Geocoding Truck Stops Documentation

William Co

2025-08-19

Table of contents

1 Setup

This study utilizes a comprehensive truck stop directory dataset containing information about individual truck stops. Notably, the original dataset does not include geographic coordinates (latitude and longitude), which are essential for spatial analysis and mapping. The primary objective of this research is to systematically extract and assign accurate geographic coordinates to each truck stop entry.

A significant challenge encountered in this process stems from the inconsistent formatting of address information. While some entries provide complete street addresses, others list only road names, highway exits, or mile markers. This lack of standardization complicates the process of automated geocoding and necessitates additional data processing steps.

Furthermore, the dataset required extensive cleaning to ensure its suitability for analysis. This included standardizing address formats.

2 Challenges

2.1 Inconsistent Addresses

Addresses in the dataset fall into the following categories:

- **Standard addresses:** These include a street number and road name, allowing for straightforward geocoding.
- **Exit-based addresses:** These reference a highway and exit number, but may lack a full street address.
- **Non-standard addresses:** These do not conform to either of the above formats, such as entries that only specify the intersection of two streets or other ambiguous location descriptions.

This variability in address formats presents a significant challenge for automated geocoding.

2.2 Age

A further complication arises from the temporal nature of the dataset. Several locations are historical or no longer in operation, which poses challenges when attempting to verify their existence using contemporary mapping services. In such cases, it was necessary to consult archival web resources and historical records to confirm the status and location of these truck stops. This process often required subjective judgment to determine whether a site remains active, has been repurposed, or no longer exists.

3 Data Scraping

To obtain accurate geographic coordinates for each truck stop, we systematically collected data from several reputable online sources. For every truck stop entry, we extracted latitude and longitude information from the following platforms:

- **Truck Stops and Services / RV and Travelers Directory:**
  We began by scraping data from Truck Stops and Services and the RV and Travelers Directory. These websites were selected for their consistent formatting and comprehensive coverage of truck stop locations. The structured nature of their listings facilitated reliable extraction of geographic coordinates and associated metadata.
- **Yelp:**
  The Yelp API was utilized due to its unique capability to perform searches based on phone numbers, which aligns well with our dataset's standardized contact information. This allowed for precise matching and retrieval of business coordinates, even in cases where address information was incomplete or ambiguous.
- **Yellow Pages:**
  We also scraped YellowPages, leveraging its support for phone number-based queries. This provided an additional layer of verification and expanded our ability to cross-reference locations.

For each source, we captured not only the coordinates but also relevant business and location attributes, as detailed in the respective data dictionaries in the Appendix. By integrating data from multiple platforms, we increased the likelihood of obtaining accurate and up-to-date geographic information for each truck stop.

3.1 Entry Matching

Following the data collection phase, three independent reference datasets were assembled from the aforementioned online sources. The next step involved systematically matching these reference datasets to the original truck stop directory to assign geographic coordinates to each entry.

To accomplish this, we developed two principal methodologies:

3.1.1 Phone Number Matching

The first approach relies on direct matching of entries based on phone numbers. Phone numbers obtained from Yelp, Yellow Pages, and Truck Stops and Services are compared to those in the original truck stop directory.

3.1.2 Place Name to ZIP Code Matching

The second approach employs a hierarchical matching strategy. Initially, entries are filtered by state or ZIP code. Subsequently, matches are refined by city or highway exit, followed by road name, and finally by business or place name.

Using these matching methodologies, each entry in the original truck stop directory could be associated with up to four potential matches: three derived from phone number-based matching (utilizing data from Yelp, Yellow Pages, and Truck Stops and Services) and one from the hierarchical place name to ZIP code matching approach. This comprehensive strategy maximized the likelihood of accurately linking each truck stop entry to its corresponding geographic coordinates.

4 Post Matching

4.1 False Positives

A notable challenge encountered during the matching process was the occurrence of false positives. Inconsistencies in phone number records, particularly within Yellow Pages, and documented inaccuracies in latitude and longitude values across all data sources resulted in multiple, and sometimes conflicting, geographic coordinates for a single truck stop entry.

To mitigate these issues, we implemented a systematic coordinate validation methodology as described below.

4.2 Coordinate Matching

Suppose we have a Yelp source, where the phone number matches to two unique latitude and longitude. Next, the Truck Stops and Services (Phone) source matches to two different latitude and longitude pairs, while Truck Stops and Services (Place Match) matches to one distinct coordinate. There is no match for Yellow Pages. Finally, we have the original entry_n from the truck stop directory. In this scenario, we observe several possible coordinates for a single truck stop, but there is no reliable method to determine which coordinate is correct.

Suppose we also entry_{n+1} where we have multiple coordinates from the Truck Stops and Services (Place Match) source but no matches from other sources. Suppose for entry_{n+2}, we have one coordinate from Yelp, one from Yellow Pages, and one from Truck Stops and Services (Phone). This scenario allows us to compare matches across three different sources.

Let C_i represent coordinate pairs (latitude, longitude) where $C_i = (\text{lat}_i, \text{lon}_i)$.

Data Source	Yelp	Truck Stops and Services (Phone)	Truck Stops and Services (Place Match)	Yellow Pages
entry_n	C_0, C_1	C_2, C_3	C_4	$\emptyset$
entry_{n+1}	$\emptyset$	$\emptyset$	C_5, C_6, C_7	$\emptyset$
entry_{n+2}	C_8	C_9	$\emptyset$	C_{10}

4.2.1 Case 1: Multiple Matches

For entry$_n$, we have the set of possible coordinates $\mathcal{C}_n = \{C_0, C_1, C_2, C_3, C_4\}$, where each coordinate represents a potential location for the same truck stop entry.

In order to discern the correct coordinate, we make the assumption that a true coordinate C^* exists where if two different sources agree on the same coordinate, the said coordinate must be correct.

Distance-Based Validation Approach:

Using this approach, we calculate the Haversine distance between all coordinate pairs. For entry$_n$, we compute all pairwise distances between coordinates originating from different data sources (i.e., different columns in the table). Specifically, we exclude distances between coordinates from the same source; for example, $D_{0,1}$ is not considered if both C_0 and C_1 are from the Yelp column. Similarly, $D_{2,3}$ is excluded as both coordinates are from the Truck Stops and Services (Phone) column. Additionally, $D_{2,4}$ and $D_{3,4}$ are also excluded, as both coordinates are derived from the Truck Stops and Services website, albeit from different matching methods (Phone and Place Match). This ensures that only distances between independent sources are evaluated for validation purposes.

The relevant distances for this case are:

- $D_{0,2}$: distance between the first Yelp coordinate and the first Truck Stops and Services (Phone) coordinate
- $D_{0,3}$: distance between the first Yelp coordinate and the second Truck Stops and Services (Phone) coordinate
- $D_{0,4}$: distance between the first Yelp coordinate and the Truck Stops and Services (Place Match) coordinate
- $D_{1,2}$: distance between the second Yelp coordinate and the first Truck Stops and Services (Phone) coordinate
- $D_{1,3}$: distance between the second Yelp coordinate and the second Truck Stops and Services (Phone) coordinate
- $D_{1,4}$: distance between the second Yelp coordinate and the Truck Stops and Services (Place Match) coordinate

The set of all relevant pairwise distances is defined as:

$$\mathbf{D} = \{D_{i,j} : C_i, C_j \text{ are from different sources}\}$$

We then identify the minimum distance $\min\{D_{i,j}\}$ in $\mathbf{D}$ and record this value in a dedicated column called `min_distance`. This enables systematic manual correction and error analysis.

Entries where `min_distance` exceeds 200 meters are flagged for manual review and correction. If two coordinates C_i and C_j from different sources are within 200 meters (i.e., $\min\{D_{i,j}\}$ is less than 200 meters), their midpoint is used as the final coordinate for that entry. If manual correction is required, the updated coordinates C_i^* are recorded as the corrected location. Where C_i^* denotes the manually corrected coordinate, obtained from online sources, most notably Google Maps satellite imagery. Additional details regarding the correction can be found in the associated `Match_Comments` field.

4.2.2 Case 2: Single or No Match

For entry_{n+1}, if only one source provides a match, or if no matches are found, the entry is flagged for manual review and correction.

4.2.3 Case 3: Hierarchical Matching

Consider the scenario for entry_{n+2} with three matched coordinates: C_8 (Yelp), C_9 (Truck Stops and Services—Phone), and C_{10} (Yellow Pages). Compute the pairwise Haversine distances $D_{8,9}$, $D_{8,10}$, and $D_{9,10}$. Because sources differ in reliability, we apply a hierarchical decision rule that prioritizes agreement with Truck Stops and Services (TS&S).

Selection rule:

1) Let $\mathcal{P} = \{(8,9),(8,10),(9,10)\}$ be all pairs and let $\mathcal{P}_{\text{TS\&S}} = \{(i,j) \in \mathcal{P} : \{i,j\} \ni 9\}$ be the subset that includes the TS&S coordinate.

2) If $\mathcal{P}_{\text{TS\&S}} \neq \emptyset$, select $(i^*, j^*) = \arg\min_{(i,j)\in\mathcal{P}_{\text{TS\&S}}} D_{i,j}$. Otherwise, select $(i^*, j^*) = \arg\min_{(i,j)\in\mathcal{P}} D_{i,j}$.

3) If $D_{i^*,j^*} \leq 200$ meters (the same threshold used in Case 1), set the final coordinate to the midpoint of C_{i^*} and C_{j^*}.

4) If all pairwise distances exceed 200 meters, flag the entry for manual review (retain the minimizing pair and distance for auditing).

In summary, when multiple matches are available, we:

- Prefer pairs that include the most reputable source (Truck Stops and Services) and, among those, select the smallest distance.
- If no reputable-source pair exists, select the absolute smallest distance across all pairs.
- Use the midpoint when the selected pair is within the 200 m threshold; otherwise, flag for manual validation.

5 Appendix

5.1 Truck Stops and Services/ RV and Travelers Data Dictionary

The following table summarizes the data fields used in the truck stop dataset:

Column Name	Description

5.1.1 General Information

Column Name	Description
`state_id`	State identifier
`state`	Name of the U.S. state
`name`	Truck stop name
`href`	Relative URL path
`full_url`	Full website URL
`stop_type`	Type of stop (e.g., fuel, full)
`Chain`	Company or chain name

5.1.2 Location Details

Column Name	Description
`Latitude`	Latitude coordinate
`Longitude`	Longitude coordinate
`Highway`	Associated highway
`Exit`	Exit number
`Mile Marker`	Highway mile marker
`Street Address`	Street address
`City`	City name
`State`	State abbreviation
`Postal Code`	ZIP/postal code

5.1.3 Contact Information

Column Name	Description
`Phone`	Main contact number
`Phone 2-5`	Additional phone numbers
`Fax`	Fax number
`Mailing Address`	Mailing address

5.1.4 Amenities & Services

Column Name	Description
`Hours of Operation`	Operating hours
`# of Parking Spots`	Total parking spaces
`# of Reserved Parking Spots`	Number of reserved spaces

Column Name	Description
# of Paid Parking Spots	Paid-only spots
# of Fuel Lanes	Fuel pump lanes for trucks
# of Showers	Total shower stalls
# of Men's Showers	Men's shower stalls
# of Truck Service Bays	Truck repair/service bays

5.1.5 Fuel Types & Links

Column Name	Description
Unleaded	Unleaded gasoline available (Y/N)
Diesel	Diesel fuel available (Y/N)
Bulk Def	Diesel exhaust fluid (DEF) availability
Propane	Propane fuel available (Y/N)
https	HTTPS version of site URL

5.2 Yelp Data Dictionary

5.2.1 General Business Information

Column Name	Description
Original_Phone	The phone number used as input for the Yelp phone search
Name	The official name of the business
Rating	Yelp rating (e.g., 4.5 stars)
Review_Count	Total number of Yelp reviews
Is_Closed	Boolean indicating if the business is permanently closed
URL	Full Yelp business listing URL

5.2.2 Location Details

Column Name	Description
Address	Street address of the business
City	City where the business is located
State	State (abbreviation)
Zip_Code	Postal or ZIP code
Latitude	Latitude coordinate
Longitude	Longitude coordinate

5.2.3 Contact & Business Attributes

Column Name	Description
Phone	Official business phone number returned by Yelp
Categories	List of categories (e.g., "Coffee & Tea", "Gas Station")
Price	Price level indicator ($, $$, etc., if available)

5.3 Yellow Pages Data Dictionary

5.3.1 General Business Information

Column Name	Description
ADDRESS	Full address of the business as listed on Yellow Pages
AKA	Alternate names or aliases for the business
BUSINESS_NAME	The primary name of the business
BUSINESS_URL	URL to the Yellow Pages business listing
CATEGORIES	Business categories (e.g., "Restaurants", "Auto Repair")
STATUS	Business status (e.g., "Open", "Closed")
WEBSITE	Official website of the business, if available

5.3.2 Location Details

Column Name	Description
JSONLD_CITY_1	City extracted from the embedded structured JSON-LD data
JSONLD_STATE_1	State extracted from the embedded structured JSON-LD data
JSONLD_STREET_1	Street address from JSON-LD
JSONLD_ZIP_1	ZIP code from JSON-LD
JSONLD_LAT_1	Latitude coordinate from JSON-LD
JSONLD_LNG_1	Longitude coordinate from JSON-LD

5.3.3 Contact & Business Attributes

Column Name	Description
ORIGINAL_PHONE	Phone number used to initiate the Yellow Pages lookup

Column Name	Description
FORMATTED_PHONE	Formatted business phone number as displayed
JSONLD_PHONE_1	Phone number from the structured JSON-LD data
EXTRA_PHONES	Any additional phone numbers found
PHONE	Phone number listed in the primary Yellow Pages HTML content
JSONLD_NAME_1	Business name from structured JSON-LD data

5.3.4 Metadata

Column Name	Description
SCRAPED_AT	Timestamp of when the data was scraped
SEARCH_URL	URL used to perform the Yellow Pages phone-based search

5.4 iExit Data Dictionary

Column Name	Description
state	U.S. state abbreviation (e.g., TX, CA) where the highway exit is located
highway	Name or number of the highway (e.g., I-10, US-101)
exit_id	Unique identifier for the highway exit as used in iExit
title	Display title or name of the exit
exit_name	Name of the exit (may include road or location name)
exit_description	Additional descriptive text about the exit or nearby services
exit_location	Textual representation of the exit's location
iexit_detail_link	URL link to the iExit detailed page for the exit
latitude	Latitude coordinate of the exit
longitude	Longitude coordinate of the exit
google_maps_link	Direct link to the exit location on Google Maps
direction	Direction of travel (e.g., Northbound, Eastbound)