1 Methodology

This section provides a comprehensive description of the methodology used to construct the IBG, including all steps for data cleaning, processing, and integration.

1.1 Data acquisition

We identified candidate operators using the ICBA and targeted web searches. Then we used national repositories such as Transitland and the Mobility Database. For carriers without published GTFS feeds, we compiled schedules from publicly available agency timetables and Transcor Data Services, then converted them to GTFS format using custom Python scripts to ensure consistency. To support traceability, we maintained a reference database with source URLs and download timestamps for each feed.

1.2 Split combined feeds into unique agency feeds

Some data repositories provide GTFS feeds that include multiple transit agencies. To keep each operator separate, we split the combined datasets so each carrier had its own feed. This split preserved internal consistency by filtering routes by `agency_id` and then extracting the corresponding trips, stop times, stops, shapes, and service calendars so that each resulting feed remained complete.

1.3 Validate GTFS feeds

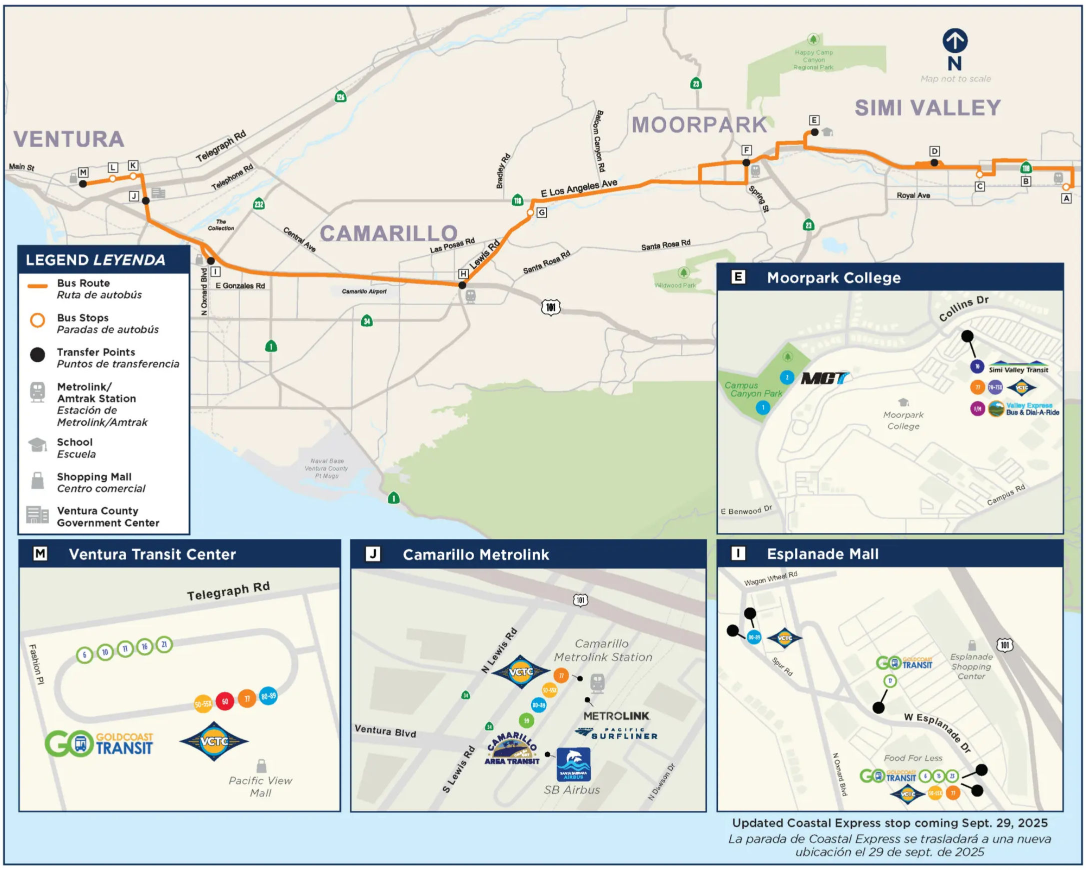
Raw GTFS feeds often contain errors that hinder automated processing. Before consolidation, we applied a standardized validation process, including:

- Presence of required GTFS files (e.g., `agency.txt`, `routes.txt`, `trips.txt`, `stops.txt`, `stop_times.txt`, `calendar.txt`).
- Presence of mandatory fields within each file.
- Uniqueness of primary identifiers (e.g., `route_id`, `trip_id`, `stop_id`).
- Referential integrity between linked files (e.g., trips referencing valid routes, stop times referencing valid trips, service identifiers in trips matching entries in the calendar).
- Valid time, date, and coordinate formats.
- Consecutive stop sequence ordering within each trip.
- Detection of duplicate shape geometries and single-day service periods (where `start_date` = `end_date`).
- Verification that all shape and stop identifiers are actively referenced by trips and stop times, respectively.

1.4 Filter intercity routes

To restrict the dataset to intercity services and exclude local transit, we filtered routes based on route characteristics. Route distance was measured using the Haversine formula, which accounts for Earth's curvature when computing distances between consecutive points in `shapes.txt`. Routes primarily functioning as local circulators were identified using distance thresholds and service descriptions and then excluded from the final dataset. Only routes connecting distinct cities or regions were retained.

Defining "distinct" metropolitan areas sometimes requires judgment in borderline cases where routes connect multiple cities within one metropolitan region. For example, the Ventura County Transportation Commission (VCTC) Cross County Limited route connects Ventura, Camarillo, Moorpark, and Simi Valley within Ventura County (Figure 1). Although the route spans several municipalities and travels more than 50 km across the county, all stops remain within the Ventura County metropolitan area and mainly serve intraregional commuting trips. Therefore, this service was classified as a regional transit route rather than an intercity bus route and excluded from the IBG dataset. In contrast, services that connect Ventura County to metropolitan areas outside the county, such as Los Angeles or Santa Barbara, were classified as intercity routes and included in the dataset.



Updated Coastal Express stop coming Sept. 29, 2025
La parada de Coastal Express se trasladará a una nueva ubicación el 29 de sept. de 2025

Fig. 1. Cross County Limited Service Map.

1.5 Consolidate feeds

After validation and filtering, we combined all 72 feeds into one consolidated GTFS feed. This step was necessary to support system-wide analysis and visualization, as most GIS tools process only one GTFS feed at a time.

1.6 Tidy and optimize the merged feed

In this step, we cleaned and optimized the consolidated feed to improve usability and performance. Initial data preparation was performed using custom Python scripts, followed by two passes of the `gtfstidy` tool for structural normalization and aggressive optimization. This process included:

- Removing invalid and orphaned records, including unused entries in `stops.txt`, `shapes.txt`, `routes.txt`, and service records in `calendar.txt` and `calendar_dates.txt` that were no longer referenced by active trips.
- Removing single-day services where the service start and end dates were identical, as these typically represent special or non-recurring trips not suitable for regular service analysis.
- Standardizing agency metadata in `agency.txt`, including assigning a uniform IANA time zone (`America/New_York`) and populating required fields such as `agency_url` to ensure GTFS compliance and prevent downstream processing failures.
- Resolving broken stop hierarchies by removing orphaned `parent_station` references in `stops.txt` that pointed to non-existent stations.
- Simplifying route geometries in `shapes.txt` by removing redundant coordinate points, fixing sequence ordering, and recalculating distances to ensure consistent route length measurements.
- Cleaning `stop_times.txt` by removing invalid distance fields and correcting stop sequence ordering to maintain proper trip progression.
- Normalizing route display attributes in `routes.txt`, including standardizing `route_color` and `route_text_color` to valid six-character hexadecimal values.
- Deduplicating identical route shapes and consolidating overlapping stop locations to reduce redundancy while preserving the underlying network topology.
- Snapping stop locations to the nearest point on their associated route geometries to improve spatial consistency between stops and routes.
- Compressing service calendars by identifying and removing duplicate service patterns across `calendar.txt` and `calendar_dates.txt`.
- Minimizing trip records by converting repeated trip patterns into frequency-based entries, reducing the total number of trips while preserving schedule information.
- Standardizing and renumbering identifiers (e.g., `route_id`, `trip_id`, `stop_id`, `shape_id`, and `service_id`) to maintain valid relationships across all GTFS files and avoid identifier collisions after consolidation.

These steps reduced file size and redundancy while retaining the full structure of the intercity bus network. As a result, we obtained a consolidated GTFS feed that serves as a core component of IBG.

1.7 Calculate weekly trip frequency

We calculated weekly trip frequency for each unique route geometry by linking trip records with service calendars (e.g., `trips.txt` and `calendar.txt`) and summing the number of operating days per week for each associated trip.

Because duplicate and overlapping service patterns were already consolidated during the optimization step (via service calendar compression), each remaining service identifier represents a distinct operating pattern. We then aggregated the weekly trips across all trips sharing the same route geometry (`shape_id`), with each travel direction treated separately since opposing directions correspond to distinct shape geometries.

1.8 Build relational GeoJSON

To support lightweight spatial analysis, sharing, and web-based visualization, we exported the finalized consolidated GTFS data as a single relational GeoJSON `FeatureCollection`. This format consolidates spatial and service information into a self-contained file that can be used directly in standard GIS software and web mapping platforms without requiring access to the full GTFS schedule tables.

The GeoJSON contains two complementary feature types:

1. **Route features (LineStrings):** Each route feature represents a unique route–geometry combination. The geometry encodes the corridor's spatial path, while the associated attributes include agency name, route identifiers, route long name, an ordered list of stop identifiers, stop count, and the estimated weekly trip frequency.

2. **Stop features (Points):** Each stop feature represents a physical boarding or alighting location. Stop attributes include the stop name and geographic coordinates, along with a list of routes and agencies that serve the stop and a route count indicating the number of distinct routes serving that location.

Each feature includes a `feature_type` property (`route` or `stop`) to distinguish the two types within the single collection. Routes and stops are linked by shared identifiers in the feature properties, such as `route_id` and `stop_id`. Route features contain arrays of associated stop identifiers, and stop features list the routes serving each location.

Through this step, we developed the second component of the IBG: a relational GeoJSON. This dataset supports advanced spatial mapping and network analysis, offering a geographic framework that complements the consolidated GTFS feed.

2 Additional Results

Table 1. Statistics by agency, sorted by weekly vehicle kilometers (VKM).

agency	weekly trips	route-km	veh-km (weekly)	# routes	# stops
Greyhound-us	3,127	192,749	2,067,655	127	731
FlixBus-us	2,701	143,736	1,474,312	105	436
RedCoach	1,303	16,731	551,497	27	39
Omnibus Express	210	18,581	261,364	15	118
Peter Pan Bus Lines	943	10,033	244,602	44	21
Jefferson Lines	370	33,168	197,004	65	171
Limousine Express	231	23,954	186,539	28	91
Salt Lake Express	797	11,561	158,745	41	133
Tornado Bus	140	10,083	141,822	10	85
ShortLine Hudson	1,463	4,251	131,056	13	364
Bustang (CDOT)	625	2,041	122,052	6	51
Barons Bus Lines	266	15,940	112,069	38	111
Peter Pan (BZ)	538	4,582	83,343	29	57
Van Galder	441	521	71,718	1	11
Academy Bus	168	4,806	65,788	14	37
Burlington Trailways	76	9,407	64,020	12	47
Adirondack Trailways	164	10,285	59,205	31	73
New York Trailways	108	8,544	58,751	16	35
Oregon POINT	248	1,980	58,677	4	41
Indian Trails	189	7,879	55,296	27	88
International Bus Lines	56	6,080	51,026	7	41
Concord Coach Lines	145	2,910	49,698	9	36
Southeastern Stages	77	6,774	47,454	11	31
Fullington Trailways	139	9,717	46,387	28	51
Peoria Charter Coach	157	5,389	44,210	19	34
Trans-Bridge Lines, Inc.	282	9,634	42,981	63	24
Plymouth & Brockton	336	613	42,832	2	9
Vonlane Bus	126	3,009	42,123	9	12
Martz Bus	235	6,137	41,925	28	24
Northwestern Stage Lines	88	4,491	38,780	10	43
C&J Bus Lines	296	860	35,360	5	10
BayRunner Shuttle	178	2,383	33,979	13	19
Go Bus	140	3,170	33,525	10	32
Wisconsin Coach Lines	258	427	31,373	2	66
Rapid Connection	28	3,052	30,346	3	33
Land to Air Express	202	3,353	28,898	23	24
Virginia Breeze	56	3,995	28,552	4	26
Pine Hill Trailways	143	3,900	27,840	18	42
Ocean Travel	56	1,823	25,522	4	6
Rochester City Lines	360	2,647	22,006	26	101
Express Arrow	34	5,318	21,964	8	31
Pacific Crest Bus Lines	110	1,616	20,715	5	40
Best Bus	45	1,524	17,647	4	7
Xe Đò Hoàng	30	2,242	16,293	4	12
Delta Bus Lines Inc.	28	2,064	14,719	4	15
Michigan Flyer	133	376	14,192	2	6
Detroit Ann Arbor Express	197	135	13,434	1	2
American Star Tours	28	1,894	13,268	4	18
Miller Transportation	56	1,827	12,827	8	22
Coach USA Erie	104	556	12,773	3	70
Sunway Charters	70	1,468	12,353	8	21
All Aboard America	28	849	11,928	2	8
Jet Set Express	28	917	11,351	2	20
CorridorRides	220	124	10,978	1	6
Badger Bus	60	601	9,680	2	9
BeeLine Express-Village Travel	42	1,351	9,532	6	10
Lamers Connect	28	1,328	9,359	4	24
Salmon Runner Bus	28	269	7,573	1	12
Reindeer Shuttle Inc.	35	583	6,607	3	7
MTR Western	14	837	5,861	2	10

continued on next page

Table 1. Statistics by agency, sorted by weekly vehicle kilometers (VKM).

agency	weekly trips	route-km	veh-km (weekly)	# routes	# stops
Eastern Sierra Transit Authority	14	686	4,816	2	31
Southern Express	40	230	4,636	2	11
Central Oregon Breeze	16	744	4,463	2	11
Vermont Translines	14	397	4,408	1	13
CYR Bus Line	14	296	4,357	1	9
Redding Area Bus Authority	50	192	4,126	2	29
W&H Bus	6	919	2,758	2	6
Diamond Express	34	110	2,545	1	13
DATTCO	24	284	2,496	1	7
Northfield Lines	36	206	2,337	3	6
Rensselaer County/Yankee Trails	25	96	1,814	1	22
Superior Tours	6	556	1,744	2	7
Total	19,063	641,789	7,203,883	1041	3919

Table 2. State-level statistics.

state	# operators	# stops	route-km
Alabama	4	25	7,434
Arizona	6	66	15,744
Arkansas	4	32	9,230
California	12	300	44,594
Colorado	6	85	9,349
Connecticut	8	31	10,385
Delaware	6	6	2,837
Dist. of Columbia	6	7	778
Florida	6	150	33,685
Georgia	7	77	20,149
Idaho	4	65	4,628
Illinois	9	88	17,709
Indiana	7	69	14,076
Iowa	6	47	9,915
Kansas	5	29	5,997
Kentucky	5	13	4,404
Louisiana	5	39	10,467
Maine	3	32	1,894
Maryland	7	42	12,843
Massachusetts	10	141	13,630
Michigan	6	98	10,455
Minnesota	8	199	16,673
Mississippi	5	24	5,886
Missouri	5	39	10,653
Montana	3	32	5,404
Nebraska	4	37	3,376
Nevada	7	32	4,138
New Hampshire	3	27	1,476
New Jersey	14	104	33,961
New Mexico	3	37	14,348
New York	21	579	39,463
North Carolina	8	105	15,222
North Dakota	1	10	2,432
Ohio	4	112	19,658
Oklahoma	7	32	6,741
Oregon	7	119	7,897
Pennsylvania	10	121	31,248
Rhode Island	3	3	729
South Carolina	5	37	8,435

continued on next page

Table 2. State-level statistics (continued).

state	# operators	# stops	route-km
South Dakota	1	23	3,395
Tennessee	5	19	8,172
Texas	9	275	61,422
Utah	3	55	8,362
Vermont	3	18	965
Virginia	7	63	13,701
Washington	4	43	9,097
West Virginia	4	22	2,769
Wisconsin	9	137	8,429
Wyoming	4	21	4,495
Total	–	3,767	608,751