Devunuri, Saipraneeth, and Lewis Lehe. 2024. "A Survey of Errors in GTFS Static Feeds from the United States." *Findings*, April. https://doi.org/10.32866/001c.116694.

TRANSPORT FINDINGS

A Survey of Errors in GTFS Static Feeds from the United States

Saipraneeth Devunuri[1], Lewis Lehe[1]

[1] Civil and Environmental Engineering, University of Illinois Urbana-Champaign

Keywords: GTFS, public transit, data quality, open data, public transport

https://doi.org/10.32866/001c.116694

Findings

This study surveys the errors in General Transit Feed Specification (GTFS) Static (Schedule) data for 632 US transit feeds. We do so using the Canonical GTFS Schedule Validator tool provided by Mobility Data, which checks feeds against GTFS standards. About 21% of GTFS feeds have at least one error. We explain what the most common errors are and provide examples. Errors related to the optional shape_dist_traveled field account for the majority of errors. Fares account for a second cluster of errors. Manual investigation can reveal errors not captured programmatically.

1. Questions

The General Transit Feed Specification (GTFS) is an Open Data standard that transit agencies use to publish data (McHugh 2013). A challenge in applying GTFS data is that agencies sometimes make mistakes in GTFS feeds. Hence, California imposes "Minimum GTFS Guidelines" to reduce errors (Cal-ITP 2024). Barbeau (2018) developed a software validator for GTFS "Realtime" feeds (which provide realtime transit information) and found errors in 54 of 78 realtime feeds tested. Since 2021, MobilityData (the organization that maintains GTFS standards) has offered an open-source Canonical GTFS Schedule Validator (MobilityData 2024a) aimed at GTFS Static[1] feeds, which documents planned service. This paper runs the Validator on all working US GTFS Static feeds listed on the Mobility Database (MobilityData 2024b). The paper answers the question: "What kinds of errors occur in US GTFS Static feeds?" The appendix shows cases from real GTFS feeds of the ten most common errors.

2. Methods

We downloaded the most recent GTFS Static data for 632 feeds (including data from 743 agencies) in the US. Included are all US feeds with either a valid or unspecified (empty) status in the Mobility Database. We run the Canonical GTFS Schedule Validator Desktop[2] app (v5.0.0) on each feed, then aggregate and analyze the results. The Validator outputs three levels of notices: errors, warnings and info. This study is limited to errors, which are violations of the specification. There are 72 errors (listed at https://gtfs-validator.mobilitydata.org/rules.html). Since some errors by their nature

1 The terms 'GTFS Static' and 'GTFS Schedule' are both used for the same set of rules.

2 The Validator has two versions: a Desktop app and a web interface to which one can upload feeds.

Table 1. Distribution of feeds by count of unique errors

Number of Unique Errors	Frequency Count	% of Feeds with Errors	% of All Feeds
0	500	-	79.1
1	83	62.9	13.1
2	34	25.8	5.4
3	8	6.1	1.3
4	5	3.8	0.8
5	2	1.5	0.3
Total	632	100%	100%

happen many times in one feed (e.g., every time a stop is recorded), rather than errors themselves we count *error occurrences*: the event that a feed exhibits some error at least once.

3. Findings

Table 1 shows the frequency distribution of error occurrences across feeds. Errors are relatively uncommon. Only 132 of 632 (21%) feeds contain errors, and most feeds with an error exhibit just one.

Errors are concentrated. Only 22 of 72 possible errors occur at all. Only ten errors occur in five or more feeds, and these ten account for 90% of all error occurrences. Figure 1 shows the distribution of error occurrences. The 'Other' category in the figure contains twelve miscellaneous errors that occur rarely (e.g., invalid URLs or colors). The ten most common errors are:

1. **equal_shape_distance_diff_coordinates**: Two points on a route shape have the same *shape_dist_traveled* but different coordinates (which is impossible).
2. **decreasing_or_equal_stop_time_distance**: For some trip, *shape_dist_traveled* decreases or stays the same from one stop to the next in `stop_times.txt`. Hence either *shape_dist_traveled* is wrongly calculated or the stops are out-of-order.
3. **trip_distance_exceeds_shape_distance**: The maximum of *shape_dist_traveled* in `stop_times.txt` exceeds the maximum of *shape_dist_traveled* in `shapes.txt`.
4. **foreign_key_violation**: Some file refers to a key which is never defined in its "parent" file: e.g., `stop_times.txt` references stop S132, but `stops.txt` does not mention S132.
5. **invalid_currency_amount**: The fare is invalid according to the ISO 4217 standard. Usually, fares are missing decimals: e.g., $2 instead of $2.00.

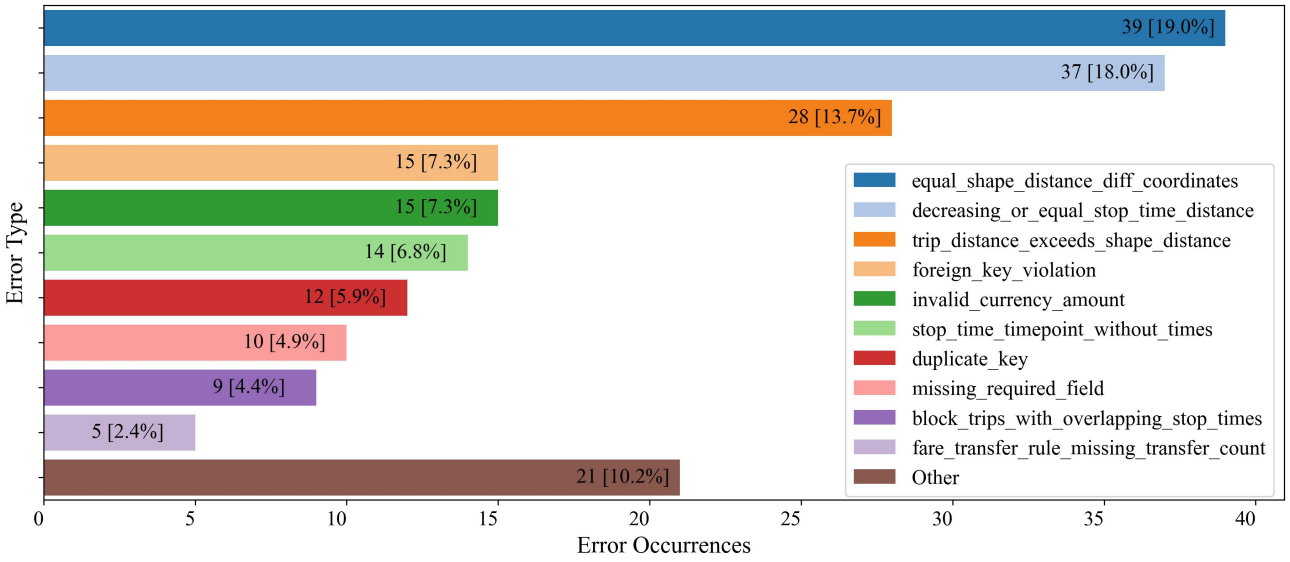

Figure 1. Distribution of error occurrences

6. **stop_time_timepoint_without_times**: An entry in `stop_times.txt` is missing either arrival or departure time, but has the field *timepoint* set to 1 instead of 0.

7. **duplicate_key**: Two entities have the same key: e.g., two trips with the same *trip_id*.

8. **block_trips_with_overlapping_stop_times**: Trips with the same *block_id* should be served by the same vehicle. This error indicates that stop times with the same *block_id* overlap (so one vehicle cannot serve them).

9. **missing_required_field**: A file is missing 'required' or 'conditionally required' fields: e.g., a trip in `trips.txt` without a corresponding *route_id*.

10. **fare_transfer_rule_missing_transfer_count**: A fare transfer rule with same *from_leg_group_id* and *to_leg_group_id* is missing *transfer_count*: the field that defines a limit for consecutive transfers.

The sources of errors are concentrated. The top three most common errors are related to the optional *shape_distance_traveled* field and account for a majority (51%) of all error occurrences. What is *shape_dist_traveled*? In `shapes.txt` file, *shape_distance_traveled* indicates how far each point on the path a vehicle travels lies from the start of the shape (moving along the path). In `stop_times.txt`, it indicates how far each stop is from the beginning of a trip. While optional, 74% of US feeds include *shape_dist_traveled* for every trip. It is best practice to include *shape_dist_traveled* when a route intersects itself, and the field also makes it possible to project stop locations from `stops.txt` onto route shapes.

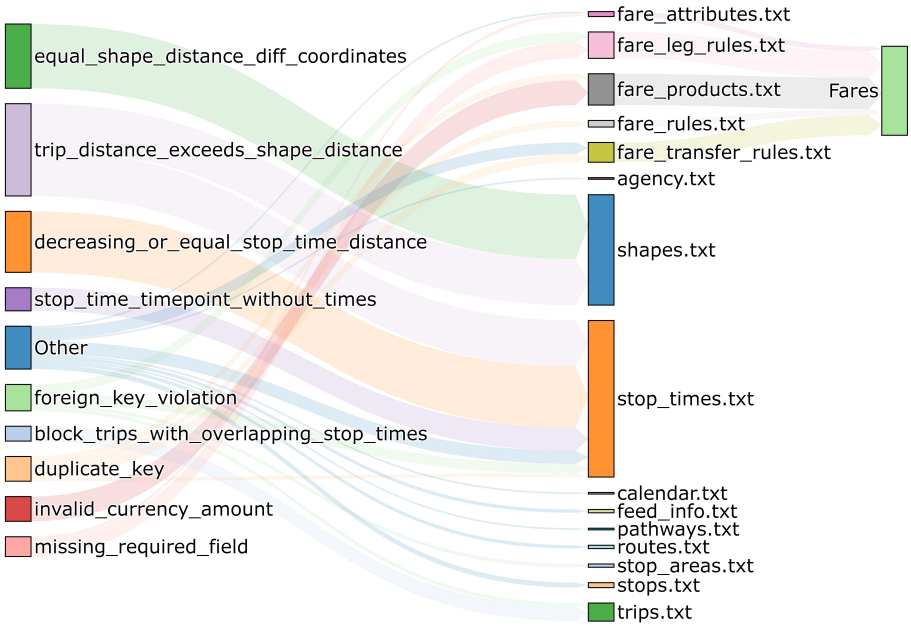

Figure 2. Error occurrences mapped to associated GTFS files

Mapping errors to the files where they occur, as in Figure 2, highlights a second common source of error: fare data. The five `fare_` files highlighted in the Figure account for 22.6% of all errors. Fares are such a major source of error because GTFS fare specifications are extremely complex so as to accommodate a wide range of fare schemes.

Since errors are concentrated among fares and *shape_dist_traveled*, it may not be hard to curtail errors by tackling these two causes. In particular, the complexity of fares specification calls for more examples and documentation. Fortunately, MobilityData is developing a new Fares V2 standard and has provided training videos and a template for it[3]. However, this may not address cases in which an agency simply does not consider it worthwhile to obey the GTFS standards to the letter. Some violations of GTFS rules are probably perceived as inconsequential. For example, a fare (field *amount*) listing of 2 instead of 2.00 triggers the **invalid_currency_amount** error, but trip planning applications can interpret 2.

Note that our survey is limited to errors that can identified programmatically, but this can pass over some severe errors discernible only by manual investigation. Figure 3 shows an example from the Dallas Area Rapid Tranist (DART) feed. The Validator gives a 'warning' **stop_too_far_from_shape** that stops 33329 and 33554 are more than 100 meters from the shape of route 421. The underlying problem, though, is that the stops lie on a street (Junius

3 https://gtfs.org/schedule/examples/fares-v2/

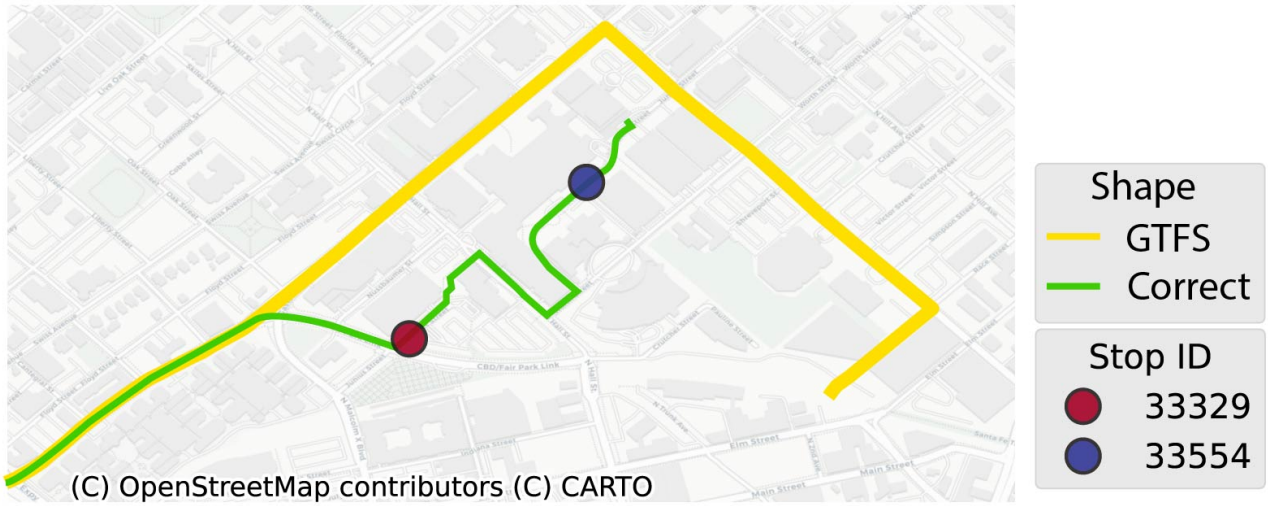

Figure 3. The shape for DART route 421 westbound is drawn incorrectly

Street) which the shape does not traverse at all. In reality, route 421 *does* travel Junius Street, taking a path different from the feed's shape. Hence our survey of errors is not exhaustive.

Submitted: April 03, 2024 AEST, Accepted: April 18, 2024 AEST

REFERENCES

Barbeau, Sean J. 2018. "Quality Control - Lessons Learned from the Deployment and Evaluation of GTFS-Realtime Feeds." In *Transportation Research Board 97th Annual Meeting Transportation Research Board*. 18–05585. https://trid.trb.org/View/1496848.

Cal-ITP. 2024. "California Transit Data Guidelines Caltrans." https://dot.ca.gov/cal-itp/california-transit-data-guidelines.

McHugh, Bibiana. 2013. "Pioneering Open Data Standards: The GTFS Story." In *Beyond Transparency: Open Data and the Future of Civic Innovation*, 125–35. Code for America Press San Francisco. https://beyondtransparency.org/part-2/pioneering-open-data-standards-the-gtfs-story/.

MobilityData. 2024a. "Mobility Database." https://database.mobilitydata.org/.

———. 2024b. *MobilityData/Gtfs-Validator: Canonical GTFS Validator Project for Schedule (Static) Files.* https://github.com/MobilityData/gtfs-validator.

SUPPLEMENTARY MATERIALS

Appendix

Download: https://findingspress.org/article/116694-a-survey-of-errors-in-gtfs-static-feeds-from-the-united-states/attachment/225750.pdf