighting factors to the segmented total for each zone and reweight for the population of each zone.

Processing the CSTDM synthetic population

This section discusses the format the CSTDM synthetic population was given in and how it was manipulated to be used for the CRRM inputs.

Existing CSTDM synthetic population

The CSTDM synthetic population was created using the 2015 PUMS sample and ACS control totals for 2015. It was provided to Steer as a database of PUMS records and weights for each record corresponding to the number of times that record is being represented in the overall synthetic population.

An evaluation of the CSTDM population totals against the existing demographic data for 2015 was performed in 2020.[50]

[50] *Socioeconomic Data and Growth Memo*, shared with Deutsche Bahn Engineering & Consulting (DB), Feb 2021

Existing columns[51]

The CRRM requires information on household income, household size, and employment status. The PUMS records, which the CSTDM synthetic population is built on top of, has columns for both the household size (NP) and household income (HINCP). Employment status, on the other hand, had to be derived.

Consideration of individuals who may fall under multiple categories.

While the household size and household income were directly used from PUMS, employment status was derived using multiple columns. This approach meant that some individuals could have been categorized as more than one column. Examples include:

- A college student who is also employed.
- A retired individual taking community college classes.
- An employed individual in their 30s that is also receiving retirement income.

To categorize these individuals, Steer has made a few assumptions to assign them into the categories that we believed to be most likely to be linked to the behavior in the choice model. Whenever an assumption as such is made, it is noted in the data manipulation process below.

Column manipulation

For each of the columns needed for the model, the PUMS columns were aggregated for reweighting. Throughout this section, the relevant PUMS columns are mentioned in parenthesis for reference.

Household size (NP)

The household size in PUMS is provided as a number between zero and 40. Households with zero individuals (e.g., a vacant unit) were excluded. Households with at least one person were aggregated into the numbers used for the model:

- 1 person
- 2 people
- 3 people
- 4 people
- 5+ people

Household income (HINCP)

The household income is provided as an actual value, including no income (0) and losses (i.e., values below zero). While the income in the model is divided into three segments, the reweighting process divided the income into the 10 segments available in the ACS:

- Less than $10,000
- $10,000 - $14,999
- $15,000 - $24,999

[51] **https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2015-2019.pdf**

- $25,000 - $34,999
- $35,000 - $49,999
- $50,000 - $74,999
- $75,000 - $99,999
- $100,000 - $149,999
- $150,000 - $199,999
- $200,000 or more

The decision to use more segments was made to capture more variation before aggregating up to the three segments in the model.

Employment status

While PUMS definition of household size and income are consistent with model use, the dataset does not have a column that corresponds to what is being referred to as the *employment status*. This status is meant to capture a primary driver in behavior. Therefore, the PUMS data needed to be processed at the individual and household level to create a new column, which could then be aggregated using CSTDM synthetic population weights. During the processing, every PUMS record was labeled as one of the following:

- **Employed.** Individuals who are in the labor force and employed.
- **Retired.** Individuals who are primarily retired.
- **Homemaker.** Individuals who are not in the labor force, likely partaking in 'homemaking' activities.
- **Student.** Individuals who are enrolled and study as their primary activity.
- **Other.** Individuals who do not fall under any of the above categories.
- **Under five years old.** Individuals who are less than five years old, who are excluded from the CRRM model. They are included here to match the correct population totals.

Each individual PUMS record is assigned a category sequentially, such that, once a record is assigned a category, it is no longer available for assignment. Generally, the assignment happens in the following order:

1. Under five years old
2. Student (up until college)
3. Employed
4. Student (graduate students and professional school)
5. Retired
6. Homemaker
7. Other

The following sections discuss each of these.

1. Label Under five years old

Based on the age column (AGEP), PUMS records of individuals under five years old are labelled as *Under five years old*.

2. Students (up until college)

Anyone between five and 16 years old is assigned as *Student*.

Based on the PUMS column that indicates the grade level that an individual is attending (SCHG), individuals between 16 and 25 that are attending school, but not *Graduate or professional school beyond a bachelor's degree* are also considered students. 25 years old is used as the upper range to capture students who take longer to complete their schooling.

PUMS also recognizes *Graduate or professional school beyond a bachelor's degree*, however these are more likely than college students to be individuals that are employed full time, retired or homemakers. Therefore, they are assigned later in the process.

3. Employed

Based on the recoded employment column in PUMS (ESR), individuals who are employed as civilians or in the army are assigned as employed. This includes those who have jobs, but are not currently at their jobs (e.g., taking unpaid leave).

All individuals between 16 and 25 who are employed, but not in grade school or in college, are labeled as *Employed*. This is capturing the assumption that college students with a side job are more likely to act as students than as employed individuals.

Individuals over 65 that are employed are labeled as *Employed*.

4. Remaining students (graduate students and professional school)

Of the remaining individuals who are categorized as *Graduate or professional school beyond a bachelor's degree* (i.e., not labeled as *Employed*), the lower 50% of those individuals by age are assigned as *Student*. The lower 50th percentile, which corresponds to individuals who are 31 years old, was chosen by Steer as a reasonable age cut off for classifying an individual solely as a student, and therefore having student behavior in the model. The remaining, older half are left unassigned such that they can be labeled as *Homemaker*, *Retired*, and *Other*.

5. Retired

Based on the age column, any individual over 65 years old that is not employed is considered *Retired*. This includes individuals who are *Unemployed* and *Not in the labor force*.

Individuals between 55 and 65 who are labeled as *Unemployed* or *Not in the labor force*, but who are receiving a non-zero retirement income—column RETP in PUMS—are labeled as *Retired*. This captures individuals who retire before the 65-year-old cutoff. The age of 55 captures 75% of the remaining individuals in the sample who receive retirement income. Note that the sample also includes dependents who may be receiving an individual's retirement benefits (e.g., spouses, children). This cutoff avoids labeling individuals who would be more appropriately labeled as *Homemaker* or *Other*.

6. Unemployed

Individuals that are *Unemployed* are labeled as *Other*.

7. Homemakers

Of the remaining individuals, those who are labeled as *Not in the labor force* are considered eligible to be labeled as *Homemaker*. The label was then decided based on household structures. For this purpose, the base assumption was that each household only had one homemaker.

The family type and employment status column (FES) was used for determining households with homemakers. The following structures are available in PUMS:

1. Married-couple family: Husband and wife in [labor force (LF)][52];
2. Married-couple family: Husband in LF, wife not in LF;
3. Married-couple family: Husband not in LF, wife in LF;
4. Married-couple family: Neither husband nor wife in LF;
5. Other family: Male householder, no wife present, in LF;
6. Other family: Male householder, no wife present, not in LF;
7. Other family: Female householder, no husband present, in LF;
8. Other family: Female householder, no husband present, not in LF; and
9. N/A[53].

The process for determining who is the *Homemaker* is summarized in Table 5.2 and is described in detail below. As noted, before, this already excludes students and retired individuals.

Table 5.2: Labeling *Homemaker* by family structure

Family structure categories	Assignment of *Homemaker*
2 and 3	One person is the *Homemaker*; assigned based on the relationship to reference person column (RELSHIPP) and the sex of the individual (SEX), such that the correct spouse is assigned as the homemaker
4	Two *Homemakers*
6 and 8	One person assigned as *Homemaker*; assigned based on age and sex
1,5,7, and 9	No *Homemaker*'s assigned

For households where the family structure includes a spouse in the labor force and a spouse not in the labor force (i.e., 2 and 3), one individual was assigned as the homemaker. When the husband is in the labor force and the wife is not, the individual in the household that has the relationship to the reference person (RELSHIPP) as a *Husband/wife/spouse* and has the sex (SEX) labeled as *Female* is labeled as a *Homemaker*. Similarly, when the wife is in the labor force and the husband is not, the individual who is a *Husband/wife/spouse* and *Male* is labeled as a *Homemaker*.

For households under the *Other family* categories, when the female or male householder was not in the labor force and was the sole individual in the household not in the labor force, they were labeled as a *Homemaker*.

In these same households, when there were multiple females or males in the household after filtering for individuals over 20 years old, only one individual was labeled as *Homemaker*. The

[52] As a check, it was confirmed that both spouses in these family structures were labeled as *Employed*.

[53] Includes same-sex couples.

individual labeled as *Homemaker* is the individual closest to 46 years old. This age was chosen as the midpoint between the typical age individuals would end college and the age when individuals would retire.

8. Other

All individuals who were not assigned to any category are assigned as *Other*.

Results

After this processing, the distribution of each category for the PUMS records and for the weighted CSTDM synthetic population are presented in Table 5.3.

Table 5.3: Distribution of employment status for the PUMS records and the CSTDM synthetic population

Employment status	Percent of California PUMS records	Percent of CSTDM synthetic population
Employed	45.6%	43.0%
Retired	11.0%	11.5%
Homemaker	5.6%	4.5%
Student	24.1%	21.0%
Other	7.4%	13.1%
Under 5 years old	6.3%	6.9%

Rates for the aggregate totals

As evidenced above, the Employment column is not directly available in the Census Bureau datasets and requires a number of assumptions to categorize individuals. The *Homemaker* status depends on household-level analysis that requires disaggregate information. Similarly, classification as *Student* requires information on age and employment.

To produce comparable aggregate totals, the processing of the CSTDM data also produced rates at which:

- Each of the household structures were labeled as having a *Homemaker*, if any (e.g., households with only one adult in the labor force have a rate of 0).
- College students who are categorized as *Employed*.
- Graduate students were categorized as *Student*, *Employed*, or any of the other categories.

Homemaker

These rates were output for the following family structures by county:

- Married-couple family: Husband in LF, wife not in LF;
- Married-couple family: Husband not in LF, wife in LF;
- Other family: Male householder, no wife present, not in LF; and
- Other family: Female householder, no husband present, not in LF.

The following section explains how these rates were used in creating the target totals for *Homemaker*.

Student & Employed

Similarly, the rates for those who are *Graduate or professional school beyond a bachelor's degree* and what they are ultimately labeled as are output from the processing of the CSTDM synthetic population. These rates are used in creating the target totals for *Student* and *Employed*. The following section explains how these rates are used in processing the target totals.

Target totals for reweighting

This section describes how the target totals that are used for reweighting were developed. It includes specific ACS column references in parenthesis where relevant. All ACS columns are 5-year estimates for 2014-2018. target totals were developed as percentages of the total population by county.

Columns used and manipulation.

Population (B01003)

Population, while not explicitly a column, is being controlled for at every step of the reweighting process. More importantly, the population column is used to develop the target totals as percent of the county population.

Number of households by household size (B11016)

ACS column B11016 provides household size for family and non-family households. These were combined for the target totals. These were further combined to match the segmentation used in the CRRM and the aggregated CSTDM synthetic population:

- 1 person
- 2 people
- 3 people
- 4 people
- 5+ people

Number of households by household income (S1901)

The PUMS records were binned into the ACS columns, which were used directly:

- Less than $10,000
- $10,000 - $14,999
- $15,000 - $24,999
- $25,000 - $34,999
- $35,000 - $49,999
- $50,000 - $74,999
- $75,000 - $99,999
- $100,000 - $149,999
- $150,000 - $199,99
- $200,000 or more

Employment status

Similar to the PUMS classification, employment status was developed using a variety of columns.

Employed

The percentage of the county population labeled as *Employed* is created with the Local Area Unemployment Statistics (LAUS) from the California Employment Development Department[54] (CA EDD). An annual average is taken for 2018.

In addition, the rate of college students who are employed but are categorized as *Student* is accounted for. For each county, the rate is multiplied by the number of college students and subtracted from the total employment for the county.

The total population is used to create the percentages with the remaining number of employed individuals.

Retired

The number of retired individuals is retrieved from the Social Security Administration[55]. The *Retired workers* column is used. The total population is used to create the percentages.

Homemakers (S2302)

Similar to the PUMS labeling process, the assumption used is that there is at most one *Homemaker* per household.

For each family structure available, the rates developed in the previous step of processing the CSTDM synthetic population is used to determine how many of those families had a *Homemaker*. These rates vary by county and family structure. The total population is then used to create the percentages with the total *Homemakers* by county.

Students (B14001)

All individuals from kindergarten through college were categorized as *Students*.

For graduate students and those undertaking professional degrees, the rate developed throughout the CSTDM labeling process is used. The county-specific percentage of these individuals who are labeled as students is the same as the percentage in the labeled CSTDM synthetic population. This accounts for approximately 15% of the total graduate students. All other graduate students are classified as *Employed*, *Retired*, *Homemaker* or *Other*.

The total population is used to create the percentages with the total students.

Under five years old (S0101)

Individuals under five years old were categorized as *Under five years old*. The total population was used to determine the share of the population of each county that is *Under five years old*.

[54] **https://data.edd.ca.gov/Labor-Force-and-Unemployment-Rates/Local-Area-Unemployment-Statistics-LAUS-/e6gw-gvii**

[55] **https://www.ssa.gov/policy/docs/statcomps/oasdi_sc/2018/ca.html**

Other

To develop the *other* share for each county, the portion of the population identified as *Unemployed* in the LAUS dataset is combined with remaining population[56].

Results

Table 5.4 below shows the final state-wide distribution of these categories for the target totals.

Table 5.4: Distribution of employment status for the target totals

Employment status	Percent of California target totals
Employed	45.6%
Retired	11.0%
Homemaker	5.6%
Student	24.1%
Other	7.4%
Under 5 years old	6.3%

Output format

The target totals were developed for each of the relevant columns as percentages of each county population. In addition to this, the total population for each county was also provided.

These outputs are used directly as targets for the reweighting process, which is described in the next section.

Reweighting the population

To create the inputs for the CRRM model, the aggregate population from the CSTDM synthetic population is reweighted according to relevant targets. The following discusses the mechanism used for reweighting and how it was applied, how well the algorithm converged, and further manipulations to create the final inputs for the model.

Iterative proportional fitting

To weight the aggregated CSTDM synthetic population, iterative proportional fitting (IPF)[57] was used to generate new weights for each segment. IPF aims at weighing the population so that it most closely resembles the target data by creating weights that are closest to the real data. That is, each segment is weighed so that the new aggregate totals are more closely representative of the target totals. The IPF algorithm is run in iterations, where every iteration creates an average weight for each segment, after which the segments get adjusted for county-level population.

[56] Checks were done to ascertain that the remaining population, excluding *Unemployed*, was not negative. That is, that the other categories did not add up to more than the total population of the county.

[57] The Python library *ipfn* was used to perform the IPF process (**https://pypi.org/project/ipfn/)**

In general, at every iteration the aggregation of the segments gets closer and closer to the target totals. Ideally, the segments will match the target totals with no discrepancies. However, this is often not possible as the distribution of the seed—that is, the distribution of PUMS and the CSTDM synthetic population—is not a perfect representation of the distribution in the total population for every county. If the initial distribution is not very different from the target totals, the algorithm eventually converges, leading to very small changes from iteration to iteration.

Variable definition

To explain the algorithm run, the variables used are defined in Table 5.5.

Table 5.5: IPF variable definition

Variable	Definition	Domain
c	County indexer	$c \in C$, all of the counties in California
z	Modeling zone indexer	$z \in [1, \dots, 1189]$, all of the modeling zones
n_c	Target population for county c	$n_c \in \mathbb{Z}^+$
n_z	Target population by modeling zone z	
n_c^{hh}	Target number of households for county c	$n_c^{hh} \in \mathbb{Z}^+$
e	Employment category indexer	$e \in E$, all of the employment categories
s	Household size category indexer	$s \in S$, all of the household size categories
i	Household income category indexer	$i \in I$, all of the household income categories
r_{ce}	Target proportion of population for employment category e in county c	$0 < r_{ce} < 1$
r_{cs}	Target proportion of population for household size category s in county c	$0 < r_{cs} < 1$
r_{ci}	Target proportion of population for household income category i in county c	$0 < r_{ci} < 1$
t	Iteration indexer	$n_c \in \mathbb{Z}$
n_{cesi}^{CSTDM}	Aggregate population from CSTDM synthetic population for county c, employment e, household size s and household income i	$n_{cesi}^{CSTDM} \in \mathbb{R}^+$
n_{cesi}^{t}	Aggregate population at iteration t for county c, employment e, household size s and household income i	$n_{cesi}^{t} \in \mathbb{R}^+$
a_{cs}	Adjustment for household of size s for county c	$a_{cs} \geq 5$

Variable	Definition	Domain
m_{ce}^{t}	Multiplier at iteration t for county c and employment e	$m_{ce}^{t} \in \mathbb{R}^{+}$
m_{cs}^{t}	Multiplier at iteration t for county c and household size s	$m_{cs}^{t} \in \mathbb{R}^{+}$
m_{ci}^{t}	Multiplier at iteration t for county c and household income i	$m_{ci}^{t} \in \mathbb{R}^{+}$
m_{cesi}^{t}	Multiplier at iteration t for county c, employment e, household size s and household income i	$m_{cesi}^{t} \in \mathbb{R}^{+}$
$m_{cesi}^{t}{}'$	Population-weighed m_{cesi}^{t}	$m_{cesi}^{t}{}' \in \mathbb{R}^{+}$

Individual- vs. household-level dimensions

Traditionally, IPF dimensions are targeting the same population. However, in this case, both household-level attributes and individual-level attributes are being used. The required input into the CRRM—and the final output of the reweighting process—are individuals. Therefore, the individual weights are carried through and are adjusted when the dimension is a household dimension (i.e., household size and household income) by dividing the size of the segment by the household size. For households over five people, the county-specific average size in the CSTDM synthetic population, a_{cs}, is being used. For reference, the state-wide average is 6.9432.

Approach

For each iteration of the IPF algorithm, a multiplier is created for each county and dimension. The employment multiplier at iteration t for county c and employment e is:

$$m_{ce}^{t} = \frac{r_{ce} \sum_{e} \sum_{s} \sum_{i} n_{cesi}^{t}}{\sum_{s} \sum_{i} n_{cesi}^{t}}.$$

Note that the size of the segment is converted to the percentage of the county to be comparable to the rates being used. For the household incomes, the number of individuals is divided by the household size. The household size multiplier for iteration t for county c and household size s is:

$$m_{cs}^{t} = \frac{r_{cs} \sum_{e} \sum_{s} \sum_{i} n_{cesi}^{t} / a_{cs}}{\sum_{e} \sum_{i} n_{cesi}^{t} / a_{cs}}.$$

Similarly, the household income multiplier for iteration t for county c and household income i is:

$$m_{ci}^{t} = \frac{r_{ci} \sum_{e} \sum_{s} \sum_{i} n_{cesi}^{t} / a_{cs}}{\sum_{e} \sum_{s} n_{cesi}^{t} / a_{cs}}.$$

Finally, the multiplier for the segment at iteration t, is the average multiplier is:

$$m_{cesi}^{t} = \frac{\left(m_{ce}^{t} + m_{cs}^{t} + m_{ci}^{t}\right)}{3}.$$

These are then weighted by the population to meet target populations for the county, creating:

$$m_{cesi}^{t}{}' = \frac{n_{c}}{\sum_{e} \sum_{s} \sum_{i} n_{cesi}^{t} m_{cesi}^{t}}.$$

Algorithm

The following algorithm is run for a predetermined number of iterations:

1. For each iteration t:
    i. Calculate multiplier:
        a. Calculate employment multipliers, m_{ce}^{t};
        b. Adjust for household size;
        c. Calculate the household size multipliers, m_{cs}^{t};
        d. Calculate the household income multipliers, m_{ci}^{t};
        e. Calculate the average weight, m_{cesi}^{t}; and
        f. Reweight for the county population to get $m_{cesi}^{t}{}'$;
    ii. Trim multiplier if needed:
        a. Trim the multiplier such that it is within the range of some cut-off. In this case, using $\frac{1}{1.5}$ and 1.5.
    iii. Multiply the trimmed multiplier ($m_{cesi}^{t}{}'$) to aggregate population at iteration t for county c, employment e, household size s, and household income i (n_{cesi}^{t}) to generate the next segment size: n_{cesi}^{t+1}.

This algorithm was implemented using the Python ipfn package[58].

For iteration zero ($t = 0$), the CSTDM totals were just scaled to match the population totals by zone for each county.

Convergence

While the theoretical convergence of IPF depends on a series of factors, numerable papers have discussed the use of IPF in spatial microsimulation. Diagnostics have found that, for most purposes concerning basic special microsimulation weighting, 10 iterations are enough to approach reasonable results using the ipfn library. To check convergence, the algorithm was run with 25 iterations. For the purposes of testing the convergence, four types of measurements were considered:

- The **mean of the segment multipliers**, $m_{cesi}^{t}{}'$, which should converge to one as the algorithm converges.
- The **standard deviation of the segment multipliers**, which should converge to zero as the algorithm converges.
- The **absolute error**, which is the absolute difference between the iteration total and the target total, normalized by the total population or number of households.
- The **distribution error**, which is the absolute difference between the iteration shares and the target shares.

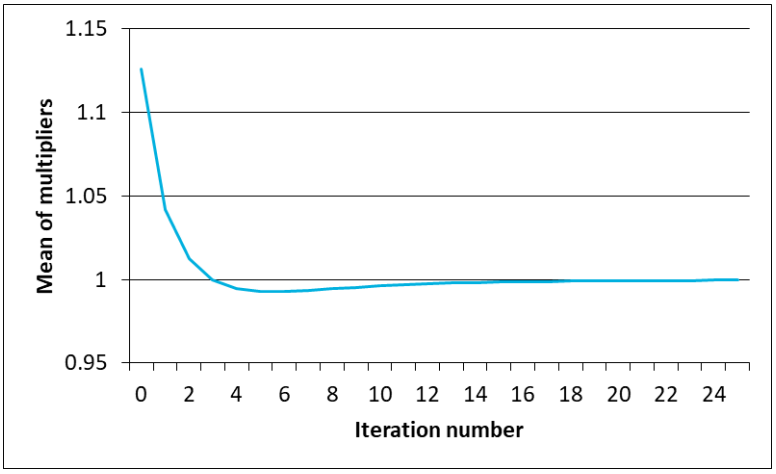
[58] **https://pypi.org/project/ipfn/**

The absolute error is subject to the different scaling effects of population and households. Since the CRRM models individuals, the input population is scaled to individuals, such that the number of households does not ever completely match up. The weighting algorithm outputs 3.4% more households than the target totals. However, the population—which was 2.0% smaller if the CSTDM synthetic population was to be used directly—is directly scaled to, making it match up perfectly. The absolute errors for Household size and Household income are not expected to converge to zero due to this scaling effect.

Alternatively, the distribution error captures whether the distribution of the population or households matches the targets, making those independent of the scaling effect. The distribution error is expected to converge to zero.

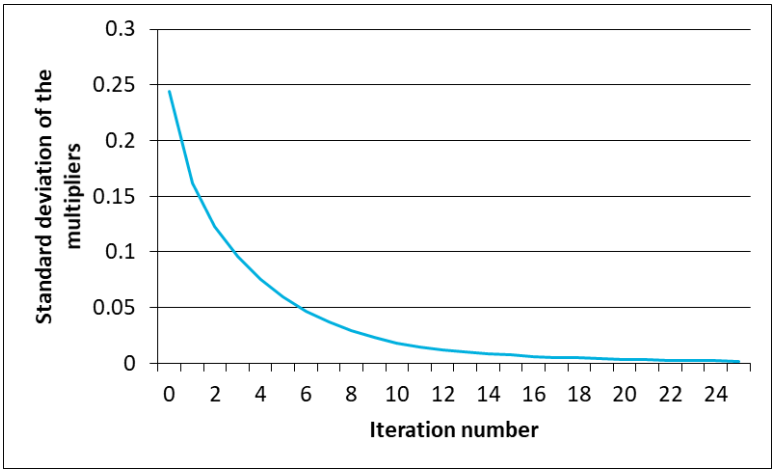
Figure 5-1 shows the mean of the multipliers for each iteration of the algorithm. As expected, the mean converges to one. Note that, because the seed of algorithm is the CSTDM synthetic population output, the weights are relatively close to one since the beginning.

Figure 5-1: Mean of the multipliers per iteration.



Similarly, Figure 5-2 shows the standard deviation of the multipliers for each iteration. As expected, the multipliers converge to zero.

Figure 5-2: Standard deviation of multipliers per iteration



Finally, Figure 5-3 through Figure 5-5 show the cumulative state-wide absolute error for Employment, Household size, and Household income, respectively. An additional line is included for the initial error of the aggregate CSDTM synthetic population. As noted previously, this absolute error is subject to scaling issues, which are seen in the absolute error for Household size.

Figure 5-3: State-wide absolute error for Employment by iteration

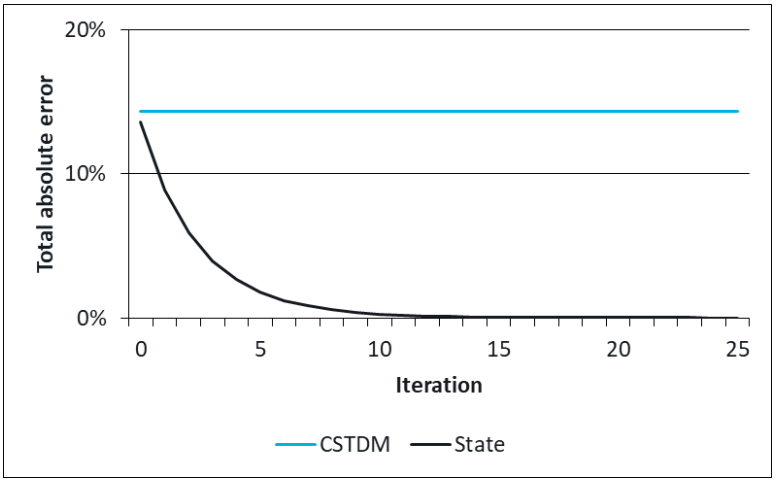


Figure 5-4: State-wide absolute error for Household size by iteration

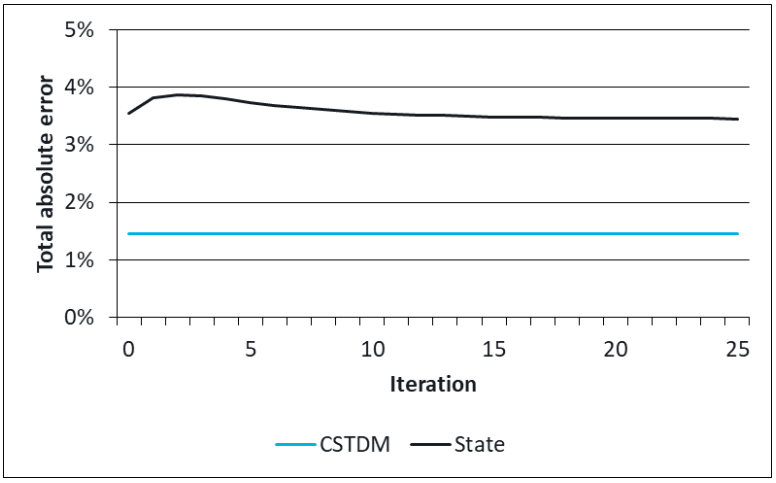
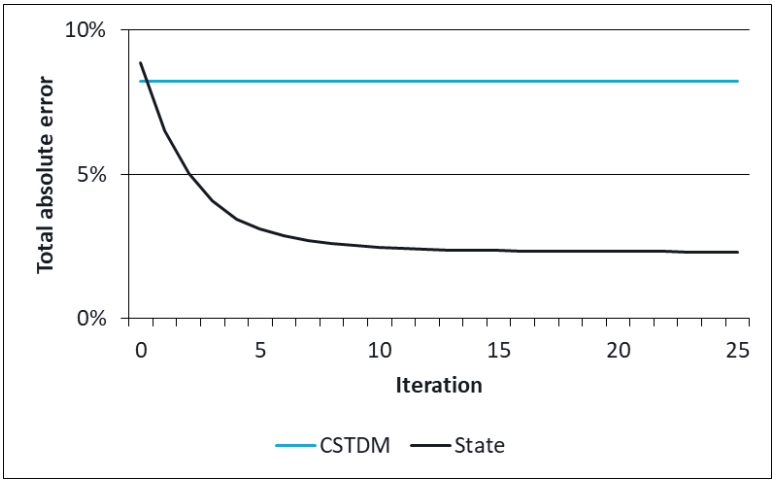


Figure 5-5: State-wide absolute error for Household income by iteration



Similarly, the distribution error plots are shown for the full state in Figure 5-6 through Figure 5-8 for Employment, Household size, and Household income, respectively. As expected, the error in the distribution of all three categories is converging to zero. This means that, while the *number* of total households is higher between the aggregated CSTDM synthetic population and the target totals, the *distribution* is more aligned with what is expected. The population size, on the other hand, directly matches the totals.

Figure 5-6: State-wide distribution error for Employment by iteration

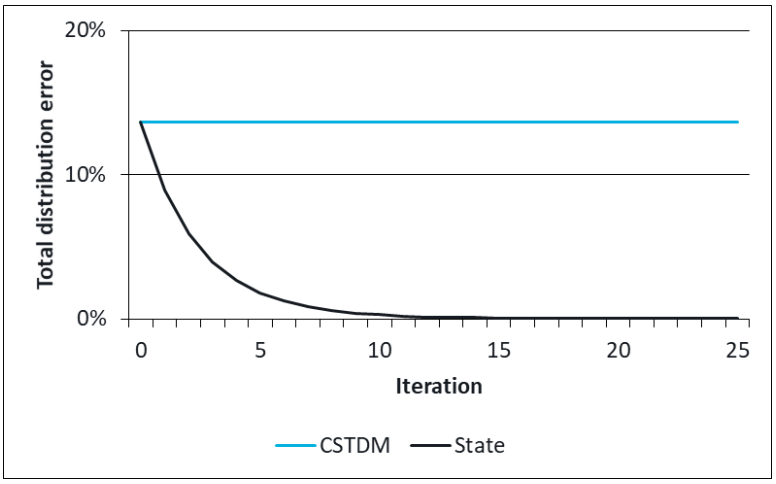


Figure 5-7: State-wide distribution error for Household size by iteration

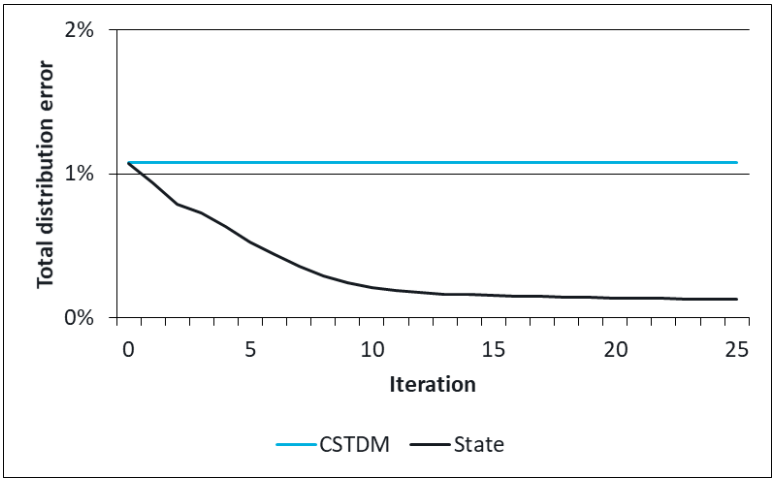
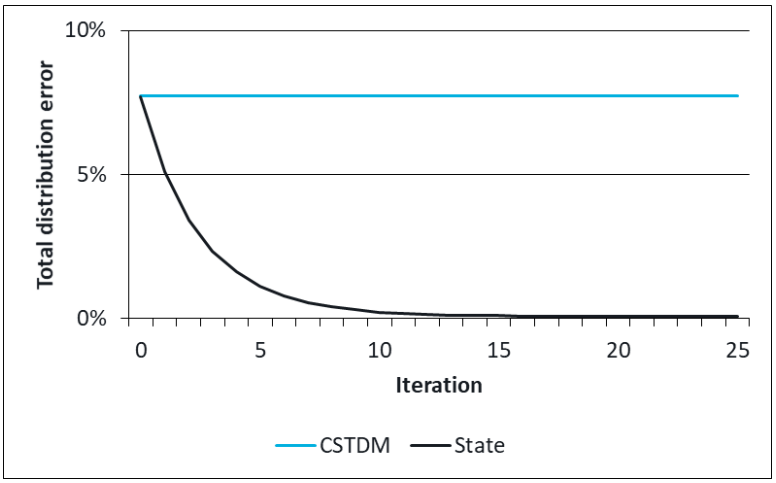


Figure 5-8: State-wide distribution error for Household income by iteration



For the purposes of the inputs, the 25th iteration is used.

Weighting by CRRM zone

The population inputs for the CRRM are disaggregated by the 1,186 modeling zones. To get to the zone-level segments, the reweighted aggregate CSTDM synthetic population is used.

Once the CSTDM has been adjusted to target totals by county, the implied multipliers are calculated by county. This multiplier is calculated as:

$$\frac{n_{cesi}^{25}}{n_{cesi}^{CSTDM}}.$$

The county-level multipliers are then multiplied by the zone-specific segments derived with the same process from the CSTDM, n_{zesi}^{CSTDM}.

$$n_{zesi}^{t*} = n_{zesi}^{CSTDM}\left(\frac{n_{cesi}^{25}}{n_{cesi}^{CSTDM}}\right)$$

Finally, the product is adjusted to account for the target zone-specific population. The final value is:

$$\hat{n}_{zesi} = n^{t*}_{zesi}\left(\frac{n_z}{\sum_e \sum_s \sum_i n^{t*}_{zesi}}\right).$$

These final inputs are used as the population inputs for the CRRM.

Final distributions

The population inputs for each category are summarized below for household size, household income, and employment status, respectively.

Table 5.6: Distribution of household size for the target totals and the CRRM population inputs

Household size	Percent of California target totals	Percent of CRRM population inputs
1 person	23.8%	25.7%
2 people	30.3%	35.1%
3 people	16.7%	14.4%
4 people	15.2%	13.0%
5+ people	14.0%	11.7%

Table 5.7: Distribution of household income for the target totals and the CRRM population inputs

Household income	Percent of California target totals	Percent of CRRM synthetic population
Less than $10,000	5.1%	5.2%
$10,000 - $14,999	4.4%	5.3%
$15,000 - $24,999	8.0%	9.3%
$25,000 - $34,999	7.9%	9.0%
$35,000 - $49,999	10.9%	12.2%
$50,000 - $74,999	15.9%	17.0%
$75,000 - $99,999	12.3%	12.3%
$100,000 - $149,999	16.2%	14.9%
$150,000 - $199,999	8.4%	6.8%
$200,000 or more	11.0%	8.1%

Table 5.8: Distribution of employment status for the target totals and the CRRM population inputs

Employment status	Percent of California target totals	Percent of CRRM synthetic population
Employed	45.6%	45.7%
Retired	11.0%	11.4%
Homemaker	5.6%	5.7%
Student	24.1%	23.7%
Other	7.4%	7.3%

Under 5 years old	6.3%	6.2%

Future year forecasts

Future year forecasts of population and employment for 2030, 2040 and 2050 were produced, using the 2018 detailed data derived above and aggregate County level forecasts provided by the client. A four-step process was undertaken as follows:

1. County Growth from the client – distribute it to BGRPs – based on NAICS Industry definitions.
   – New total by zone/BGRP – total jobs for 2030, 2040, 2050
2. Look at the ratio of all to primary jobs from 2002-2018 using Longitudinal Employer-Household Dynamics (LEHD LODES) (2002-2018), from Census Bureau[59] – identify ratio for 2030, 2040, 2050.
   – Applying the ratio to total jobs for 2030, 2040, 2050 to convert those to total workers.
3. Translating from the work zone to home zone using OD data, split by age.
4. Population growth to home zones and distribute employment.

Summary

The resulting data utilized in the model is summarized by county in Table 5.10.

Non-resident and external trips

As previously noted, growth of non-resident and external trips uses a different and simpler approach. These use the 'global' growth factor of the resident population total, with the exception of the Las Vegas/Clark County external zones, which use the Nevada population forecasts. These are summarized in Table 5.9 below.

Table 5.9: Non-resident and external growth factors

Year	Non-Resident and External	Clark County
2018	1.0000	1.0000
2030	1.0084	1.1811
2040	1.0256	1.2892
2050	1.0242	1.3551

[59] https://lehd.ces.census.gov/data/

Table 5.10: Summary of socio-economic data

County	Population				Households				Employment			
	2018	2030	2040	2050	2018	2030	2040	2050	2018	2030	2040	2050
Alameda	1,596,130	1,670,455	1,795,198	1,898,488	585,990	621,433	672,826	718,556	830,893	871,743	899,954	916,246
Alpine	3,369	1,200	1,187	1,201	1,461	521	516	541	423	791	810	831
Amador	37,634	41,584	40,621	38,929	16,169	17,849	17,339	16,683	11,868	13,340	14,040	14,510
Butte	211,127	211,002	224,028	242,078	86,829	86,838	92,070	100,717	81,508	87,605	89,146	91,392
Calaveras	42,997	43,735	40,752	37,686	18,642	19,036	17,552	16,321	9,204	10,520	10,430	10,430
Colusa	21,818	22,135	21,532	20,406	7,439	7,667	7,512	7,175	8,429	10,200	10,160	10,200
Contra Costa	1,184,957	1,171,945	1,274,708	1,361,137	429,151	429,892	467,934	503,818	362,821	389,962	399,929	409,267
Del Norte	27,423	24,738	23,347	21,836	11,527	10,533	9,975	9,421	8,392	8,120	7,970	7,850
El Dorado	186,885	185,434	179,456	168,423	75,531	75,542	72,280	68,313	46,995	64,647	67,053	69,087
Fresno	963,206	1,047,382	1,083,901	1,098,206	316,891	347,675	361,522	370,250	377,351	453,247	474,262	488,680
Glenn	26,194	29,182	28,513	26,584	9,590	10,828	10,645	10,033	9,030	9,920	10,000	10,070
Humboldt	135,532	131,729	126,479	121,539	58,917	58,302	57,065	55,841	48,446	51,997	51,970	52,012
Imperial	184,406	184,997	189,972	192,294	59,216	60,191	61,994	63,245	41,751	67,938	69,203	75,466
Inyo	10,815	18,887	18,552	18,093	5,278	9,367	9,255	9,106	7,290	7,630	7,670	7,650
Kern	850,789	940,257	966,310	969,968	277,463	308,815	317,512	321,524	325,123	373,223	387,175	397,088
Kings	140,830	157,531	161,190	160,446	45,810	51,789	53,355	53,520	53,706	53,523	56,301	58,571
Lake	64,775	68,446	67,564	67,065	27,384	28,965	28,385	28,120	16,043	18,370	19,010	19,730
Lassen	32,096	25,708	21,772	17,983	13,986	11,446	9,762	8,194	7,141	8,610	8,980	8,610
Los Angeles	10,067,183	9,566,663	9,306,759	8,877,939	3,542,900	3,457,683	3,428,661	3,344,049	4,696,616	4,741,504	4,844,201	4,918,527

County	Population				Households				Employment			
Madera	152,427	161,980	163,345	161,937	48,962	52,414	52,849	52,644	72,446	58,171	62,393	64,845
Marin	276,295	244,319	245,498	243,295	116,577	107,274	109,130	109,964	113,873	127,772	133,914	137,214
Mariposa	17,858	17,017	16,588	16,372	7,909	7,594	7,329	7,167	4,640	5,440	5,290	5,200
Mendocino	90,500	88,789	89,200	89,697	37,868	37,569	38,037	38,607	30,936	34,924	35,057	35,073
Merced	280,028	311,578	329,168	336,170	87,275	96,571	101,475	104,142	77,159	91,804	93,622	95,589
Modoc	9,248	8,346	7,463	6,464	4,222	3,841	3,388	2,940	2,343	2,880	2,890	2,870
Mono	21,353	12,987	12,068	10,881	10,381	6,391	6,017	5,495	2,684	8,240	8,220	8,260
Monterey	438,913	434,506	436,307	430,706	142,056	143,089	145,719	146,171	179,296	208,438	213,639	217,996
Napa	152,976	132,087	131,600	128,515	57,746	50,944	50,994	50,643	74,896	82,875	85,857	88,160
Nevada	88,270	97,464	94,444	89,649	39,040	43,245	41,499	39,411	30,930	35,395	35,773	35,965
Orange	3,141,992	3,201,361	3,283,811	3,307,387	1,078,185	1,124,328	1,168,774	1,196,457	1,608,654	1,758,299	1,783,555	1,796,951
Placer	391,217	443,936	474,905	490,667	153,335	174,574	184,749	191,615	160,753	202,468	208,375	214,099
Plumas	18,460	17,530	15,319	13,712	8,581	8,197	7,110	6,441	5,818	6,740	6,410	6,100
Riverside	2,408,962	2,540,559	2,637,463	2,670,068	798,257	851,792	889,154	913,435	701,382	899,463	948,171	972,532
Sacramento	1,500,992	1,611,309	1,708,461	1,782,519	559,286	605,994	644,964	676,798	674,370	749,989	779,615	802,687
San Benito	47,705	71,265	75,452	76,959	14,882	22,362	23,688	24,382	13,365	19,439	19,836	20,269
San Bernardino	2,104,994	2,257,518	2,302,286	2,287,280	670,580	724,620	740,484	742,337	795,651	943,601	1,004,816	1,060,036
San Diego	3,290,379	3,373,792	3,416,779	3,394,592	1,211,368	1,267,872	1,303,671	1,319,161	1,508,363	1,621,686	1,663,295	1,702,140
San Francisco	870,037	837,021	845,589	848,071	384,214	377,001	383,957	386,357	759,701	864,799	905,910	944,639
San Joaquin	721,097	831,956	896,033	942,102	236,047	272,591	293,112	310,861	259,362	312,086	327,842	336,253

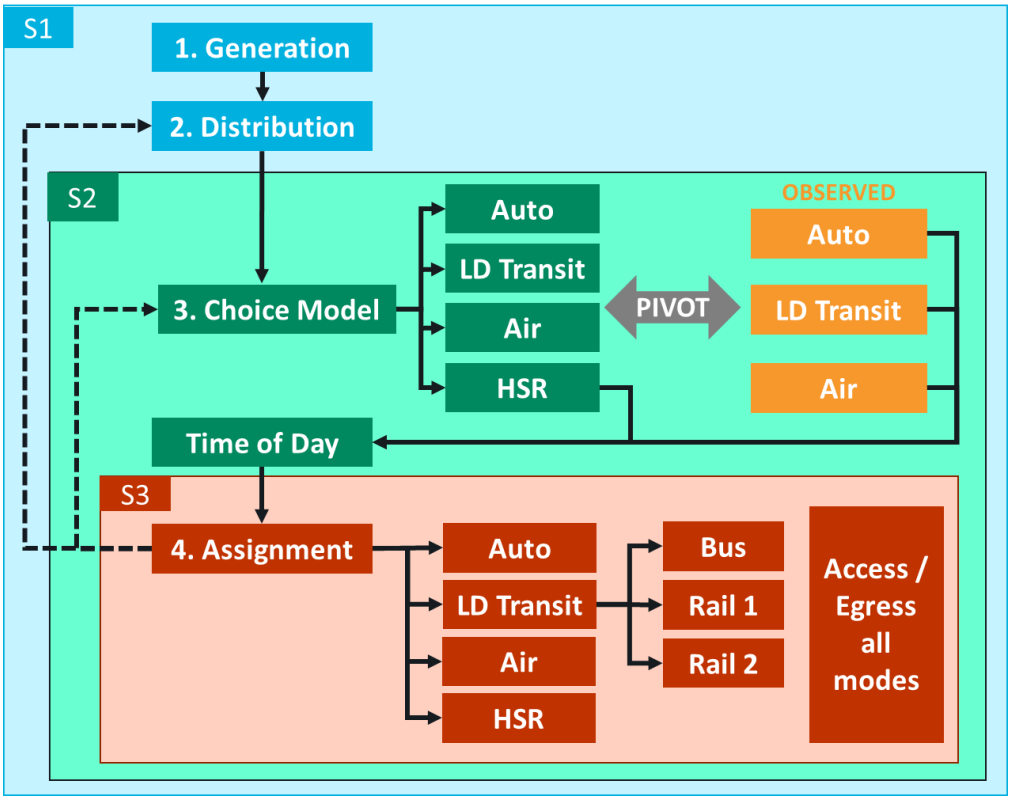
County	Population				Households				Employment			
San Luis Obispo	288,265	286,547	287,621	279,398	118,568	119,585	120,918	120,483	95,991	129,941	131,335	132,181
San Mateo	750,613	721,006	728,934	726,771	275,112	272,120	279,660	281,902	411,344	443,543	451,754	457,690
Santa Barbara	431,369	459,727	475,401	486,994	150,530	162,093	169,647	178,888	170,756	231,192	234,786	237,977
Santa Clara	1,973,286	1,900,159	2,009,127	2,075,768	688,267	675,903	716,945	744,545	1,111,857	1,190,517	1,215,461	1,240,316
Santa Cruz	264,909	268,734	269,540	266,117	99,682	102,986	104,320	105,667	86,247	115,175	118,275	120,773
Shasta	175,025	178,722	180,245	181,492	72,145	74,258	74,599	75,540	63,998	72,963	73,227	73,098
Sierra	3,051	3,132	2,944	2,844	1,570	1,655	1,542	1,479	468	626	564	527
Siskiyou	43,013	43,068	41,085	39,107	19,415	19,611	18,546	17,810	13,184	13,880	13,370	12,980
Solano	423,809	451,280	476,163	494,487	152,938	163,720	172,822	180,668	153,039	150,504	155,944	158,777
Sonoma	484,610	475,831	459,445	434,406	193,086	194,090	189,702	182,854	205,019	223,418	228,989	233,813
Stanislaus	530,603	558,565	577,523	593,396	178,546	189,083	195,857	203,030	198,813	211,373	216,468	221,046
Sutter	96,640	104,005	105,803	104,604	33,148	36,110	36,797	36,806	28,414	35,189	36,821	37,943
Tehama	84,698	65,151	64,900	64,129	33,804	26,172	26,007	25,854	18,126	22,440	23,320	23,840
Trinity	12,840	16,042	15,727	15,442	6,025	7,564	7,343	7,142	2,334	2,750	2,750	2,770
Tulare	505,901	487,378	487,888	472,966	157,481	153,105	154,190	151,600	147,315	180,237	185,163	189,344
Tuolumne	54,137	50,082	48,956	48,542	23,649	21,973	21,595	21,875	16,523	17,870	17,510	17,030
Ventura	881,412	805,456	789,877	758,161	302,671	283,215	281,103	274,187	313,479	346,210	366,197	382,302
Yolo	212,092	230,484	240,261	243,409	77,418	84,223	88,032	90,990	82,216	127,714	131,586	134,333
Yuba	75,767	87,172	91,389	94,142	26,655	30,677	32,294	33,871	16,962	21,400	22,700	23,850

6 Generation and Distribution

Introduction

The California Rail Ridership model (CRRM) is a 4-Step travel demand model. Figure 6-1 shows the flow chart of the steps for the model. In this section, we will discuss the first two steps – Generation and Distribution.

Figure 6-1: model flow chart



In the Generation step, we create trip production and attraction rates by zones in the model via trip production models and trip attraction models. The approach used for the estimation/calibration/validation of these production and attraction models is set out below.

In the Distribution step, we assign the zonal level trip productions and attraction into a production-attraction trip table. This is done via gravity models. The approach used within the gravity models and their calibration/validation against observed data is set out below.

Sources of data

The estimation of the trip generation and distribution models is based on trip making data obtained from various sources. The key sources are discussed below.

2017 National Household Travel Survey – California Add-On

The National Household Travel Survey (NHTS) is a national trip making survey conducted by the Federal Highway Administration (FHWA). The survey was last conducted between 2016-2017. Each state has add-on information that shows state specific travel patterns. For this work we used the California Add-On of NHTS.

The survey estimates a total of over 120 million trips[60] for an average weekday in entire state of California, based on a sample of 23,391 households and 43,850 individuals. The NHTS contains very detailed information regarding persons, trips, and vehicles of surveyed households. However, for the generation model we will only consider the following information:

- **Trip purpose**: NHTS includes trip purpose as part of the surveyed data and contains 20 different categories. Since the scope of the generation models are daily trips, the travel day trip purpose variable is used to identify the purpose of the outbound and return legs.
- **Origin/Destination trip location identifier and household location**: These are used to identify location type, county, and zone location. An additional process was required to map the ZIP Code Tabulation Areas (ZCTAs) to our zoning system.
- **Trip mode**: There are 25 different responses in the *mode* field of the NHTS. These were grouped into 9 different categories to have a more manageable mode analysis. This trip characteristic is used for identification of in-scope trips, validation, and data processing only since the generation models are agnostic of the mode.
- **Trip Time and day characteristics**: The NHTS includes the starting and ending time of each trip entry, as well as a flag to identify weekend trips.[61] This information is used to characterize trip time period and identify direction (outbound/return).
- **Trip Distance**: Traveled distance is used mainly to identify in-scope trips.
- **Household family income category**: The NHTS includes 14 categories for family income. This information is used as one of the variables for the cross-classification analysis. However, they are aggregated into 3 different categories (high, medium and low income).
- **Household size**: Four different categories are defined (i.e., 1, 2, 3, and 4+ household members).
- **Number of workers in the household**: Similarly, six different categories are defined.

Socioeconomic data

There are multiple sources available for the main socioeconomic variables in California, varying in the level of aggregation, geographic coverage, publication year, years projected (when available) and segmentation. As part of the development of the CRRM, a thorough research and

[60] For the purposes of this analysis, both weekdays and weekends are considered, therefore the 7-day weights are used in the total trip estimation.

[61] The weekend is defined as Saturdays, Sundays, and Fridays after 6 PM.

reconciliation process has taken place. The Population and Employment Data section of this report describes these data in more detail.

The main descriptive variables used for the generation models and the public data that is part of the process, are the following:

- Population
  - 2011-2015 American Community Survey 5-Year Estimates (2015), from American Factfinder[62]
  - Historical Population Estimates by Decade (2010-2019), from California Department of Finance[63]
  - Population and Housing Estimates with Census Benchmark (2010-2019) from California Department of Finance[64]
- Households
  - 2011-2015 American Community Survey 5-Year Estimates (2015), from American Factfinder[65]
- Employment
  - Current Employment Statistics (CES) (2020), from CEDD/BLS[66]
  - Quarterly Census of Employment and Wages (2019), from CEDD/BLS[67]
  - Long-Term Occupational Employment Projections (2019), from CEDD/BLS[68]
  - Longitudinal Employer-Household Dynamics (LEHD LODES) (2002-2017), from Census Bureau[69]
- Income
  - 2011-2015 American Community Survey 5-Year Estimates (2015), from American Factfinder[70]
- School enrollment
  - Enrollment in California Public School Districts – 1718 (2017-2018), from CDE's DataQuest[71]

---

[62] https://data.census.gov/cedsci/

[63] http://dof.ca.gov/Forecasting/Demographics/Estimates/

[64] http://dof.ca.gov/Forecasting/Demographics/Estimates/e-5/

[65] https://data.census.gov/cedsci/

[66] https://data.edd.ca.gov/Industry-Information-/Current-Employment-Statistics-CES-/r4zm-kdcg

[67] https://data.edd.ca.gov/Industry-Information-/Quarterly-Census-of-Employment-and-Wages-QCEW-/fisq-v939

[68] https://data.edd.ca.gov/Employment-Projections/Long-Term-Occupational-Employment-Projections/4yzm-uyfq

[69] https://lehd.ces.census.gov/data/

[70] https://data.census.gov/cedsci/

[71] https://dq.cde.ca.gov/dataquest/content.asp

Observed trips tables.

Origin-destination trip tables have been developed as part of the base demand estimation for auto, rail, long-distance bus, and air. Each trip table was estimated using multiple data sources and different estimation, calibration and validation procedures depending on the available information. A detailed description of each process and sources can be found in the relevant sections of this report.

Limitations/Caveats

- The NHTS dataset includes long-distance trips, however the sampling design did not include explicit variables to accurately represent long-distance trips. Nonetheless, the dataset is the most recent available, and is therefore considered the most appropriate to use.
- Observed Trip Tables were built with a level of granularity at the county level, i.e., they are most suitable to represent up to county-to-county movements. Below this level of detail there will be a greater level of inherent uncertainty.
- Socioeconomic characteristics reflect only the data sources publicly available at the time of development of the CRRM.

Generation model approach

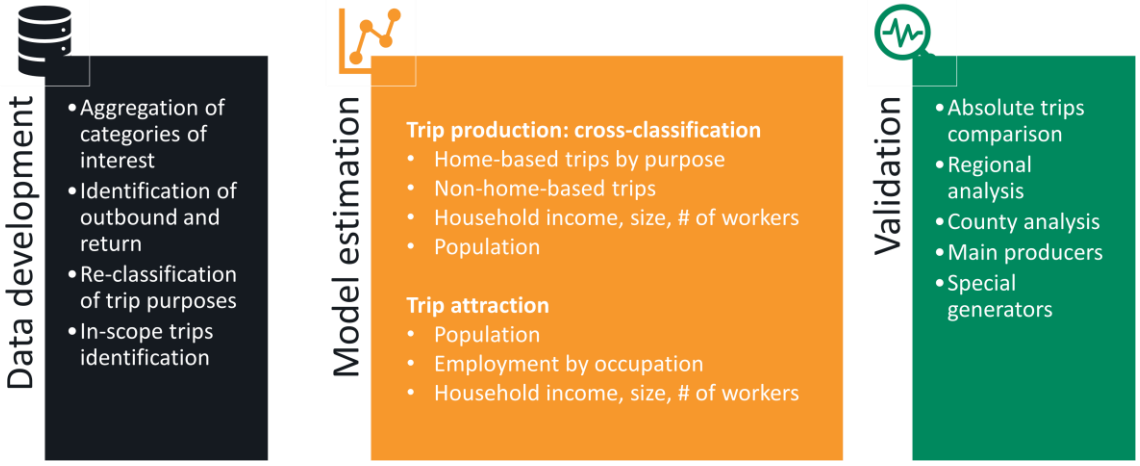
Overview

The purpose of the generation model is to represent as close as possible the trip-ends for daily in-scope trips[72] in California. In the CRRM generation model the demand is represented in Production Attraction (PA) format.

The production and attraction models are built based on the observed travel behavior obtained from the NHTS. The models are then validated based on observed trip tables. Trip production rates are developed using the cross-classification of household socioeconomic characteristics by trip purpose. The attraction model is developed based on regression analysis using household survey data, employment by occupation and population data.

The overall approach for the development of the generation model is shown in the following diagram. A more detailed description of each step is included in the following sections.

Figure 6-2: Generation model development approach



Source: Steer, 2021

Methodology

Data development

The basis for the construction of the generation models is the NHTS, given that it is the most recent travel study in the state of California.

Given the PA approach, it is important to identify within a person's trip journal which trips can be considered as outbound and return, especially for home-based, since at a daily level it is appropriate to assume that there is a returning trip.

[72] Given the Zoning System defined and the main use of this model for intercity rail forecasting.

Identification of outbound and return

In a first exploration phase it was clear that the trip-purpose attributes within the NHTS were mostly assigned to each individual leg of a trip and additional processing was required to analyze the data and appropriately identify outbound and return trips. For example, it was identified that 33%[73] of commute trips stopped on the way to their main destination (e.g., for shopping or to pick-up/drop-off someone), and only the initial and final leg were assigned as daily purpose "To/from work," providing inaccurate origin and/or destination for commute trips. This information is paramount for the generation process since the characteristics of the origin and destination are used to explain the production and attraction of trips for each identified purpose.

Given that similar trends were found for other purposes, we identified the outbound/return trips by purpose for home-based trips, as follows:

- **Step 1. Identify home-based trip chain.** Identify the consecutive set of trips that begin and end at home location. For example, given the following trip chain: 1) Home->Shop, 2) Shop->Work, 3) Work->Home, 4) Home->School, 5) School->Home; there would be two resulting home-based trip chains. The first one including 1-3 and the second 4-5.
- **Step 2. Identify the *main destination* for each trip chain.** The main destination is defined as the location where the individual spends most of their time during the day (i.e., maximum *dwell time*). For example, if the first leg starts at 9 am, takes 30 minutes and the next leg starts at 12 noon, then the dwell time is 12-(9+0.5) = 2.5 hours. This analysis is done on a person-by-person basis, checking the starting time and end time of each trip entry.
- **Step 3. Identify Outbound and Return for each trip chain.** All the trips that take place before arriving to the *main destination* are categorized as Outbound, and all the rest are categorized as Return.
- **Step 4. Identify and extract *inner loops*.** Identify sets of trips that start and end at the same location within the home-based trip chains. Even if a set of consecutive trips begin and end at the home location (Step 1), it is possible that this contains *inner loops*, where the person starts from an intermediate location (not home), go to a different location (not home either), or comes back. For example, in the following trip chain: 1) Home->Work, 2) Work->Coffee, 3) Coffee->Work, 4) Work->Home; trips 2 and 3 are clearly not home-based trips even though they seem to be within a home-based "work" trip chain. In these cases, a completely independent outbound/return non-home-based trip is identified with its own purpose.
- **Step 5. Collapse home-based trips into outbound and return.** Once *inner loops* are extracted,[74] all the trips within the outbound (or return) leg are collapsed into a single trip, whose origin is home and destination are the *main destination* (or vice versa). Additional rules for the remaining attributes are applied:
  - For numerical attributes, such time and distance, the sum of the single leg's values are assigned as the collapsed-trip attribute value.

[73] The ratio was obtained based on the weighted trips, with the 7-day weights. This subset corresponds to ~35% of trip entry points (i.e., unweighted trips).

[74] These are set apart to be analyzed as part of the non-home-based trips.

- For the trip mode: If multiple modes are found, a new category of "multimode" is defined and an additional attribute with the concatenation of modes is included. E.g., "Bus-Walk-Auto".
- For the trip purpose the following approach is assumed: If "Commute" purpose is present in the trip chain, the purpose is set to "Commute". Otherwise, the purpose of the last trip of the outbound leg is assumed to be the purpose for outbound and return.

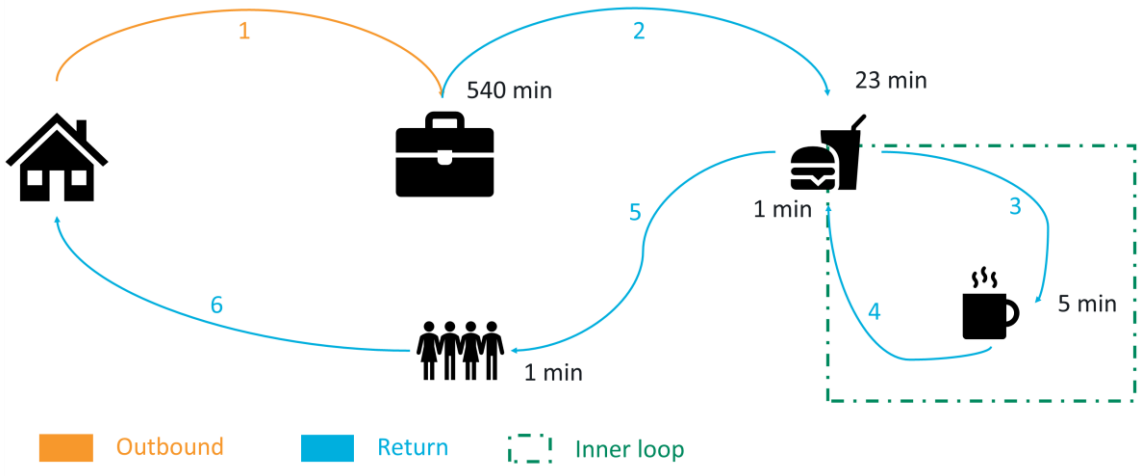
Consider the example described in Table 6.1 and all the steps of the proposed approach illustrated in Figure 6-3.

Table 6.1: Example #1 of home-based trip chain processing

Trip No.	Origin	Destination	Mode	Purpose	Departure time	Arrival time	Trip duration (min)
1	Home	Work	Bus	Commute	5:30	7:00	90
2	Work	Other # 1	Bus	Social	16:00	16:30	30
3	Other # 1	Other # 2	Walk/Bike	Social	16:53	16:58	5
4	Other # 2	Other # 1	Walk/Bike	Social	17:03	17:08	5
5	Other # 1	Other # 3	Walk/Bike	Social	17:09	17:20	11
6	Other # 3	Home	Auto	Commute	17:21	18:51	90

Source: Steer, 2021

Figure 6-3: Proposed approach applied over Example #1



Source: Steer, 2021

- **Step 1 and 2**: From the example in Table 6.1, this is a home-to-home trip chain, and the highest dwell time is in the Work location, spending 9 hours there. Therefore, the main destination is Work.
- **Step 3**: Given that Work is the main destination, trip 1 is the outbound leg, and trips 2-6 are the return leg (see blue links in Figure 6-3).

- **Step 4**: In Figure 6-3, trips #3 and #4 form a loop and it is completely within the return leg, therefore, it is extracted from this trip chain as a non-home-based trip. The remaining trips conform to the home-based trip of analysis.
- **Step 5**: Final outbound trip is uniquely trip # 1, whereas the return trip is an aggregation of trips 2, 5, and 6 (see Table 6.1: Example #1 of home-based trip chain processing).

Table 6.2: Collapsed outbound and return trip from Example #1

Trip No.	Origin	Destination	Mode	Mode description	Purpose	Departure time	Arrival time	Travel time
1	Home	Work	Bus	Bus	Commute	5:30	7:00	90
2	Work	Home	Multimode	Bus-Walk/Bike-Auto	Commute	16:00	18:51	131

Source: Steer, 2021

Note that the approach considered the identification of outbound and return of home-based trips and non-home-based trips that have both legs (extracted in Step 4). However, trip chains that either contain only Outbound (from home) or Return (to home) legs are not included as part of this analysis. This correspond to 5% of the trip entries and are treated as non-home-based trips.

A key assumption for PA matrix representation is that, at a daily level, the trips are balanced by direction. Therefore, including one-directional trips would unbalance the PA matrices. For this reason, we assume that non-home-based trips are also balanced.

So, for non-home-based trips (i.e., home location is not part of the trip chain), after the extraction of inner loops there are two possible cases: a) trips that have outbound and return (~98% of non-home-based observations), and b) are one-directional trips (~2% of observations). For the non-home-based trips that have outbound and return, the collapsing process explained before is carried over and a single leg per direction represents the trips. On the other hand, for single leg trips the opposite directional trip is synthetically created, thereby maintaining the total number of trips. In other words, each direction now has half the trips of the original single-leg weight. Finally, most attributes are carried from the single leg trip: purpose, distance, travel time. However, the starting time of the synthetic leg is assigned using the County-County time-of-day average distribution of the synthetic direction.

In-scope trips identification

Once a complete dataset with outbound-return trips was obtained, the in-scope trips for the rail forecasts were extracted. The following filters, which are similar to the ones applied to obtain the observed trip tables, were applied to the extracted trips:

- Exclude all trips whose origin and destination are in the same zone according to the CRRM Zoning System definition, i.e., intrazonal trips.
- Exclude all trips that are entirely made by bike or walk modes, since only motorized vehicles were considered in the base demand analysis.

This analysis resulted in a total of 50,191,290 in-scope trips, made by California residents within California, which is approximately 42% of the original trips in NHTS. However, the in-scope trips in the base demand were only 5,629,496 trips. To identify the in-scope trips for this study, we compared the trip-length distribution of base demand and NHTS processed dataset (see Table 6.3). The data indicates that the shorter the distance, the higher the overestimation of trips (within the NHTS database relative to the in-scope demand estimate).

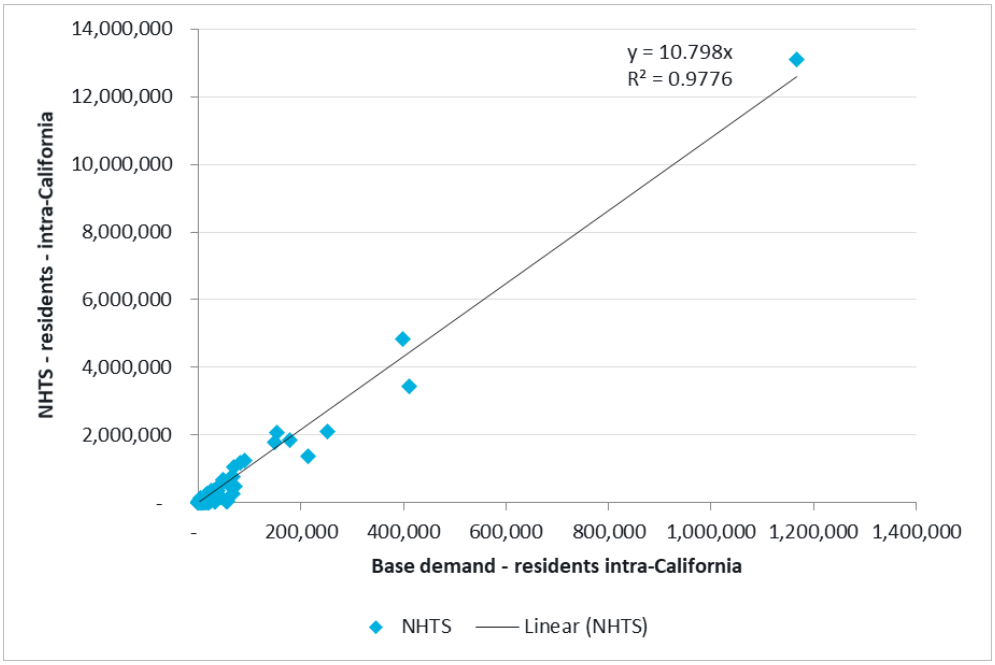
Table 6.3: Trip-length distribution comparison. Base demand vs NHTS

Range	Base demand	Processed NHTS	Difference %
[0,10)	2,686,737	28,115,584	946%
[10,20)	976,523	11,177,380	1045%
[20,30)	422,325	4,484,829	962%
[30,40)	286,698	2,230,233	678%
[40,50)	196,121	1,216,650	520%
[50,60)	175,815	879,628	400%
[60,70)	151,880	408,226	169%
[70,81)	127,627	323,099	153%
[81,90)	100,913	251,289	149%
[90,100)	76,846	206,679	169%
[100,125)	116,166	305,007	163%
[125,150)	73,276	223,181	205%
[150,175)	41,608	70,064	68%
[175,200)	25,019	71,771	187%
[200,250)	44,226	71,606	62%
[250,300)	20,007	19,003	-5%
[300,350)	23,012	38,577	68%
[350,400)	35,325	33,754	-4%
[400,500)	37,674	54,516	45%
>=500	11,698	10,214	-13%
Total	**5,629,496**	**50,191,289**	

Source: Steer, 2021

In addition, the relative trip patterns between the base demand and the NHTS-processed data was evaluated, by comparing county pair trip data (see Figure 6-4). The table illustrates R^2 statistic of the linear fit is very close to 1 (0.9776), which indicates that the patterns are consistent between the base demand and the NHTS data. However, the slope is ~10.8, which means that the NHTS demand is generally significantly higher than the base demand.

Figure 6-4: Linear correlation between county-county base demand and NHTS-processed data



Source: Steer,2021

With these results, it is concluded that the most appropriate way to represent in-scope trips was to scale the data by distance bins. This ensures that other trip patterns within each distance bracket remained unchanged, while the balance by distance is realigned to be consistent with the base demand trip-length distribution.

The distance brackets defined for scaling of the NHTS data are outlined below:

Table 6.4: In-scope base demand targets for scaling

Distance bracket	Targets (In-scope observed demand)
[0,10)	2,686,737
[10,30)	1,398,847
[30,60)	658,635
[60,100)	457,265
[100,150)	189,442
[150,200)	66,627
>200	171,943
Total	**5,629,496**

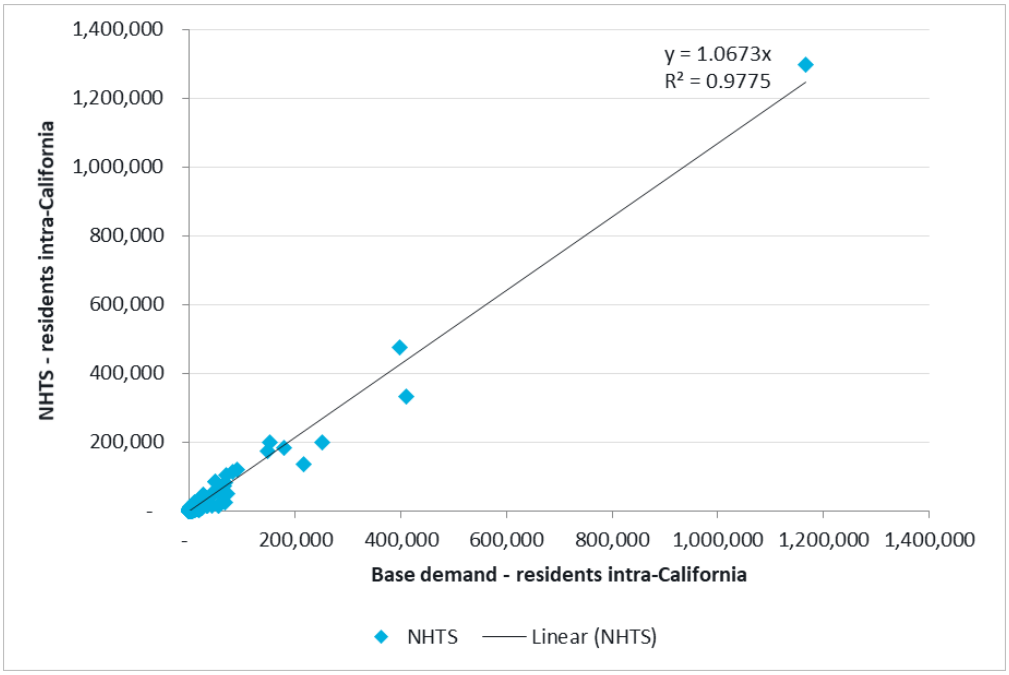
Source: Steer, 2021

Trip patterns scaled processed NHTS.

In Figure 6-5 the final trip patterns after adjustment are shown. The patterns show similar correlation to base demand (as shown in Figure 6-4: Linear correlation between county-county

base demand and NHTS-processed data), but with a slope closer to 1, for all the county-county trips; meaning that the adjustment is successfully implemented. Also, the relative demand of the different counties remained unchanged, by comparing before and after adjustment.

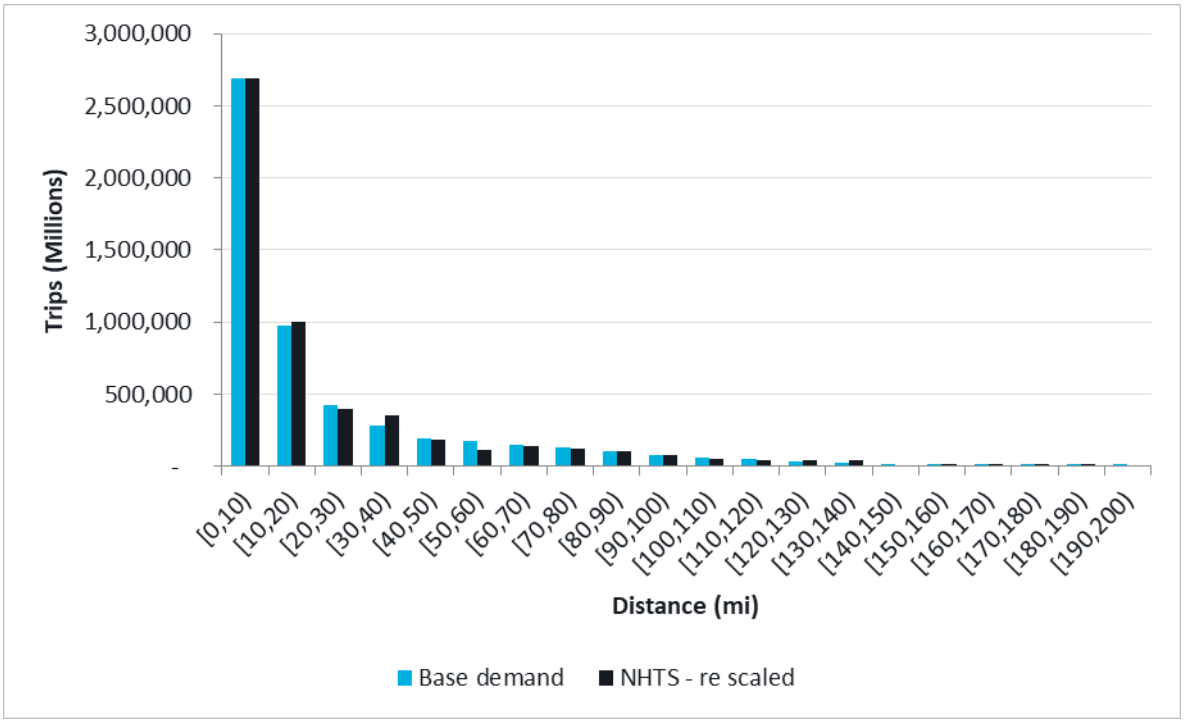
Figure 6-5: Linear correlation between county-county base demand and NHTS-scaled data



Source: Steer, 2021

As expected, the trip-length distribution is almost identical to the base demand. Figure 6-6 shows the trip-length distribution at a more disaggregated level than the adjustment, which accounts for the differences.

Figure 6-6: Trip length distribution of base demand and NHTS scaled data.



Source: Steer, 2021

Model estimation

Trip productions

Cross-classification analysis is used to estimate trip productions. Different studies for both urban and long-distance demand forecasting have identified common transferable parameters to be used as part of cross-classification such as household size, household income, number of household workers and number of vehicles. Two-dimensional trip tables using household income as the first variable and household size as the second variable was implemented.

Different trip rates are estimated for home-based (68% of trips in NHTS) and non-home-based trips (32% of trips in NHTS). This differentiation is particularly important given that home-based trips are more likely to be commuter trips, while non-home based tend to be driven by different purposes, e.g., leisure and other.

Trips rates ($\boldsymbol{T_{pk}}$) were estimated for each trip purpose $\boldsymbol{p}$ (e.g., home-based commute, home-based leisure, non-home-based, etc.) and household category $\boldsymbol{k}$, which is a combination of the cross-classification variables (e.g., low-income and size 1 households). The trip rates are a ratio between the number of trips identified ($Ntrips$) and the number of persons within households (NH) for the particular $\boldsymbol{p}$ and $\boldsymbol{k}$ subsets, obtained from the NHTS data analysis:

$$T_{pk} = \frac{Ntrips_{pk}}{NH_k}$$

Segmentation of income was defined consistently with the categories used within our behavioral analysis, namely:

- Low-income households: Less than $50,000
- Mid-income households: $50,000 - $100,000
- High-income households: More than $100,000

In terms of household size segmentation, the following segments showed the most consistent trips rates based on the available data:

- 1 person households
- 2 person households
- 3 person households
- 4 or more person households

The resulting trips rates by purpose are shown in the following tables. As it can be seen, non-home-based trip rates were created as a single rate per purpose, as opposed to cross-classification approach, since there were not enough data points to provide consistent trip rates across different classes.

Table 6.5: Average daily trip rates for home-based Commute purpose

Income/HH size	1	2	3	4+
Less than $50,000	0.0388	0.1092	0.1465	0.1805
$50,000 - $100,000	0.0755	0.0797	0.0922	0.1016
Greater than $100,000	0.0572	0.0779	0.0939	0.0928

Source: Steer, 2021

Table 6.6: Average daily trip rates for home-based Business purpose

Income/HH size	1	2	3	4+
Less than $50,000	0.0024	0.0023	0.0051	0.0065
$50,000 - $100,000	0.0039	0.0036	0.0061	0.0023
Greater than $100,000	0.0042	0.0049	0.0059	0.0050

Source: Steer, 2021

Table 6.7: Average daily trip rates for home-based Leisure purpose

Income/HH size	1	2	3	4+
Less than $50,000	0.0493	0.0539	0.0442	0.0427
$50,000 - $100,000	0.0535	0.0583	0.0597	0.0351
Greater than $100,000	0.0431	0.0560	0.0505	0.0447

Source: Steer, 2021

Table 6.8: Average daily trip rates for home-based other purpose

Income/HH size	1	2	3	4+
Less than $50,000	0.0645	0.0698	0.0753	0.0649
$50,000 - $100,000	0.0419	0.0515	0.0532	0.0476
Greater than $100,000	0.0319	0.0419	0.0542	0.0508

Source: Steer, 2021

Table 6.9: Average daily trip rates for non-home-based all categories

Commute[75]	Business	Leisure	Other
0	0.00114	0.00129	0.00237

Source: Steer, 2021

Trip attractions

To estimate trip attractions, we used regression analysis. As part of the socioeconomic analysis for the CRRM model building, different data has been collected for the state of California, at different levels of granularity (some at a Census Tract, some at a zone level). The regressions are built based on the most aggregated geographic representation within these datasets for consistency.

Usually, trip attraction models are based on linear regressions using descriptive variables available for the geographic areas. For this model the available data is population, household characteristics (number, size, income level, workers composition), employment by occupation and school enrollment.[76] The regression model estimates, for each purpose, the coefficients of the descriptive variables (i.e., C_r^p of equation below). Therefore, the attracted trips can be obtained as follows:

$$Nattr_j^p = \sum_r C_r^p * v_{rj}$$

Where:

- $Nattr_j^p$ is the number of trip-ends attracted for the purpose $\boldsymbol{p}$ in geographic area $\boldsymbol{j}$
- C_r^p is the estimated coefficient for descriptive variable $\boldsymbol{r}$ (e.g., population) and trip purpose $\boldsymbol{p}$
- v_{rj} is the value of descriptive variable $\boldsymbol{r}$ in geographic area $\boldsymbol{j}$

[75] Even though the processing resulted in very few commute non-home-based trips, these were manually assigned to "Business," as most probably were miscategorized.

[76] Provided in the 21 North American Industrial Classification System (NAICS) categories.

The variables tested within the estimation is a compilation of 52 descriptive variables for the 1,181 zones, based on the aggregation of data provided at Block level and the geographic relation between the CRRM Zoning system and 2019 TIGER/Lines shapefiles.[77]

Given the number of variables available, the following methodology was used to select the independent variables used in the regression analysis:

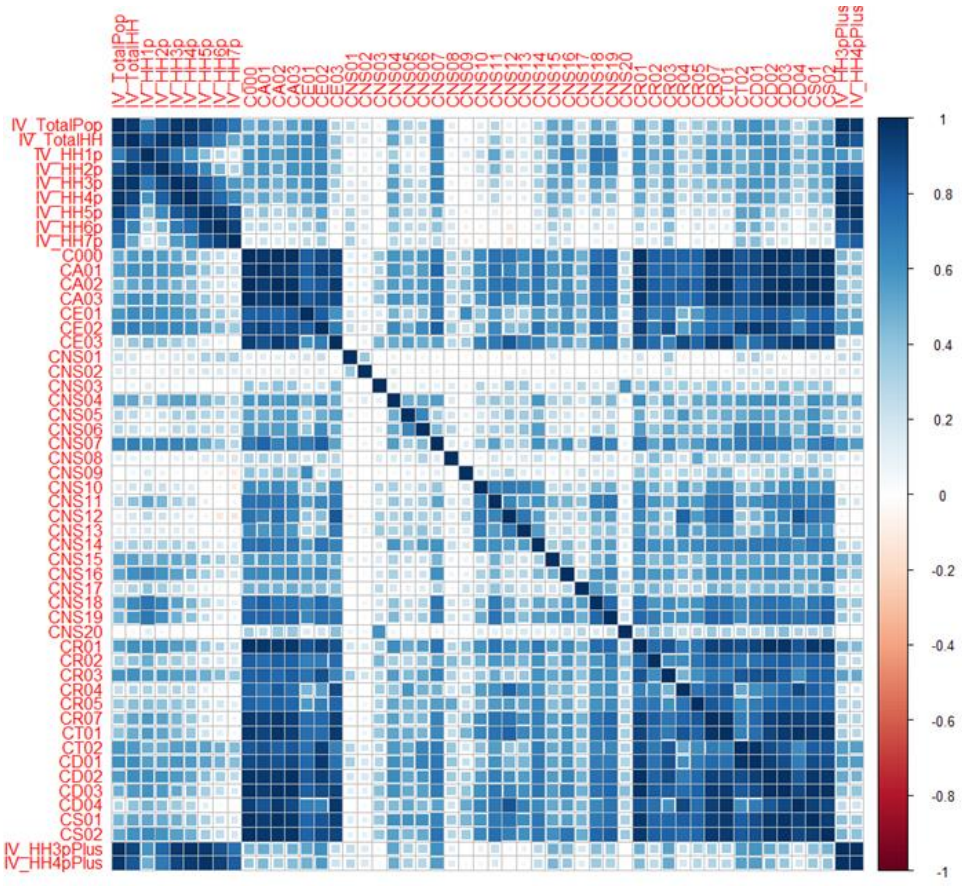
1. Estimated the linear correlation between variables.
2. Estimated single-variable regression.
3. Based on 1 and 2, selected a subset of (maximum) 4 lowly-correlated variables,[78] starting with the variables exhibiting best goodness of fit (i.e., R2 and t-test results) in the single-variable regression.
4. Ran the multi-variate regressions and sort the results based on R2.
5. Checked the reasonableness of the independent variables based on expert knowledge; and kept only expressions where all the independent variables are statistically significant.
6. Selected the best fit for each purpose.

Figure 6-7 shows the linear correlation between the different variables. As expected, the number of households and derivatives are highly correlated between them and with respect to total population and total households. This is also the case between the different classes of employment (by age, by income, by industry, etc.).

[77] See www.census.gov/geo/maps-data/data/tiger-line.html for more information on the TIGER/Line data products.

[78] Cutoff of 0.5 Pearson correlation is used as ample boundary.

Figure 6-7: Correlation analysis of socioeconomic variables



Source: Steer, 2021

Even though multicollinearity does not necessarily imply worse prediction value in linear regressions, it does affect the readability of the results and result in spurious conclusions regarding the data. Therefore, for the selection of variables included only have variables whose correlation is below 0.5.

A total of 91 combinations of models among the different purposes were tested. For Leisure and Other purposes, the correlation cutoff of 0.5 was relaxed to 0.55 to incorporate variables that might represent better the travel behaviors. An additional metric was included to guarantee low multicollinearity: variance inflation factor (VIF)[79]. A literature rule-of-thumb of VIF<10 was used to test all the regressions.

The models selected and main statistics per independent variable are show in Table 6.10.

[79] For more information visit http://www.how2stats.net/2011/09/variance-inflation-factor-vif.html

Table 6.10: Attraction model parameters and results

Purpose	IV	Description	Estimate	Intercept	Std. Error	t-value	Pr(>|t|)	R2	VIF
Commute	CA02	Number of jobs for workers aged 30 to 54	0.1092	39	0.0045	24.04	1.075E-94	0.440	N/A
Business [80]	CA02	Number of jobs for workers aged 30 to 54	0.0044	0	0.0002	25.25	8.906E-26	0.942	N/A
Leisure	CNS18	Number of jobs in NAICS sector 72 (Accommodation and Food Services)	0.3654	452	0.0380	9.63	9.94E-21	0.192	1.41
Leisure	CNS17	Number of jobs in NAICS sector 71 (Arts, Entertainment, and Recreation)	0.2284		0.0853	2.68	0.0075	0.097	1.41
Other	CNS07	Number of jobs in NAICS sector 44-45 (Retail Trade)	0.7242	0	0.0202	35.81	2.748E-164	0.230	N/A

Source: Steer, 2021

The results of the attraction models show rather weak goodness of fit at a TAZ level. The main reason for these results is the lack of trip data points for the in-scope trips at every zone. For the purpose of the CRRM model, the results are found to be acceptable mainly because the distribution model is singly constrained for most of the purposes, except commute; and therefore, it relies more on the production end than the attraction end.

Furthermore, it is found that at the county-level the attraction model exhibits acceptable model fitness. The next section covers these results.

Model validation

Using the production trips rates, a validation process was carried out to find the estimated trip productions per zone, where zonal properties (such as household descriptive totals by zone) are taken from the Synthetic Population description. The number of produced trips in a zone $\boldsymbol{i}$ is given by:

[80] Business purpose model was estimated at the County level, since not enough data was available to predict at the TAZ level.

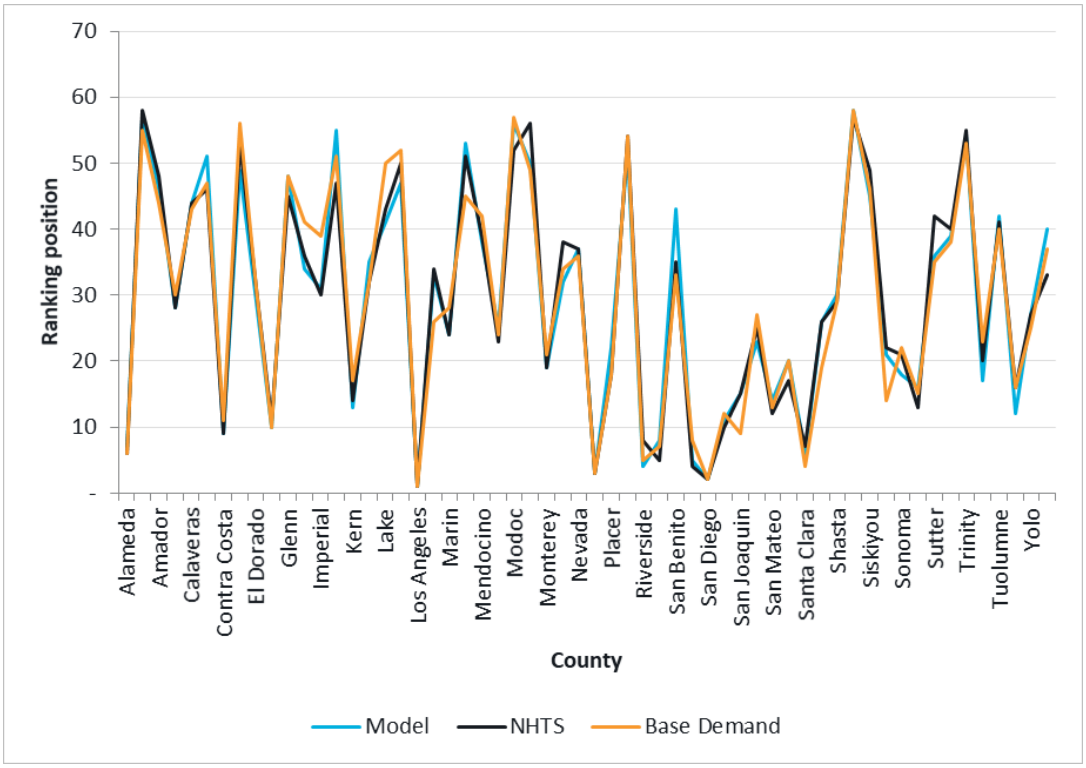
$$Nprod_i^p = \sum_k T_k^p * h_{ik}$$

Where:

- $Nprod_i^p$ is the number of trip ends produced for the purpose $\boldsymbol{p}$ in zone $\boldsymbol{i}$.
- T_k^p are the trip rates estimated through cross-classification from NHTS, for purpose $\boldsymbol{p}$ and household category $\boldsymbol{k}$.
- h_{ik} is the number of people in zone $\boldsymbol{i}$ within households that belong to category $\boldsymbol{k}$.

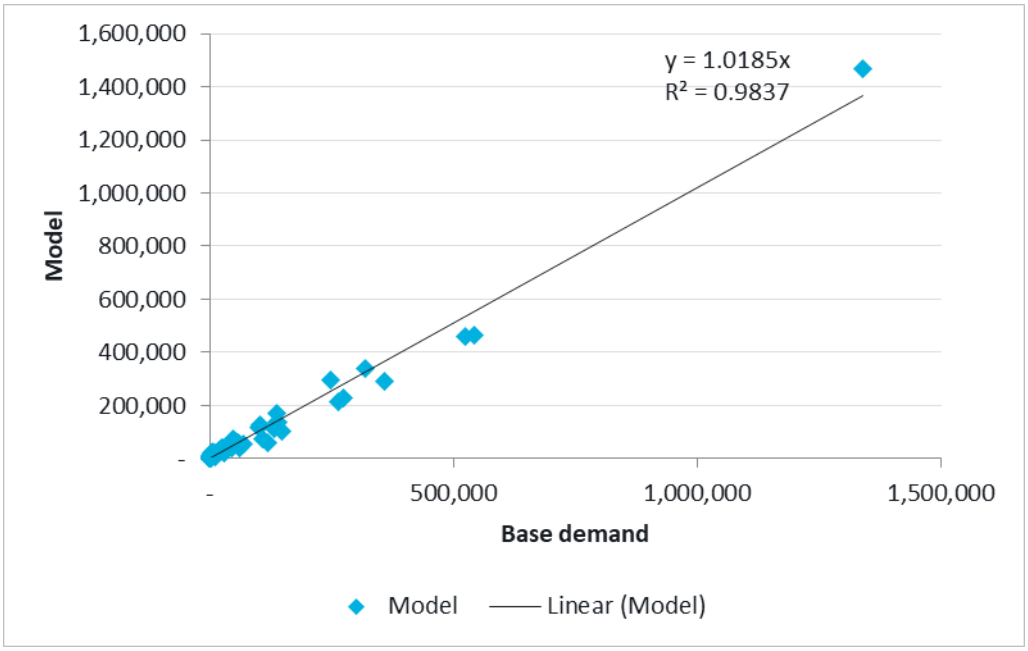
The results from the modeled trips show a good approximation to production both in terms of relative numbers, as shown by the ranking per county in Figure 6-8; as well as absolute numbers, shown by the linear comparison in Figure 6-9.

Figure 6-8: Comparison of county trip production ranking for modeled, NHTS and Base demand



Source: Steer, 2021 (note not all counties are shown on the x axis simply due to the fit on the chart)

Figure 6-9: Linear correlation of estimated trips produced by county and Base demand.

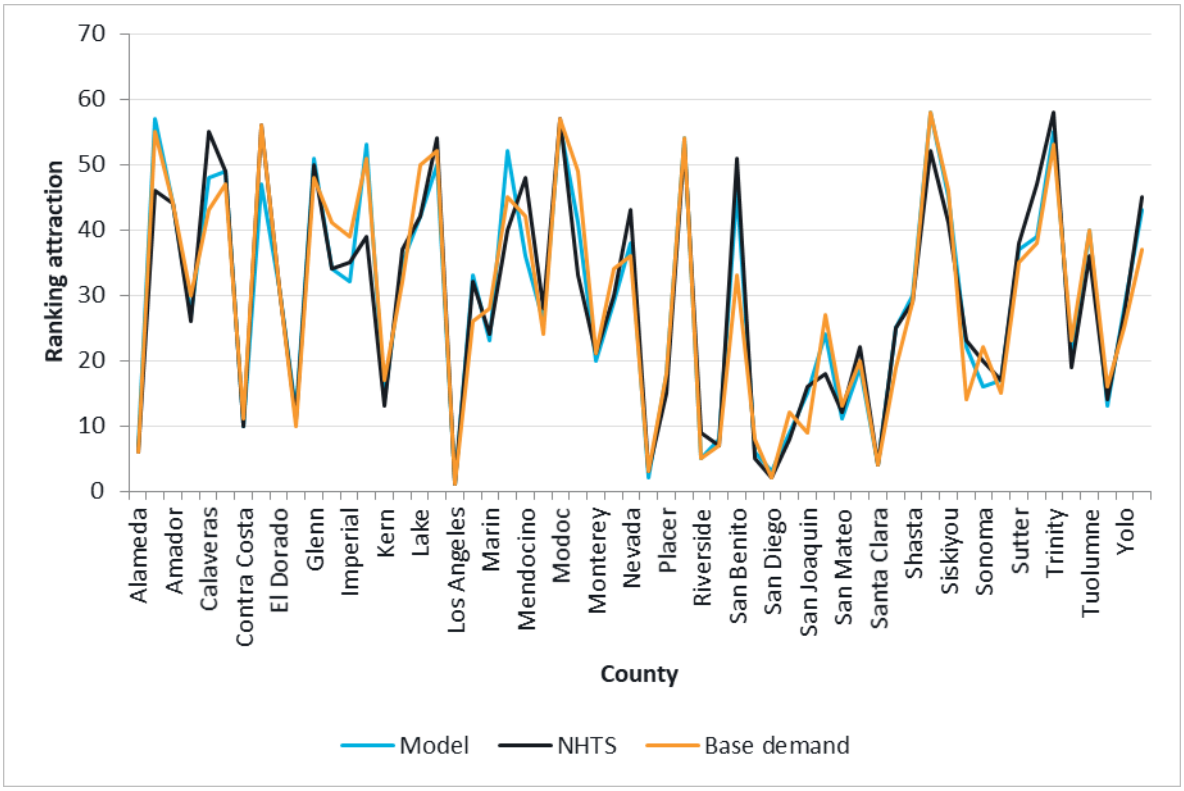


Source: Steer, 2021

Similarly, the total attracted trips were estimated using the resulting linear functions from the regression analysis and validated against the Observed Trip Tables. The comparative ranking of attracted trips is shown in Figure 6-10. As it can be seen, for most of the counties, the attraction trip ends provide similar relative results.

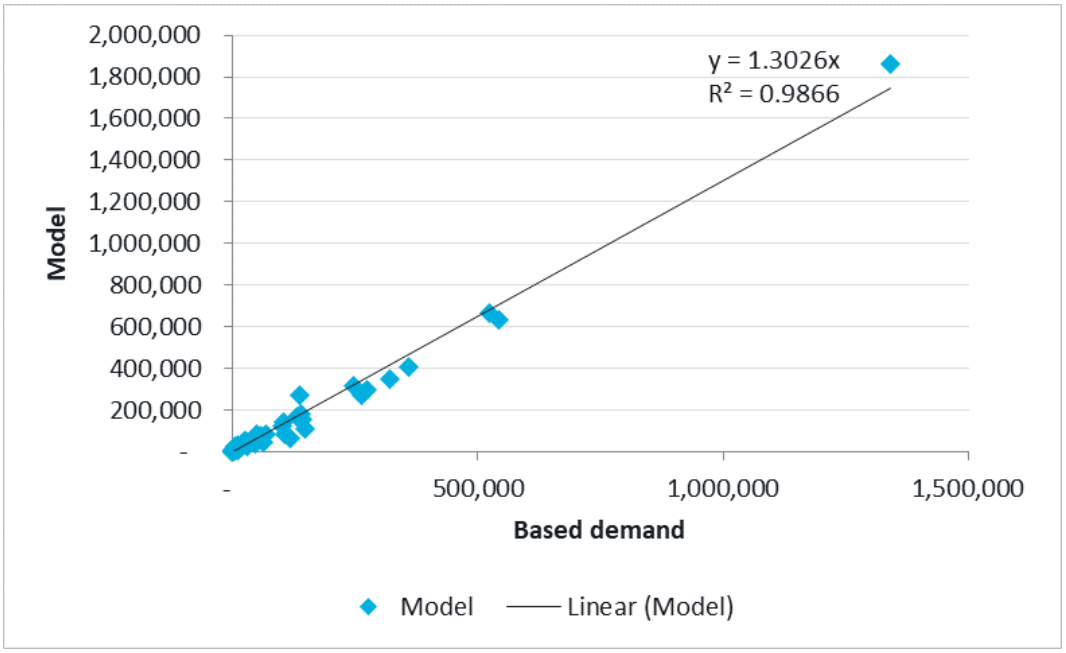
Figure 6-11 shows the linear correlation between estimated attracted trips and base demand. First, the R^2 tells us that the fitness of the model is very good at the county level. Second, by looking at the slope (1.3026) it can be concluded that the total of attracted trips estimated is higher than the observed. However, given that the relative dimensions with respect to other counties has a good fit, these models are considered to be appropriate. A global scaling factor to decrease the total estimated trips can be incorporated to address the difference.

Figure 6-10: Comparison of county trip attraction ranking for modeled, NHTS and Base demand.



Source: Steer, 2021 (note not all counties are shown on the x axis simply due to the fit on the chart)

Figure 6-11: Linear correlation of estimated trips attracted by county and Base demand.



Source: Steer, 2021

Limitations/Caveats

- The outbound and return process consisted of a list of rules applied to the full NHTS trips dataset, considering the most reasonable processing assumptions based on analysis and professional judgement. However, there are a small percentage of exceptions to these rules, which are difficult to trace given the size of the dataset.
- Given the extent and scope of the NHTS, long-distance movements are underrepresented. In addition, daily trip making behavior is mostly dominated by short-distance trips. So, even if the intra-zonal trips are removed, trips rates will most likely be biased by short-distance movements. To correctly assign the generated trips, targets were reweighted based on base demand.
- Population information by zone and by intersection between household characteristics was not readily available from Census Bureau. Instead, synthetic population from the CSTDM model was updated with targets from the Census Bureau regarding each individual break down, at different levels of granularity – specifically, household characteristics (size, income) distribution by location, total population by age, distribution employment status by age and location.
- The descriptive socioeconomic variables defined the level of granularity provided by the attraction models. Although the model was built on a zone-to-zone basis, the accuracy of the model estimation is mostly at a county level due to the lack of data points of available source data, regarding in-scope trips.
- Even though attraction models are not extremely accurate at the zone level, these results are used loosely in the distribution model for business, leisure and other purposes, given that these are singly constrained and will rely more strongly on the production results.
- This approach does not include car ownership as a descriptive variable since it is assumed to be at saturation point for California.

Distribution model approach

Overview

The Trip Distribution model estimates the number of trips that occur between each origin zone and each destination zone. These trips are generated based on the estimated trip productions (or origins) and trip attractions (or destinations) from the Trip Generation model. In the CRRM, the trip productions and attractions are stratified by trip purpose; therefore, the trip distribution produces trip tables by trip purpose.

This distribution of trips among destinations is done by a “Gravity model,” which is a form of a deterrence function that disincentivizes travel as distance, time or costs increase.

The output from the trip distribution is validated against the observed trips by comparing the patterns at the county-to-county levels and the trip lengths of the modeled and the observed data. While the trip ends from the first step of the model (trip generation) represents relatively robust estimates, it is acknowledged that the distribution of those trips is harder to estimate, particularly at a statewide level. The reason for this is that trip distribution (i.e., choice of destination) by its nature is based on many factors – not simply travel costs.

Sources of data

Data from NHTS was used to estimate the trip distribution model. A combination of national (2016-2017) and California Add-on data was used to assess the travel behavior of residents of California. There were 26,095 household surveyed in California and their travel was expanded to the entire population of the state. The NHTS is fairly comprehensive, covering people across demographics and geography, hence the behavior of people observed is considered to provide a good representation of the population. There are always outliers in the data, and these outliers cannot be captured in the models; the models are built for representing general observed behaviors.

To validate the origin destination, travel patterns from the trip distribution process, observed trips between zones are compared with those modeled. This observed data is described in the relevant base demand appendices.

Methodology

There are various trip distribution methods used in travel demand models. Most common among them are **Growth Factor Method** and **Gravity Method** models. While the growth factor method is based on relative growth of the origins and destination, it has limitations when there are significant changes in the destination choice over time and hence can only reasonably be used for short term forecasting. The gravity model method on other hand is more generic and can be extended to more long-term forecasting as the model accounts for changes in destination choice based on the travel costs, changes in housing and economic activity locations, and travel choices over time. Therefore, in this project we decided to use a gravity model approach for the trip distribution.

Gravity model

Trip generation tables are used to determine how many trips access the network at each TAZ and how attractive those TAZs are. Trip generation tables are grouped in five main trip-distribution segments: Commute, Business, Leisure, Other and Non-home based. Cost functions (i.e., impedance functions) were established for each trip-distribution segment based on average daily composite impedances skimmed from the scenario service data inputs.

The gravity model estimates production-attraction trip tables from zone level estimates of productions and attractions and travel cost between the zones. The trips assigned between two zones are directly proportional to the productions in the originating zone and to the attractions in the destination zone and inversely proportional the travel impedance or deterrence between the zone pairs. The deterrence can be directly the distance between zones, or travel time, or a more complex function. In most sophisticated models, the impedance or deterrence has a logistical function which represents the travel impedance as a function of a combination of travel distance, cost, and time.

There are many formulations of a gravity model: they can be constrained by productions of zones, attractions of zones, or doubly constrained by both productions and attractions.

$$T_{ij} = \alpha\, O_i\, D_j\, f(C_{ij})$$

Where:

- T_{ij} is the number of trips traveling between zones ***i*** and ***j***.
- O_i is the number of trips produced in zone ***i***.
- D_j is the number of trips attracted by zone ***j***.
- $f(C_{ij})$ is a generalized function of the travel costs.
- α is a calibration parameter.

The deterrence function is based on the travel cost or time between zones and has an inverse relationship with the travel cost between a zone pair. The frequency of trips between a zone pair increases with lower cost and decreases with higher cost. This function is estimated for each trip purpose because trip purposes have different sensitivities to cost and perceived costs of travel. For example, the frequency of leisure trips is more likely to be impacted by a given cost change than that of commuting trips. Due to its incremental nature, it can either be implemented with an incremental logit form or as an elasticity to changes in service. We reviewed the deterrence functions used in the CSTDM model – the model used by Caltrans for long-distance highway forecasting – and used the costs from the base year assignment for calibration purposes, adjusting from this base position as required. We calibrated the synthetic matrices to match as close as possible the trip patterns in the observed demand matrices.

For all trip-distribution segments, a specific cost function was estimated based on a power function form, as follows:

$f(C_{ij}) = \gamma\ {C_{ij}}^{-\beta}$ Where:

- C_{ij} is the *logsum* of the average daily composite impedance for traveling from i to j across all available modes.
- γ and β are calibration parameters.

Calibration of the trip distribution model happens recursively. The gravity model which is formulated for estimation of the trips between OD pairs has parameters which are calibrated based on a comparison of the trip lengths between the modeled and observed trips by trip purpose. The deterrence function is estimated to match the trip length frequency distributions by purpose using observed data and other regional travel surveys.

After the trip distribution model is calibrated, the county-to-county modeled trip tables are validated against the observed trip tables by adjusting the location-specific and county-to-county K-factors (discussed below) to modify the synthetic distribution to better represent the attractiveness of different zones and thus match better the observed distribution. This step also considers the impact of special generators and physical traffic impeding factors like bridges or tunnels.

Gravity model estimation

As discussed earlier, separate gravity models were developed for each trip purpose – Commute, Business, Leisure, Other and Non-Home-Based purposes. To calibrate the gravity model, we compared the trip length frequency distribution between the modeled PA trips produced from distribution step with the observed trip length frequency distribution from the observed data

Table 6.11 shows the estimated values of Gamma and Beta in the power function for each trip purpose.

Table 6.11: Gravity model estimation parameters

Purpose	Γ	β
Commute	1	3.1
Business	1	1.9
Leisure	1	1.7
Other	1	3.2
Non-Home Based	1	1.9

K-factors

K-factors correct for residual differences in trip distribution, usually at the district level. The trip distribution models generally do not account for many factors that might affect the destination choice of people and K-factors can be used in those cases. Sometimes the travel costs between two zone or district pairs are not accurate in calculations as many other factors might contribute to the attractiveness of a zone other than travel costs. K-factors are intended to account for the choice factors that are not able to be included in the models. Since trip distribution models have relatively few input variables, it is reasonable to believe that other factors that affect location choice are not included in the models. In many cases they cannot be measured, quantified, or forecasted. K-factors provide a means for accounting for these factors, although they are then assumed to remain fixed over time and across all scenarios.

In the model, apart from capturing county level travel patterns, the K-factor estimation also accounts for the special generators like convention center's, hotel clusters, and sporting event venues. The intra zonal trip assignment and zone pairs with greater or fewer trips due to factors not accounted for in the distribution model are represented by the K-factors. The model also checks travel impedances between zone pairs as a part of K factor calculations.

Induced demand

Induced demand is included in the model at the OD level, with the demand applied to HSR, rail and combo only (as all other modes are assumed not to change).

The change in utility at the distribution model stage is used to estimate an induced demand factor through the application of an elasticity, as follows:

$$Induced\ demand\ factor = \left(\frac{Test\ utility}{Comparator\ utility}\right)^{Elasticity}$$

The elasticity values are shown in the table below for each purpose.

Table 6.12: Induced Demand Elasticity

Purpose	Elasticity
Commute	0
Business	-0.5
Leisure	-0.5

Purpose	Elasticity
Other	-0.5
Nonresident and External	-0.5

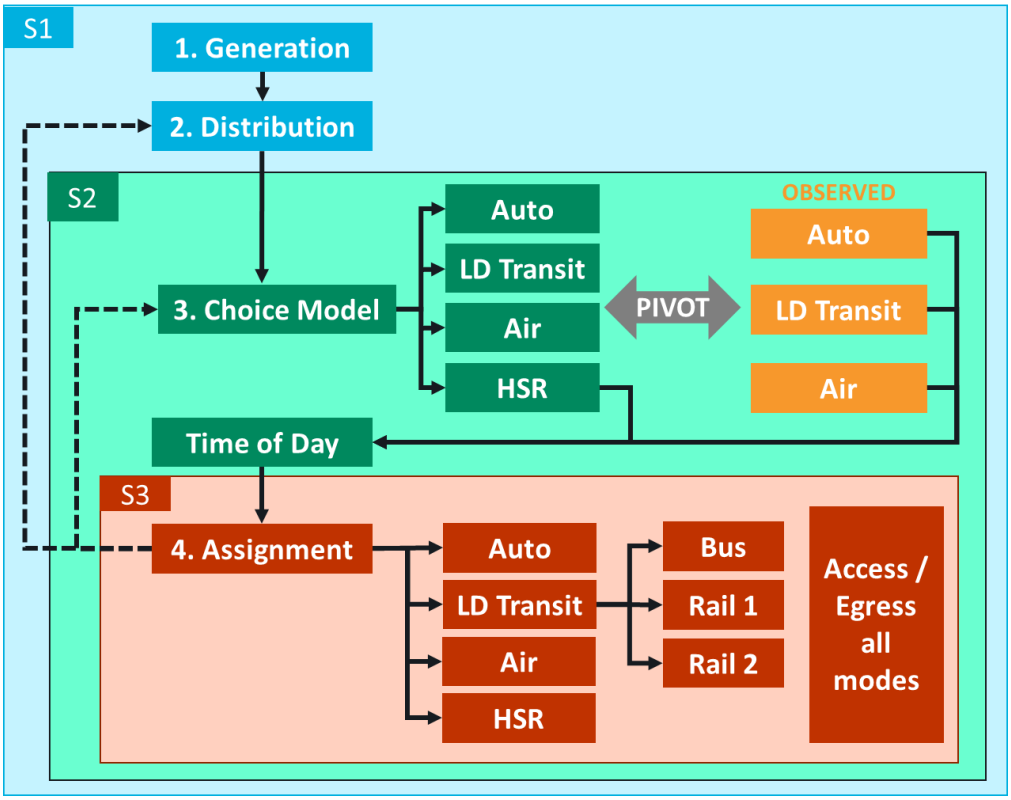
The total induced demand is then spread across rail/combo/HSR based on the share of induced within these three modes.

7 Choice Modeling

Introduction

The California Rail Ridership model (CRRM) is a 4-Step travel demand model. Figure 7-1 shows the flow chart of the steps for the model. This section, will briefly discuss the choice modeling step of the CRRM.

Figure 7-1. CRRM Model flow chart



The choice modeling step of the CRRM estimates the choices in selecting or changing the mode to complete the trip. The mode choice includes auto, HSR, long-distance transit and air. The choice model represents the impact of travel costs on travel behavior decisions. The structure of the choice model has been determined through our behavioral research (Stated Preference survey). The inputs to the choice model are demand from the distribution step and costs from the assignment step.

Stated preference survey

As part of the choice model development, Steer conducted online stated preference (SP) surveys between August 21st and October 2nd, 2020. Data was collected from a paid online panel provided by a third-party vendor Dynata and through the email contact databases of the California High-Speed Rail Authority. Of the final 7,014 respondents, 4,336 remained usable after the data was filtered and cleaned. Filters included survey completion times and answer consistency tests. A companion report[81] details the results of the individual survey questions, the weighting methodology used with the filtered sample, and the results for the weighted sample.

Choice model development

Model structure and formulation

The model structure is a multinomial choice model, with a 0.5 nesting parameter for distribution and it models the choice between combinations of main mode and the access and egress modes. The main modes are auto, bus, rail, air, high-speed rail and combo, where combo being a multi-mode travel option of bus/rail/high-speed rail modes. The access/egress modes in the model are car, taxi, and local transit (e.g., subway, ferry).

The MNL model estimates the coefficients for following parameters:

- Main modes
  - In-vehicle time (IVT) coefficient by purpose
  - Cost coefficient by income and employment
  - Modal constants by urban/rural split
  - Delay time coefficient (to account for reliability)
  - Transfer penalties (for multi-step)
- Access & egress modes
  - Access and egress time coefficients
  - Access and egress mode constants (transit and TNC)

Standard formulation of the mode choice utility equation is as follows:

$$V_i = mode\ constant_i + access\ mode\ constant + egress\ mode\ constant + b_{ivt} * IVT + b_{cost} * cost + b_{access} * access\ time + b_{egress} * egress\ time + b_{delay} * delay\ time + b_{transfer} n_{transfers}$$

Main Mode

Table 7.1 shows the utility equations for mode choice by main modes. The equation for auto, transit and combo modes are discussed in the Table. For auto as a main mode the access and egress does not apply, hence the utilities do not consider and access egress costs. For transit including rail, air, bus and high-speed rail has all mode specific coefficients including access and

[81] California rail ridership modeling: stated preference survey summary report, Steer, February 2021

egress. The cost function for the combo mode is more complicated due to multiple main modes assigned for the trip.

In these equations [mode] is shown, substitute in the parameter name with the mode specific parameter. For example, if air is the main mode, use “a_air_urban” where the formula has “a_[mode]_urban”.

Table 7.1: Mode Choice Utility Equations

Mode	Equation
Auto	V_[mode] = beta_ivt_business * [mode]_travel_time * (purpose == 1) + beta_ivt_commute * [mode]_travel_time * (purpose == 2) + beta_ivt_leisure * [mode]_travel_time * (purpose == 3) + beta_ivt_other * [mode]_travel_time * (purpose == 0) + beta_cost_employed_low_income * [mode]_total_cost * (income_class == 1) * (employment == 1) + beta_cost_employed_medium_income * [mode]_total_cost * (income_class == 2) * (employment == 1) + beta_cost_employed_high_income * [mode]_total_cost * (income_class == 3) * (employment == 1) + beta_cost_student * [mode]_total_cost * (employment == 2) * (income_class > 0) + beta_cost_retired * [mode]_total_cost * (employment == 3) * (income_class > 0) + beta_cost_homemaker * [mode]_total_cost * (employment == 4) * (income_class > 0) + beta_cost_emp_other * [mode]_total_cost * (employment == 5) * (income_class > 0)
Rail/Air/ Bus/HSR	V_[mode] = a_[mode]_urban * (urban == 1) + a_[mode]_rural * (urban == 0) + dist_[mode] * dist_fct_[mode] + beta_ivt_business * [mode]_travel_time * (purpose == 1) + beta_ivt_commute * [mode]_travel_time * (purpose == 2) + beta_ivt_leisure * [mode]_travel_time * (purpose == 3) + beta_ivt_other * [mode]_travel_time * (purpose == 0) + beta_cost_employed_low_income * [mode]_total_cost * (income_class == 1) * (employment == 1) + beta_cost_employed_medium_income * [mode]_total_cost * (income_class == 2) * (employment == 1) + beta_cost_employed_high_income * [mode]_total_cost * (income_class == 3) * (employment == 1) + beta_cost_student * [mode]_total_cost * (employment == 2) * (income_class > 0) + beta_cost_retired * [mode]_total_cost * (employment == 3) * (income_class > 0) + beta_cost_homemaker * [mode]_total_cost * (employment == 4) * (income_class > 0) + beta_cost_emp_other * [mode]_total_cost * (employment == 5) * (income_class > 0) + utility_access + utility_egress
Combo	V_[mode] = (a_bus_urban * (urban == 1) + a_bus_rural * (urban == 0)) * multi_frac * Multi_Mode_Bus + (a_rail_urban * (urban == 1) + a_rail_rural * (urban == 0)) * multi_frac * Multi_Mode_Rail + (a_hsr_urban * (urban == 1) + a_hsr_rural * (urban == 0)) * (1-multi_frac) + (dist_bus_rural * (urban == 0) + dist_bus_urban * (urban == 1)) * dist_fct_bus * multi_frac * Multi_Mode_Bus+ (dist_rail_rural * (urban == 0) + dist_rail_urban * (urban == 1)) *

Mode	Equation
	dist_fct_rail * multi_frac * Multi_Mode_Rail + (dist_hsr_rural * (urban == 0) + dist_hsr_urban * (urban == 1)) * dist_fct_hsr * (1-multi_frac) + beta_ivt_business * [mode]_travel_time * (purpose == 1) + beta_ivt_commute * [mode]_travel_time * (purpose == 2) + beta_ivt_leisure * [mode]_travel_time * (purpose == 3) + beta_ivt_other * [mode]_travel_time * (purpose == 0) + beta_cost_employed_low_income * [mode]_total_cost * (income_class == 1) * (employment == 1) + beta_cost_employed_medium_income * [mode]_total_cost * (income_class == 2) * (employment == 1) + beta_cost_employed_high_income * [mode]_total_cost * (income_class == 3) * (employment == 1) + beta_cost_student * [mode]_total_cost * (employment == 2) * (income_class > 0) + beta_cost_retired * [mode]_total_cost * (employment == 3) * (income_class > 0) + beta_cost_homemaker * [mode]_total_cost * (employment == 4) * (income_class > 0) + beta_cost_emp_other * [mode]_total_cost * (employment == 5) * (income_class > 0) + utility_access + utility_egress

Access/egress approach

Access and egress are a key component in the choice model. Access and egress are considered separately, with the choice model considering the nine permutations of transit, auto and taxi/TNC.

The SP work included a focus on access and egress, with a view to ascertain the relative perception of access/egress relative to the main mode, as well as understanding if and how there was any impact on their time. The formulation adopted is set out below and includes a greater perceived time as the actual IVT exceeds 35 minutes.

The utility function for access and egress part of the choice model is based on the travel distance and the formulation is as follows:

$$Utility_{acc/egr} = \min\left\{\beta_{ivt_{mainmode}} * (t_{ivt} + t_{walk+wait}), P_{\frac{acc}{egr}} + t_{walk+wait} * \beta_{ivt_{mainmode}}\right\}$$

$$P_{acc/egr} = \begin{cases} t_{ivt} * \beta 1_{\frac{acc}{eg}}, & t_{ivt} < 25 \\ (t_{ivt} - 25) * \beta 3_{\frac{acc}{eg}}, +25 * \beta 1_{\frac{acc}{eg}}, & 25 \le t_{ivt} < 35 \\ {t_{ivt}}^2 * \beta 2_{\frac{acc}{eg}} + 30 * \beta 1_{\frac{acc}{eg}}, & t_{ivt} \ge 35 \end{cases}$$

$$\beta 3_{\frac{acc}{eg}} = \frac{1225 * \beta 2_{\frac{acc}{eg}} + 6 * \beta 1_{\frac{acc}{eg}}}{11}$$

Where:

- t_{ivt} is the Access/Egress IVT **real**
- $t_{walk+wait}$ is then Access/Egress walk and wait time **perceived**
- $\beta_{ivt_{mainmode}}$ is the Main mode IVT coefficient
- $\beta 1_{\frac{acc}{eg}}$ is the first access/egress coefficient
- $\beta 2_{\frac{acc}{eg}}$ is the second access/egress coefficient

The beta coefficients in the above formula have alternate parameter names, which are described in the following table. The access/egress mode is either “car” for auto or “nocar” for transit or TNC.

Table 7.2: Beta Coefficients

Beta Coefficient	Alternate Parameter Name
β_{ivt}	beta_ivt_[purpose]
$\beta 1_{acc}$	beta_access_time_[access_mode]
$\beta 2_{egr}$	beta_access_time_[access_mode]2
$\beta 1_{acc}$	beta_egress_time_[egress_mode]
$\beta 2_{egr}$	beta_egress_time_[egress_mode]2

Model Parameters

Modes

The main modes considered in the model are:

- Auto
- Bus (long distance bus, such as Greyhound)
- Air (also called flight)
- Rail (conventional rail)
- HSR
- Combo

The access and egress modes in the model are:

- Auto
- TNC (for example, taxi or uber)
- Transit (local transit)

Both TNC and transit are considered as “no car” for the purpose of which coefficients to use.

Input Skims

The following attributes are the input trip components that are skimmed from the network:

Parameter	Description
TripDistance	Distance in miles between origin and destination. For all modes, this is based on the auto distance.
[mode]_travel_time	Travel time for the main mode leg.
[mode]_total_cost	Total OD costs for the entire journey.
[mode]_access_time	Travel time for the access leg.
[mode]_egress_time	Travel time for the egress leg.

Distance Function

The dist_fct_[mode] is the distance function and it varies depending on the main mode. The formulas are as follows:

Main Mode	Distance Function
Auto	N/A
Bus	$ln\left(\frac{TripDistance}{100}\right)$
Air	$\left(\frac{TripDistance}{100}\right)$
Rail/HSR	$\left(1+e^{-a(b-c)}\right)^{-1}$ where: a = steepness_[mode] b = TripDistance (in miles) c = midpoint (200 miles)

Flags

The following are a list of flag attributes used in the choice model. The urban flag is OD dependent. The purpose, employment and income class flags are based on the demand segment.

Flag	Description
urban	Home region is urban (LA, SF, or SD) or rural. 1 = Urban 0 = Rural Note that for bus, there is no urban/rural split and the same coefficient (a_bus) is used for both in the formula.
purpose	Trip purpose

Flag	Description
	0 = Other 1 = Business 2 = Commute 3 = Leisure
employment	Employment category 0 = N/A 1 = Employed 2 = Student 3 = Retired 4 = Homemaker 5 = Other
income_class	Income class 0 = N/A 1 = Less than $50k 2 = $50k - $100k 3 = More than $100k

Other Parameters

The following describes the rest of the parameters in the model:

Variable	Definition
multi_frac	Fraction of multi that is rail or bus
Multi_Mode_Bus/Rail	1/0 or 0/1 if multi is bus+HSR or rail+HSR, respectively
[mode]_travel_time	In-vehicle time in minutes

Choice model coefficients

Two sets of choice model coefficients are used in the model. The first set has segment specific coefficients and are used for the first 20 segments that have different purposes, employment and income categories. The second set has segment agnostic coefficients and are used for the 21st segment (nonresident and external trips).

Note that the “beta_cost_emp_other” and “beta_delay_time” are not used in the current version of the model.

Table 7.3: Segment Specific Coefficients

Coefficient	Value
a_air_rural	-2.226708
a_air_urban	-2.029057
a_bus	-1.294277
a_hsr_rural	-0.192063
a_hsr_urban	-0.030169
a_rail_rural	-0.490289
a_rail_urban	-0.338035

Coefficient	Value
beta_access_time_car	0.00005
beta_access_time_car2	-0.000117
beta_access_time_nocar	-0.008522
beta_access_time_nocar2	-0.000043
beta_cost_emp_other	-0.018562
beta_cost_employed_high_income	-0.005998
beta_cost_employed_low_income	-0.0106
beta_cost_employed_medium_income	-0.010552
beta_cost_homemaker	-0.015844
beta_cost_retired	-0.005192
beta_cost_student	-0.014913
beta_delay_time	-0.006546
beta_egress_time_car	-0.011611
beta_egress_time_car2	-0.000034
beta_egress_time_nocar	-0.013302
beta_egress_time_nocar2	-0.000068
beta_ivt_business	-0.007752
beta_ivt_commute	-0.006931
beta_ivt_leisure	-0.005281
beta_ivt_other	-0.005539
dist_air	0.374009
dist_bus	0.248972
dist_hsr	-2.185973
dist_rail	-2.165587
steepness_hsr	-0.008107
steepness_rail	-0.005148

Table 7.4: Segment Agnostic Coefficients

Coefficient	Value
a_air_rural	-2.289005
a_air_urban	-2.163887
a_bus	-1.305698
a_hsr_rural	-0.262141
a_hsr_urban	-0.028879
a_rail_rural	-0.134095
a_rail_urban	0.019371
beta_access_time_car	0.000589
beta_access_time_car2	-0.000125

Coefficient	Value
beta_access_time_nocar	-0.009772
beta_access_time_nocar2	-0.000032
beta_cost	-0.012744
beta_delay_time	-0.006426
beta_egress_time_car	-0.011631
beta_egress_time_car2	-0.000018
beta_egress_time_nocar	-0.012714
beta_egress_time_nocar2	-0.000072
beta_ivt	-0.006059
dist_air	0.443091
dist_bus	0.202583
dist_hsr	-2.104333
dist_rail	-2.765088
steepness_hsr	-0.00826
steepness_rail	-0.00303

These parameters are applied to the 20 segments in the model as set out below.

Table 7.5: Segment specific coefficients

Segment	Purpose	Employment	Income	beta-ivt	beta_cost	beta cost value used	VOT (2018$)
1	Commute	Employed	Low	-0.00693	-0.02120	#N/A	$19.62
2	Business	Employed	Low	-0.00775	-0.01060	beta_cost_employed_low_income	$43.88
3	Leisure	Employed	Low	-0.00528	-0.01060	beta_cost_employed_low_income	$29.89
4	Other	Employed	Low	-0.00554	-0.01060	beta_cost_employed_low_income	$31.35
5	Commute	Employed	Middle	-0.00693	-0.02110	N/A	$19.71
6	Business	Employed	Middle	-0.00775	-0.01055	beta_cost_employed_medium_income	$44.08
7	Leisure	Employed	Middle	-0.00528	-0.01055	beta_cost_employed_medium_income	$30.03

Segment	Purpose	Employment	Income	beta-ivt	beta_cost	beta cost value used	VOT (2018$)
8	Other	Employed	Middle	-0.00554	-0.01055	beta_cost_employed_medium_income	$31.50
9	Commute	Employed	High	-0.00693	-0.01200	#N/A	$34.67
10	Business	Employed	High	-0.00775	-0.00600	beta_cost_employed_high_income	$77.55
11	Leisure	Employed	High	-0.00528	-0.00600	beta_cost_employed_high_income	$52.83
12	Other	Employed	High	-0.00554	-0.00600	beta_cost_employed_high_income	$55.41
13	Leisure	Student	All	-0.00528	-0.01491	beta_cost_student	$21.25
14	Other	Student	All	-0.00554	-0.01491	beta_cost_student	$22.29
15	Leisure	Retired	All	-0.00528	-0.00519	beta_cost_retired	$61.03
16	Other	Retired	All	-0.00554	-0.00519	beta_cost_retired	$64.01
17	Leisure	Homemaker	All	-0.00528	-0.01584	beta_cost_homemaker	$20.00
18	Other	Homemaker	All	-0.00554	-0.01584	beta_cost_homemaker	$20.98
19	Leisure	Other	All	-0.00528	-0.01856	beta_cost_emp_other	$17.07
20	Other	Other	All	-0.00554	-0.01856	beta_cost_emp_other	$17.90

Mode constant

The SP survey and choice model development explicitly allowed for the derivation of a mode constant for each main mode (using auto as the reference mode), a perceived value reflected model preferences of travelers beyond the actual time and cost of the trip. These were derived based on trip length, again using auto trip length as the reference.

The formulation of the constant value differs by main mode, as follows:

Rail/HSR: $$\beta_{constant} + \beta_{distance} * \frac{1}{1+\exp(-\beta_{steepness}*(distance-200)}$$

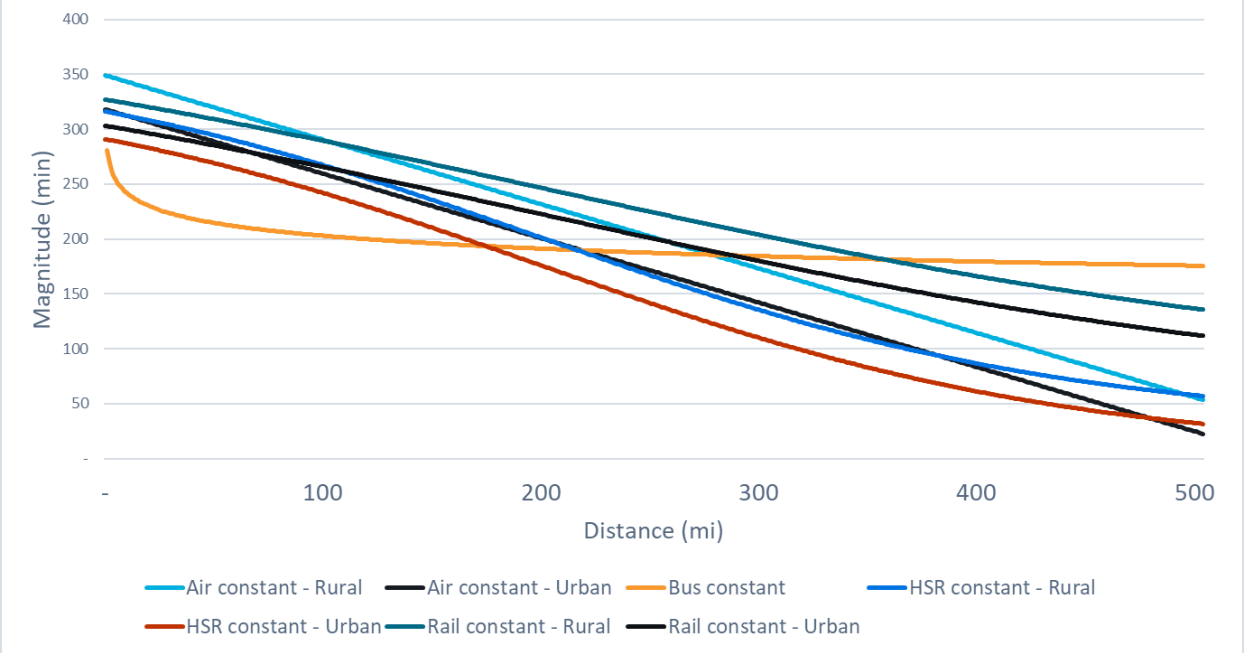
Bus: $$\beta_{bus\ constant} + \beta_{bus\ distance} * log\ \frac{distance}{100}$$

Air: $$\beta_{air\ constant} + \beta_{air\ distance} * \frac{distance}{100}$$

The constant for the combo mode is the sum of the constant for each mode within combo, plus a transfer penalty.

The resulting constant values (in minutes) by trip distance are illustrated in Figure 7-2. With auto as the reference mode, the constants are more than 200 minutes for trips less than 100 miles. All but bus reduce in value with longer distances, reaching 50-150 minutes at 500 miles; bus is relatively constant at or around 200 minutes for trips more than 100 miles.

Figure 7-2: Temporal mode constant values



Out of scope movements

To avoid the choice model apportioning demand to unrealistic options, a set of out-of-scope rules were developed. These review the distances of various journey elements and where they are

considered unrealistic, the option is deemed out of scope to ensure essentially zero share in the choice model.

In summary, the rules available are as follows, although not all are active by default:

1. Access + Egress distance > auto distance
2. Access distance > Main mode distance AND Egress distance > Main mode distance
3. Access + Main + Egress distance > 2x auto distance
4. Transfer distance > 1 mile
5. For Bus main mode only: access OR egress distance > 7.5 miles
6. For Bus main mode only: out of scope for employed/homemaker segments.
7. For Commute purpose: no auto as an egress mode (transit/TNC allowed)
8. For Other purpose: no auto as an egress mode
9. For Business/Leisure purpose: no auto as an egress mode
10. For Other/Leisure purpose: restrict egress modes based on a 20-mile cut-off.
11. Choice of rail/HSR vs combo to minimize unreasonable use of both.
12. For all purposes: restrict egress modes based on a 20-mile cut-off.
13. All non-auto trips below 10 miles out of scope
14. Non-auto trips below 20 miles out of scope if Access_+ Egress distance > 40% auto distance

If one or more of these rules is met, then the OD is out-of-scope for that specific mode under consideration.

Note, these rules only apply to the specific time period under consideration (i.e., it is possible for an OD to be in-scope in one time period but out-of-scope in another – most likely because the service level offered in time periods can differ). It is also possible for an OD to be out-of-scope in one direction, but to be in-scope in the reverse direction.

These rules are therefore applied separately by time period AND separately by direction.

The following provides an expanded definition of rule 11, given its complexity.

Rule 11

This rule only applies to the rail and combo modes in the base year, and to rail, combo and HSR in the future year (i.e., they do not apply to intercity bus, air or auto).

The primary purpose is to seek to ensure there is no unreasonable use of both rail/HSR and combo (noting that the combo mode, by definition, will also utilize either rail or HSR at least part of the way).

Definitions

- RA_D: Access distance for rail (i.e., distance from origin zone to access station).
- CA_D: Access distance for combo mode (i.e., distance from origin zone to access station).
- HA_D: Access distance for HSR (i.e., distance from origin zone to access station).
- RE_D: Egress distance for rail (i.e., distance from egress station to destination zone).
- CE_D: Egress distance for combo mode (i.e., distance from egress station to destination zone).
- HE_D: Egress distance for HSR (i.e., distance from egress station to destination zone).
- R_{OUT}: Rail is out-of-scope.
- C_{OUT}: combo mode is out-of-scope.

- H_{OUT}: HSR is out-of-scope.
- R_{IN}: Rail is in-scope.
- C_{IN}: Combo mode is in-scope.
- H_{IN}: HSR is in-scope.
- C_R: The combo-mode route only uses rail and bus (not HSR).
- C_H: The combo-mode route only uses HSR and bus (not HSR).
- C_B: The combo-mode route uses both rail and HSR (and potentially also bus).
- 30m: 30 miles.

Part 0

Base year	Future year (i.e., including HSR)
• R_{OUT} • C_{OUT}	• R_{OUT} • C_{OUT} • H_{OUT}

Designated as Part 0, because it is not really a rule – rather it is a statement reconfirming the validity of the other rules, i.e., if any of the modes (rail, combo or HSR) are out-of-scope based on prior rules, then they remain out-of-scope. This is re-stated since there are instances below which indicate a given mode is in-scope, but this is negated if it is already considered out-of-scope based on prior rules.

Part 1

Base year	Future year (i.e., including HSR)
• IF R_{OUT} THEN C_{IN}	• IF R_{OUT} AND H_{OUT} THEN C_{IN}

The logic here is that there cannot be double-counting if both rail and HSR (where applicable) are not in-scope, hence no further action is required regarding the combo mode.

Part 2

Base year	Future year (i.e., including HSR)
• IF R_{IN} THEN – IF RA_D <= 30m AND RE_D <= 30m THEN C_{OUT} – ELSE IF CA_D <= 30m AND CE_D <= 30m THEN C_{IN} AND R_{OUT} – ELSE C_{OUT}	IF C_R THEN • IF R_{IN} THEN – IF RA_D <= 30m AND RE_D <= 30m THEN C_{OUT} – ELSE IF CA_D <= 30m AND CE_D <= 30m THEN C_{IN} AND R_{OUT} – ELSE C_{OUT}
	IF C_H THEN • IF H_{IN} THEN – IF HA_D <= 30m AND HE_D <= 30m THEN C_{OUT} – ELSE IF CA_D <= 30m AND CE_D <= 30m THEN C_{IN} AND H_{OUT} – ELSE C_{OUT}

	IF C_B THEN • IF R_{IN} OR H_{IN} THEN – IF RA_D <= 30m AND RE_D <= 30m OR HA_D <= 30m AND HE_D <= 30m THEN C_{OUT} – ELSE IF CA_D <= 30m AND CE_D <= 30m THEN C_{IN} AND R_{OUT} AND H_{OUT} – ELSE C_{OUT}

The logic for each condition within the base year is as follows:

- If an attractive rail option exists (based on both the access and egress legs being relatively short), then the combo mode is not considered to be a viable option that people would consider (given people's aversion to transferring modes, especially if an attractive direct option is available to them).
- If an attractive rail option does not exist (based on one or more of the access/egress legs being long) but an attractive combo mode option does exist (based on both the access and egress legs being relatively short), then the combo mode is considered to be a viable option, but the rail mode is not.
- If an attractive option does not exist for either rail or the combo mode (based on one or more of the access/egress legs being long in each case), then neither are likely to be considered viable options for most people (and this should be shown in the outputs from the choice model). Given rail is already in-scope, however, the unattractive combo-mode option was eliminated to remove any risk of double-counting.

The logic for the future year is the same, except it applies to either rail, HSR or both depending on which of these modes are used within the combo mode routing for the OD under consideration:

- When the combo mode only uses rail (and bus), the conditions are identical to those used in the base year.
- When the combo mode only uses HSR (and bus), the conditions are the same as those used in the base year, except that it relates to HSR as opposed to rail.
- When the combo mode uses both rail and HSR (and potentially bus), then this considers both rail and HSR, with the combo mode being out-of-scope if either is considered to provide an attractive option, and similarly with both rail and HSR being out-of-scope if they are not attractive, but the combo mode is.

8 Model Calibration/Validation

Introduction

The preceding section set out the calibration and validation of the generation and distribution model. This next section sets out the validation of mode choice using the choice model described earlier, as well as the validation of the assignment of the final (post-pivot) modal matrices onto the network[82]. To date, the CRRM model validation has focused on mode and County level replication only, with no network level validation undertaken. Detailed results of earlier mode- and county-level validation are not repeated here. This section discusses the more detailed, validation undertaken for model refinement.

Approach

The focus of this effort was primarily the rail mode, with secondary review of air given its good data. There is little robust long distance bus data available, with the observed bus matrix derived using schedule and assumed load factors analysis and hence bus is not deemed a material mode to review further.

Given the 2018 Base CRRM uses a pivot process that essentially replicates the observed County-County demand by main mode, the focus is on the routing through the network and resulting demand at the station/airport level and the network level for all modes.

Rail

- **Observed rail station boardings and alightings** – observed station level data is available and was compared with modelled boardings and alightings. This was done across all stations, with a focus on the busiest (LA Union) and/or notable stations (Bakersfield). Observed vs. modelled are presented graphically as a scattergram, with goodness of fit statistics. Given the wide range of observed volumes, scattergrams are done for all stations keeping in mind that there is wide variation in stations volumes.
- **Service level demand** – boardings at the line level are compared with observed. This comparison was done at the individual route level for the inter-City routes operated by Amtrak (Capital Corridor, San Joaquin's, etc.) and the commuter routes operated by Caltrain and others.

Air

- **Observed airport boardings and alightings** – observed airport level data is available through the analysis of airport pair demand information used in the construction of the observed air travel demand matrices and are compared with modelled boardings and alightings. This is

[82] Based on Steer run 868.

done across all airports, with a focus on the busiest (such as LAX, SFO and SMF) and/or notable locations (such as Bakersfield, BFL). Observed vs. modelled can be presented graphically as a scattergram, with goodness of fit statistics.

- **Air link demand** – air demand used in the CRRM is based on FAA route level demand data and hence can be processed to derive link level demand. These can then be reviewed individually and observed vs. modelled can be presented graphically as a scattergram, with goodness of fit statistics.

Model improvements

Model refinements that were implemented in CRRM v5 versus CRRM v4 to improve validation and improve the model fit included the following:

- Use of a 5-minute transfer penalty on local transit for access and egress, and between main mode and access / egress (all modes). This lower value for transit access and egress transfer reflects the greater regularity and accommodation of transfers than typical on main modes.
- Statewide average $2.38 transit fare for access and egress trips was added to capture the cost associated with using local transit.
- Addition of a standard airport time of 45 minutes, reduced as applicable by an attractiveness benefit for FAA defined large and medium size airports (large-large 20 minutes, large-medium 5 minutes). The attractiveness benefit reflects how air travelers favor mainstream airlines between dominant hub airports to capture loyalty and frequent flyer privileges, as well as the greater range of alternatives when delays and cancellations occur.
- Inclusion of auto tolls (using an equivalent average time penalty) for Bay Area bridges and the tolled freeways in Orange and San Diego counties.
- The auto base demand matrix was re-adjusted within the Central Valley Region. It was observed that the share of inter-county auto trips was much higher than intra-county trips, while the observed LBS data from Streetlight showed that the larger share of auto trips in the Central Valley are intra-county. Hence the shares were readjusted to more closely replicate Streetlight shares.

Mode choice

The overall mode split is set out in Table 8.1. Of the 6.2 million daily trips in the model, auto is by far the dominant mode, with a 97% share; rail and flight are the dominant non-auto modes. Looking at the fit between observed and modeled, auto has a very good fit, reflecting its dominance. Both rail and flight achieve a good fit. Bus particularly and combo have the poorest fit, reflecting their low and dispersed volumes. Overall, the modeled split is considered reasonable.

Table 8.1: Mode choice validation

Trips	Observed	Modeled	Difference	% Difference
rail	106,151	94,230	-11,921	-11%
auto	6,004,387	6,029,670	25,283	0%
bus	8,944	11,521	2,577	29%
flight	88,759	83,135	-5,624	-6%

Trips	Observed	Modeled	Difference	% Difference
combo	2,661	1.04E+03	-1,618	-61%
Total	6,210,902	6,219,599	8,697	0%

A more detailed review of rail is shown in Table 8.2. This shows the top five regional rail observed demand and the equivalent modeled demand. All show a reasonable fit, with the poorest being the Intra-SANDAG movement.

Table 8.2: Regional level daily rail demand

Movement	Observed	Modeled	Difference	% difference
Intra-SCAG	32,510	21,725	-10,785	-33%
Intra-MTC	53,748	43,237	-10,511	-20%
SCAG-SANDAG	5,347	6,752	1,405	26%
Intra-SANDAG	4,247	2,678	-1,569	-37%
San Joaquin-MTC	4,122	3,889	-233	-6%

Assignment

In the 2018 base year, the pivot process reproduces the observed modal demand at the County level. Below this at the zone level, the patterns of OD travel retain the modelled pattern. Thus, at a high level, the assignment should validate well, as it will match the observed County level demand; however, at a more granular level, the modeled OD pattern will likely match less well, impacting detailed replication, such as at station or line level for rail.

Rail

Lines

Assigned ridership by line against observed is set out in Table 8.3 and presented graphically in

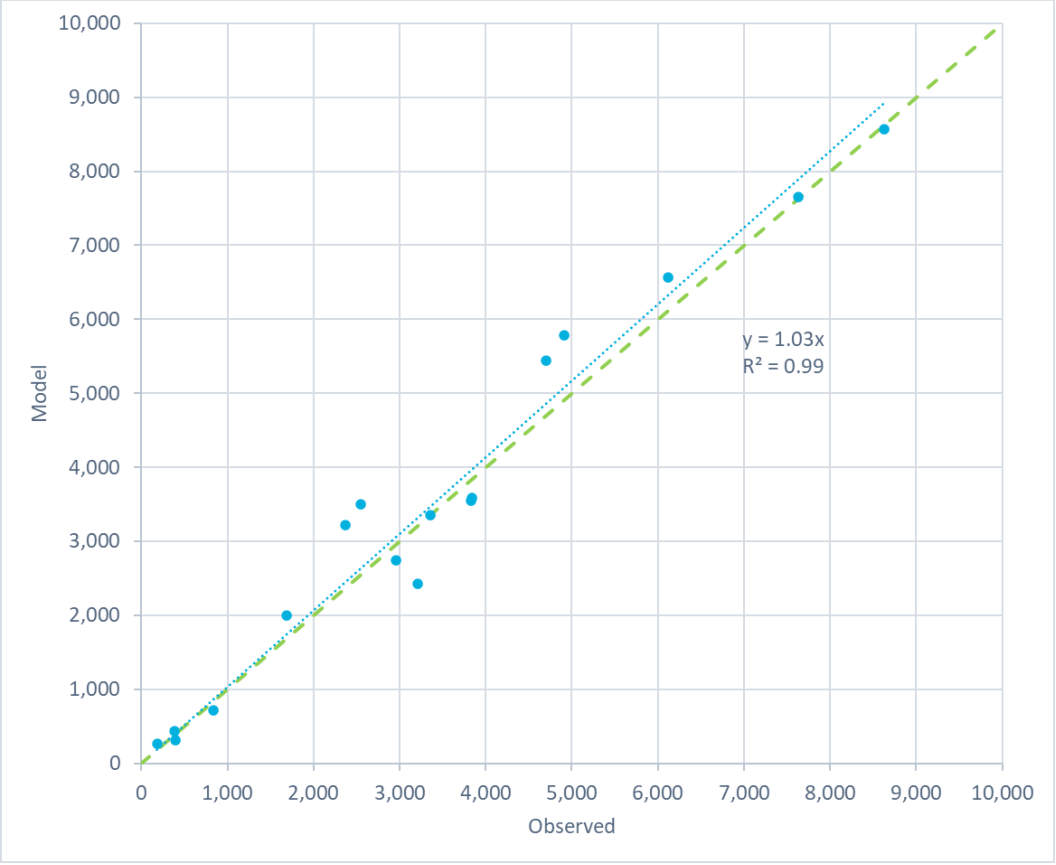
Figure 8-1 (excluding Caltrain given its scale). The poorer performing services are generally those which are part of the Metrolink network, although overall Metrolink has a good fit. The long-distance Amtrak routes are reasonable, with the exception of the Sunset Limited / Texas Eagle, albeit this only operates daily and has a notably low observed volume. Overall, the line validation is reasonable.

Table 8.3: Rail line validation

Service	Route	Observed	Modeled	Difference	% Difference
ACE	Altamont Corridor Express	3,833	3,552	-281	-7%
Amtrak	California Zephyr	396	311	-85	-22%
Amtrak	Southwest Chief	384	440	56	15%
Amtrak	Coast Starlight	833	724	-109	-13%
Amtrak	Pacific Surfliner	8,628	8,572	-57	-1%

Service	Route	Observed	Modeled	Difference	% Difference
Amtrak	Capitol Corridor	4,703	5,446	742	16%
Amtrak	Sunset Limited / Texas Eagle	180	261	81	45%
Amtrak	San Joaquins	2,961	2,749	-213	-7%
SMART	Main Line	1,686	2,004	318	19%
Metrolink	San Bernardino Line	7,627	7,647	20	0%
Metrolink	Ventura County Line	2,544	3,504	960	38%
Metrolink	Antelope Valley Line	4,907	5,780	873	18%
Metrolink	Riverside Line	3,211	2,432	-779	-24%
Metrolink	Orange County Line	6,119	6,563	444	7%
Metrolink	Inland Empire	3,354	3,358	4	0%
Metrolink	91/Perris Valley Line	2,373	3,221	848	36%
Metrolink	Metrolink network	30,135	32,503	2,368	8%
NCTD	COASTER	3,836	3,589	-246	-6%
Caltrain	Caltrain	54,301	51,187	-3,114	-6%

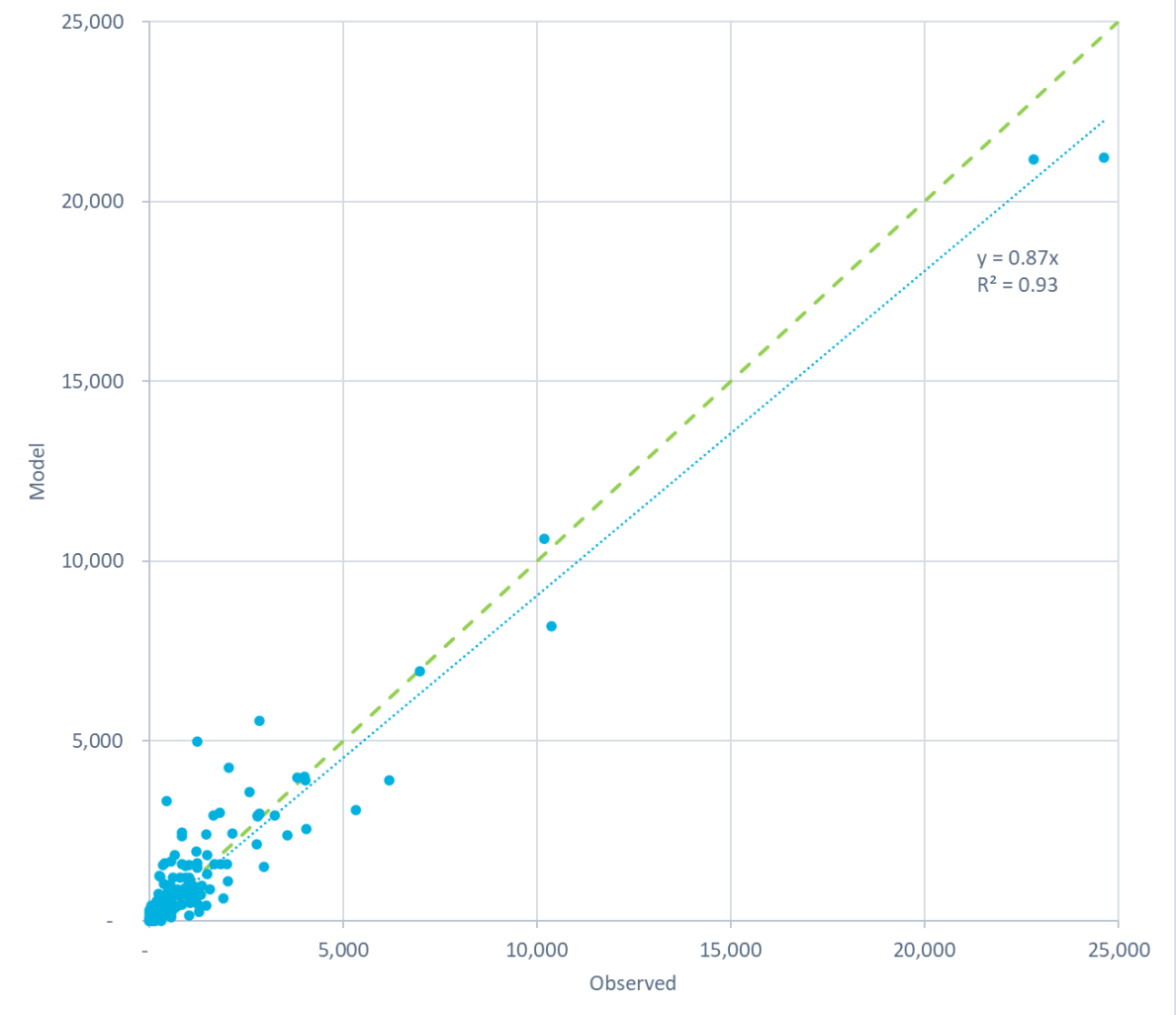
Figure 8-1: Rail line validation (excluding Caltrain)



Station level demand

Station level demand (total boardings and alightings) is shown in Figure 8-2. The two stations with the highest demand are Los Angeles Union station and San Francisco 4th & King, with the other stations having half or less the demand. For the stations with lower demand (in the range of a few thousand a day), the model fit becomes more variable, reflecting the use of the model zone level patterns by rail noted above. In addition, there are lines where station spacing is very short (e.g., within a mile or two, around San Jose for example, and the Metrolink network), making a better fit more challenging.

Figure 8-2: Rail station daily demand

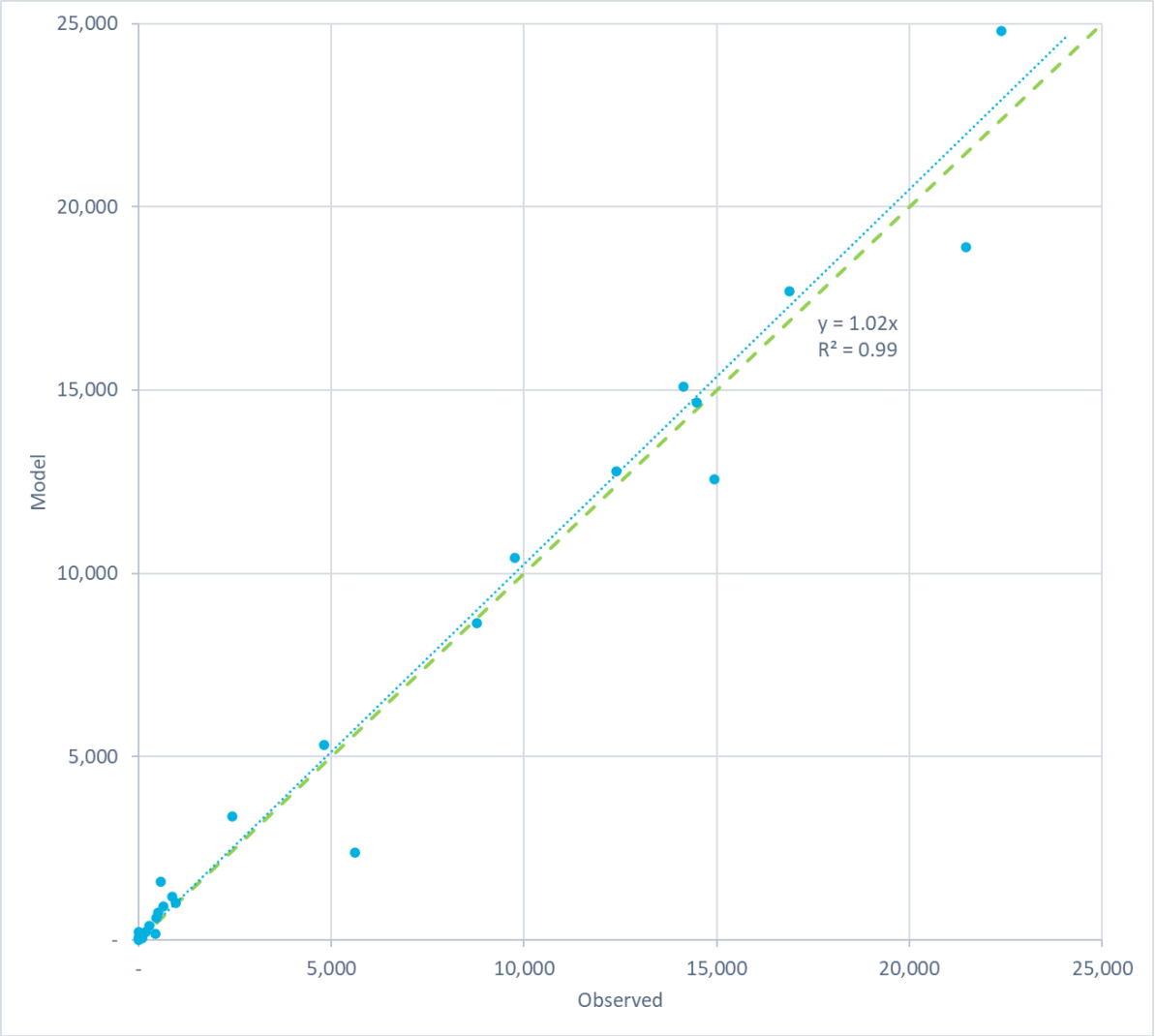


Air

Airport demand

Airport demand (total boardings and alightings) is shown in Figure 8-3. Overall, this demonstrates a good fit with observed data, with only minor variance resulting from the airport pair choices that exist in the State.

Figure 8-3: Airport daily demand



Airport pairs

At an airport pair level, the fit is comparable, as demonstrated in Figure 8-4. Overall, this demonstrates a good fit with observed data, with only minor variance in the smaller airport pair demands resulting from the airport pair choices that exist in the State, notably between the MTC and SCAG regions.

Figure 8-4: Airport-Airport daily demand

Auto

The auto time skimming uses fixed times developed from CSTDMv2 and MPO model data and hence auto times do not change with marginal changes in the modeled auto demand. Furthermore, the auto assignment is of a very partial auto matrix of trips deemed in-scope; consequently, assigned flows at a link level do not represent the total flow. Thus, any comparison between modeled and observed link level demand would not be meaningful and is not shown here.

COVID-19

2 As of the date of distribution of this document, the COVID-19 outbreak still has some impacts on global economic and political affairs including having a significant impact on the passenger rail industry in California where passenger volumes have fallen drastically – like other rail services and other modes elsewhere in the country. The situation remains dynamic and rapidly evolving with no real precedent and is subject to significant changes. There is currently limited authoritative information regarding the long-term impacts of this outbreak. Moreover, any third-party inputs will probably reflect a wide range of views. Hence, any analyses proposed from this work would include inherent uncertainties and may also include wide ranges of likely outcomes.

3 To ensure that the stakeholders are aware of these uncertainties, Steer had conducted meetings with the Authority and Caltrans to address overall status including risks and challenges. The model is a long-term forecasting tool. However, short-term impacts of the COVID-19 situation will be considered (subject to the limitations described in the paragraphs above) in the development of the long-term model data as with any impacts during economic downturns.

Control Information

Prepared by

Steer
800 Wilshire Blvd, Suite 1320,
Los Angeles, CA 90017
USA
+1 (213) 425 0990
www.steergroup.com

Prepared for

DB E.C.O. North America Inc. on behalf of California High-Speed Rail Authority

Steer project/proposal number

23454511

Version control/issue number

02

Date

March 2023

steer

QUALITY SYSTEM CERTIFICATION
DNV·GL
ISO 9001:2015

steergroup.com

steer