GPS Data Segmentation

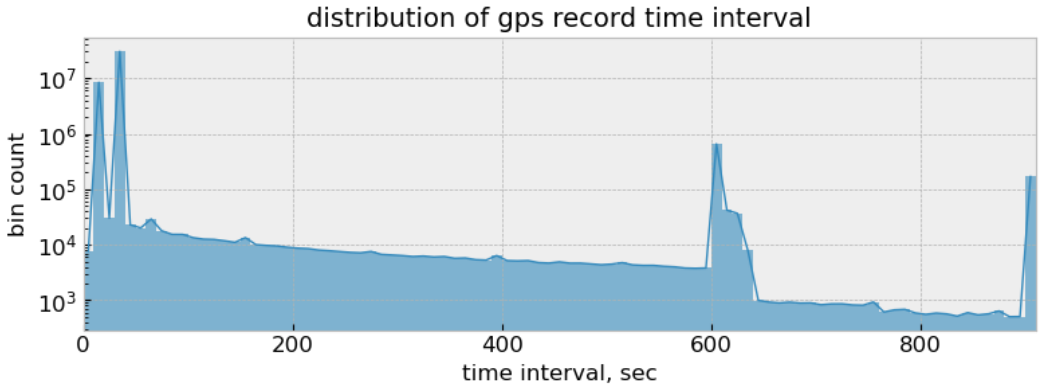
Russell Burdt

9/27/2021

GPS data segmentation means breaking up raw GPS data for a single vehicle into a distinct number of segments, where each segment is a contiguous set of GPS records meeting specific data quality criteria. The method is a data cleaning step that can improve the accuracy of subsequent data processing or modeling, and may also be used to identify vehicles with unstable GPS data streams.

Introduction

The time interval for contiguous GPS records is generally thought to be 10sec or 30sec, depending on device configuration. A distribution of time intervals based on appx 40M GPS records is below. Bin width is 10sec, first bin is [0sec, 10sec), second bin is [10sec, 20sec), and so on. Last bin is all time intervals ≥ 900sec. See Appendix A for the population definition.

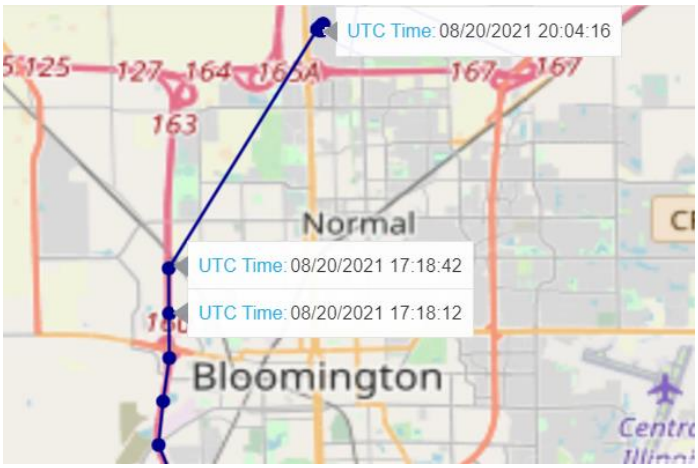


The distribution indicates 10sec and 30sec are indeed the most common time intervals, but that a range of other intervals less than 10sec and greater than 30sec (up to and exceeding 900sec) are possible. One interpretation of a long time interval is the vehicle is off and parked. In this scenario the distance interval aligned with the long time interval should be near zero, and many of the long time intervals in the distribution represent this case.

However many long time intervals are also aligned with long distance intervals which cannot be interpreted as the vehicle is off and parked; see the example above and raw query in Appendix B based on VehicleId 1C00FFFF-59AE-49C9-4E99-4663F0100000.

TS_SEC	timestamp	lat	lon	time_interval_sec	distance_interval_miles
1629479892	08/20/2021 17:18:12	40.49	-89.03	30.00	0.60
1629479922	08/20/2021 17:18:42	40.50	-89.03	30.00	0.60
1629489856	08/20/2021 20:04:16	40.55	-88.99	9934.00	3.69
1629489886	08/20/2021 20:04:46	40.55	-88.99	30.00	0.04
1629489916	08/20/2021 20:05:16	40.55	-88.99	30.00	0.07

The above example represented on a map is to the right. Contiguous GPS records are blue circles connected by lines. There is unresolved vehicle motion between 08/20/2021 17:18:42 and 08/20/2021 20:04:16, appx 3.7 miles over 2.8 hours. There are many possible physical explanations of the vehicle motion in this case, eg the vehicle was off and being towed, the driver was using the vehicle for personal use and unplugged the telematics device, a loss of coverage, etc.

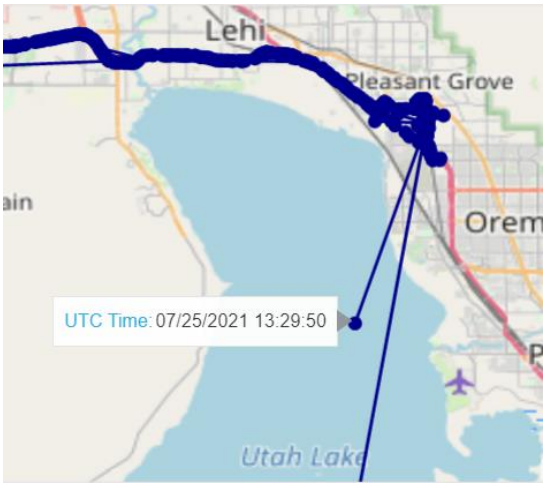


The dominant root cause of this behavior is currently suspected to be loss of coverage, or some device issue where data are not sent back, or some issue in receiving the data properly in the database (Appendix B provides additional query details). In reality Lytx can never control for some of the possible root causes (vehicle off and being towed, driver personal use, etc) so the behavior should always be expected in data at some rate. Statistics on frequency of occurrence of the behavior is in a subsequent section.

Unresolved vehicle motion may also have a non-physical origin. The example to the right indicates appx 7 miles over 5 min and then again over 30 sec. See Appendix B for the raw query based on VehicleId 1C00FFFF-59AE-48C9-716B-5D43E7800000.

TS_SEC	timestamp	lat	lon	time_interval_sec	distance_interval_miles
1627218793	07/25/2021 13:13:13	40.34	-111.74	328.00	0.02
1627219498	07/25/2021 13:24:58	40.34	-111.74	705.00	0.03
1627219790	07/25/2021 13:29:50	40.25	-111.78	292.00	6.93
1627219820	07/25/2021 13:30:20	40.34	-111.74	30.00	6.96
1627220488	07/25/2021 13:41:28	40.34	-111.74	668.00	0.07
1627221704	07/25/2021 14:01:44	40.34	-111.74	1216.00	0.06

The above example represented on a map is to the right. The non-physical motion is the GPS record at 07/25/2021 13:29:50 (driving into the lake). The subsequent record at 07/25/2021 13:30:20 is likely a physial GPS record as it is physically close to other records, and on land.



The GPS data segmentation method currently does not distinguish physical / non-physical motion, however it may be possible to do so based on additional data in GPS records. For example, (HDOP, NUM_SATS) for the record in the lake are (-1, -1) meaning not registered by the device, whereas values for the subsequent record are (0.8, 10) which indicates a good record. In the previous example of physical unresolved motion, HDOP and NUM_SATS consistently represent good records. More examples would need to be validated and limits for HDOP / NUM_SATS / etc would need to be carefully tuned as a next step though.

Data Segmentation Definition

The GPS data segmentation method assumes full duplicate rows and partial duplicate rows are already filtered. A 'partial duplicate' row means same vehicle-id and TS_SEC (epoch timestamp) where other data may be different, eg latitude and longitude. Partial duplicate rows are most likely unique GPS records sampled within the same second, which could be resolved if TS_USEC data were non-zero (currently TS_USEC appears always as zero). For the population defined in Appendix A (appx 40M records), appx 200k rows (0.5%) were full duplicates and appx 200 rows (0.001%) were partial duplicates.

Data segmentation then means identifying the set of GPS segments for all vehicles in a population, where a GPS segment is defined as a **contiguous set of GPS records with no unresolved large changes in position**. GPS segments are identified based on thresholds applied to time and distance intervals in GPS records. Current logic is to exclude from GPS segments those GPS records leading to the combination of:

- time interval greater than 61 sec, AND
- distance interval greater than 0.1 miles

Under this definition the scenario of 'the vehicle is off and parked' (long time interval, near zero distance interval) will not be excluded from GPS segments, whereas the simultaneous long time and distance intervals in the above examples will be excluded.

Data segmentation algorithm

Data segmentation is implemented in a new column in the GPS table 'segmentId' that is initialized as all zeros and then updated based on observed time and distance intervals. As an example, the table below represents 7 contiguous GPS records for the same vehicle-id (1C00FFFF-59AE-4AC9-C248-4663F0800000) with derived time and distance intervals, and where segmentId has been initialized as all zeros. The 4th record is a time / distance interval that should be excluded according to the above data segmentation definition.

TS_SEC	timestamp	lat	lon	time_interval_sec	distance_interval_miles	segmentId
1627828385	08/01/2021 14:33:05	36.20	-120.10	30.00	0.25	0.00
1627828416	08/01/2021 14:33:36	36.20	-120.10	31.00	0.19	0.00
1627828446	08/01/2021 14:34:06	36.20	-120.10	30.00	0.18	0.00
1627829715	08/01/2021 14:55:15	36.18	-120.15	1269.00	3.08	0.00
1627829745	08/01/2021 14:55:45	36.18	-120.15	30.00	0.00	0.00
1627829776	08/01/2021 14:56:16	36.18	-120.15	31.00	0.00	0.00
1627829806	08/01/2021 14:56:46	36.18	-120.15	30.00	0.00	0.00

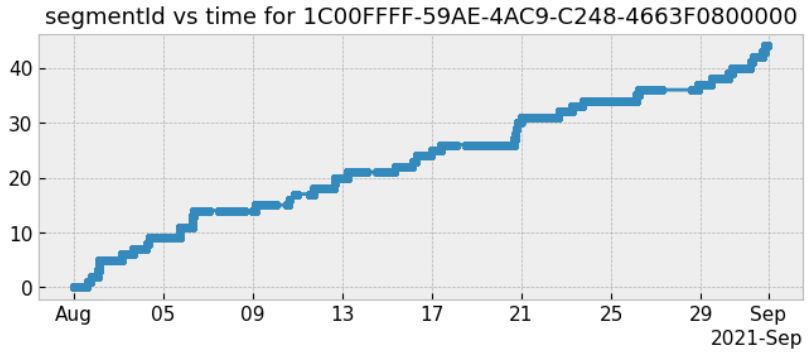
The 4th record is then assigned a null segmentId to identify its exclusion from the set of GPS data segments for this vehicle. Importantly, the 4th record is not removed from the GPS table so that the total volume of records excluded from GPS data segments can be quanitifed later on, eg to validate not too much data are being removed or to identify devices with excessive

time / distance intervals in GPS records. The segmentId for 5th and all subsequent records is then incremented, in this case to a value of 1. Further, time and distance intervals for the 4th record as well as for the 1st record of the new segment are then undefined and set to null accordingly. Based on the algorithm, the update to the table is below.

TS_SEC	timestamp	lat	lon	time_interval_sec	distance_interval_miles	segmentId
1627828385	08/01/2021 14:33:05	36.20	-120.10	30.00	0.25	0.00
1627828416	08/01/2021 14:33:36	36.20	-120.10	31.00	0.19	0.00
1627828446	08/01/2021 14:34:06	36.20	-120.10	30.00	0.18	0.00
1627829715	08/01/2021 14:55:15	36.18	-120.15	NaN	NaN	NaN
1627829745	08/01/2021 14:55:45	36.18	-120.15	NaN	NaN	1.00
1627829776	08/01/2021 14:56:16	36.18	-120.15	31.00	0.00	1.00
1627829806	08/01/2021 14:56:46	36.18	-120.15	30.00	0.00	1.00

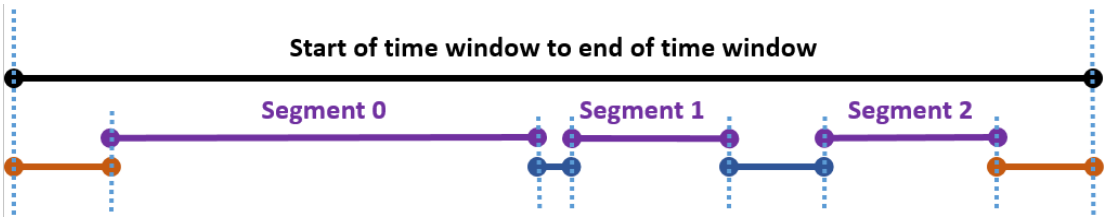
The constraint of a segment being at least two GPS rows (so time and distance intervals can be calculated) leads to the final step in the algorithm to remove any 1-row segments (ie set the segmentId to null in those cases), which may be present in any row depending on properties of the data. The algorithm is then applied by vehicle for all rows identified for exclusion according to the GPS data segmentation definition.

The figure to the right represents the time history of segmentId for the example vehicle in all of Aug 2021. Data segmentation in this case led to 45 distinct GPS segments, with a max segment duration of 3.3 days (segmentId=26) and a min segment duration of 61 sec (segmentId=28).



Data segment model

Many telematics applications assume complete knowledge of vehicle activity over 24 hours per day, such as providing usage and risk metrics to insurance providers. However the presence of long time and distance intervals in raw GPS data means a telematics device may not record 24 hours per day of data coverage (eg at 10sec or 30sec resolution). That is, there may be gaps such as '3.7 miles over 2.8 hours' or '7 miles over 5 min' as in the above examples where the telematics device / data-transfer system, for some reason, do not produce full data coverage in the database as expected. For this reason it is critical to quantify the data coverage that does exist and validate on an application-basis whether that coverage is sufficient. The data segment model serves this purpose; see the figure below.



The data segment model is a physical model that classifies every second of a time window into one of three categories. The data segment model asserts (represented in the above image) the sum of durations of all individual components to equal the duration of the time window (eg a query time window). The data segment model categories are:

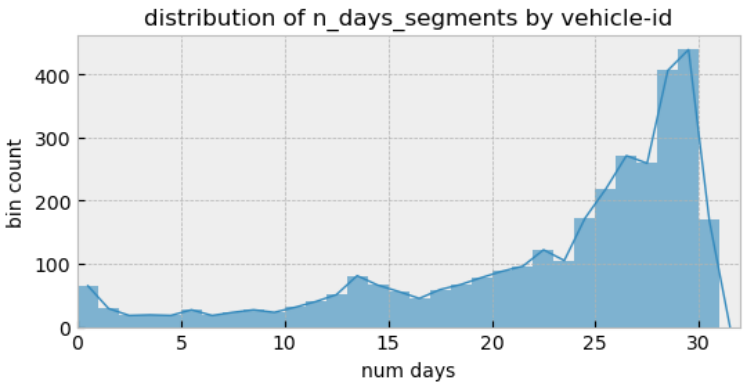
- time between left window boundary and start of first GPS segment, OR
  time between right window boundary and end of last GPS segment **(orange lines)**
- time within a GPS segment **(purple lines)**
- time between GPS segments **(blue lines)**

For the example vehicle in the previous section (1C00FFFF-59AE-4AC9-C248-4663F0800000), exactly 31 days of GPS data were queried from the database representing all of Aug 2021. The data segment model statistics are to the right. Appx 29.8 days (96.1%) of the 31 requested days are represented by 45 GPS segments. Appx 1.2 days (3.8%) of the same 31 days are represented by long time intervals aligned with long distance intervals meeting the data segmentation definition. And appx 0.02 days (0.1%) of the same 31 days are represented by time before / after GPS segment bounds. Cases for other vehicles will be different (see next section on population statistics).

left_window_to_first_segmented_record	0.013
last_segmented_record_to_right_window	0.006
n_segments	45
n_days_segments	29.797
n_days_no_segments	1.185
total_days	31.000

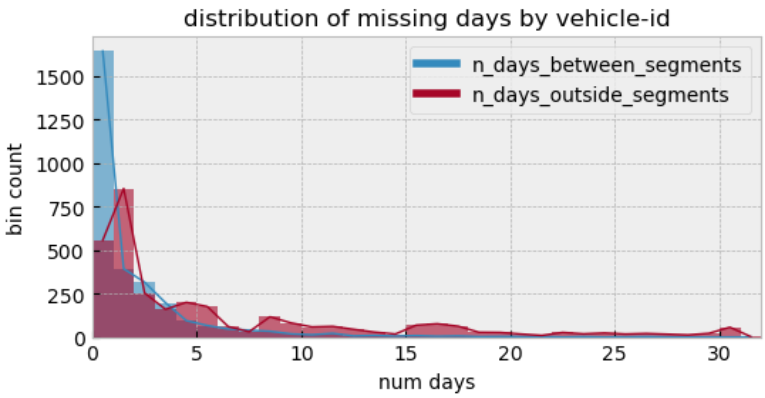
Data segmentation population statistics

The data segment model was applied to all vehicles for the population definition in Appendix A (31 days for each of 3186 vehicles). The distribution to the right represents the number of days covered by GPS segments by vehicle-id. The most common bin is associated with 29 to 30 days covered by GPS segments, ie 94-97% of the 31 requested days. However other vehicles have very low data coverage, and median data coverage by vehicle-id is 83%. Why are data missing? According to the data segment model, there are two reasons why data can be missing (the latter is more common):



- data are missing because removed by data segmentation algorithm
- data are missing because none returned by database before/after segment boundaries

The distribution to the right represents the number of days of missing data by vehicle-id for both of the reasons above. The blue distribution indicates the data segmentation algorithm most often accounts for 0-1 days of missing data for the 31 day time window, whereas 'no data in the database' most often accounts for 1-2 days of missing data and includes a more significant tail out to extreme values such as 30 days. The takeaway is that the more common source of 'missing data' for a vehicle in a time window is due to the database not returning data covering the full time window, as compared to data being removed by data segmentation.



Impact of data segmentation on base usage metrics

Data segmentation will impact metrics derived from GPS data. One example is the metric 'total number of days in motion', which is of interest to insurance providers due to a likely association with collisions. Here 'time in motion' is defined as any full GPS time interval where the corresponding distance interval indicates speed greater than 0.1 mph (appx 10 feet-per-min). Without data segmentation, long time / distance intervals such as '3.7 miles over 2.8 hours' will be recorded as 2.8 hours in motion whereas in reality the full 2.8 hours was not likely in motion.

Total number of days in motion for all vehicles by industry was calculated for the population defined in Appendix A, with and without data segmentation applied; see results below. The impact due to data segmentation is to typically reduce the metric for days in motion by appx 20%, which can be several 100s of days depending on the industry and number of vehicles.

IndustryDesc	n_vehicles	days_motion_without	days_motion_with	diff_abs	diff_%
Waste	143	836.8	692.9	143.9	17.2
Transit	1755	3801.2	3060.9	740.3	19.5
EMS	33	110.4	83.1	27.3	24.7
Distribution	171	1217.8	1017.9	199.9	16.4
Construction	39	68.9	60.3	8.6	12.4
Utilities	5	28.5	21.9	6.6	23.2
Freight/Trucking	1031	7106.8	6187.7	919.1	12.9

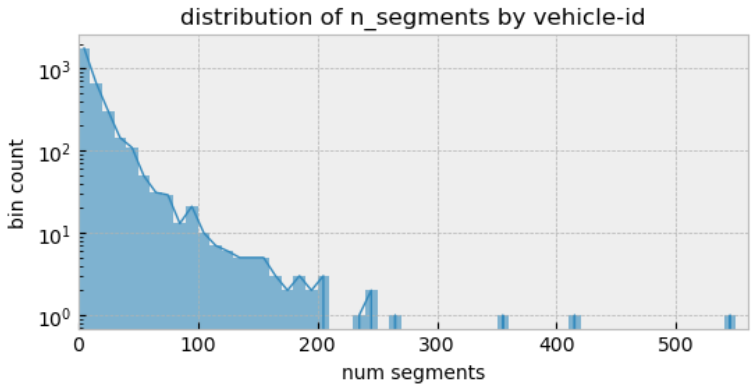
Data segmentation can also impact normalized metrics such as 'miles per day' that is a primary metric of interest to insurance providers; see table below. The column 'mpd_without' means miles per day without data segmentation, where miles is the cumulative sum of all gps record distance intervals, and day is the 31 day query window. The column 'mpd_with' means miles per day with data segmentation, where miles and days are the cumulative sum of all gps record distance and time intervals, respectively, within identified gps data segments only. Differences in miles per day according to these two interpretations can be significant, such as appx 40 miles per day for the Freight/Trucking industry.

IndustryDesc	n_vehicles	mpd_without	mpd_with	diff_abs	diff_%
Waste	143	78.3	92.4	14.0	17.9
Transit	1755	31.0	45.0	14.0	45.4
EMS	33	60.6	78.9	18.3	30.2
Distribution	171	147.5	161.3	13.8	9.3
Construction	39	51.0	80.7	29.7	58.3
Utilities	5	166.3	195.6	29.3	17.6
Freight/Trucking	1031	216.9	255.4	38.5	17.7

Other interpretations of miles per day are also possible, eg by interpreting both as a cumulative sum of gps record distance / time intervals from the first to last GPS record. However in that case non-physical vehicle motion, such as the '7 miles in 5 min' in the above example of driving into the lake, may be included.

Identification of vehicles with unstable GPS data streams

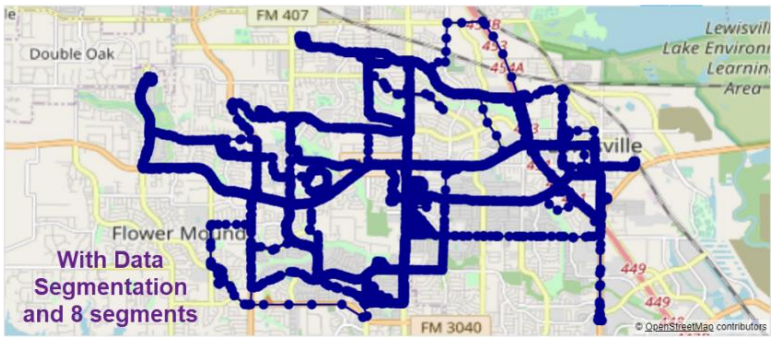
This section explores individual vehicles with a relatively large number of GPS segments, which may be useful as a diagnostic for an unstable GPS data stream. To the right is the distribution of the number of GPS segments by vehicle-id in the population defined in Appendix A (bin width is 10 segments, y-scale is log). A median case (8 segments) and an outlier case (> 200 segments) are then explored individually.



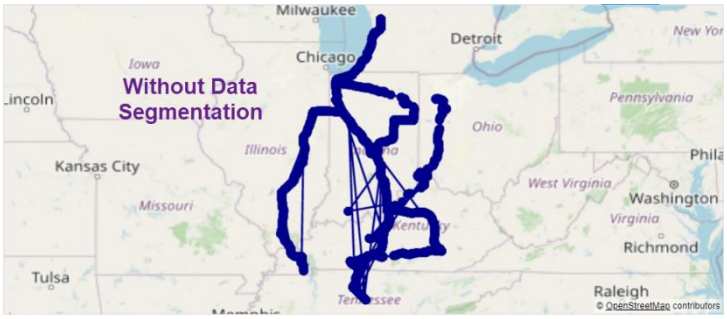
The example to the right represents a median case of 8 segments over 31 days for VehicleId 1C00FFFF-59AE-4BC9-9DBF-4663F0800000, without and then with data segmentation.

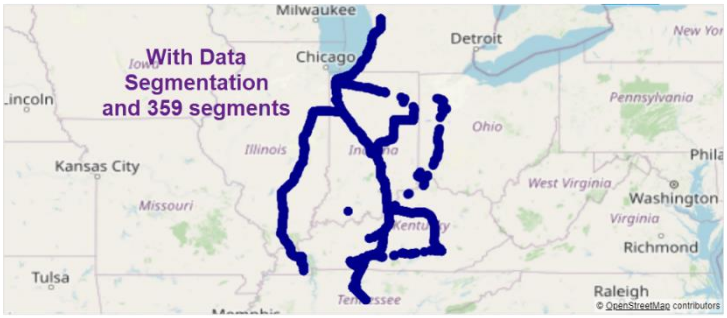
For the case without data segmentation, lines connect all contiguous GPS records, which leads to the long distance intervals marked by arrows.

For the case with data segmentation, lines connect all contiguous GPS records within distinct GPS segments (lines do not connect segments). The long distance intervals no longer show up on the map as they are all now gaps between distinct GPS segments.



The example to the right now represents an outlier number of 359 GPS segments for VehicleId 9100FFFF-48A9-D463-B2D8-9543E7480000, without and then with data segmentation.





As before, without data segmentation lines connect all contiguous GPS records, which in this case leads to many erroneous changes in position as seen on the map. Those erroneous changes in position disappear on the map when data segmentation is applied, however there are now isolated segments of vehicle motion with no indication of how the vehicle arrived at those locations.

A zoomed-in view of the 'with data segmentation' case is below, showing 5 contiguous GPS segments. In some cases, the vehicle device/data system is recording only a few GPS records before cutting out for some reason and then recording again at some point, which is an indication of an unstable GPS data stream.

Data segment model metrics for the same vehicle are to the right. Of 31 days of data requested for this vehicle, only appx 16 days are covered by segments, with appx 13 days in between GPS segments. The implication is there is no GPS data coverage for appx 40% of the time this vehicle was on the road. In this case, the root cause of the missing GPS data coverage is very unlikely to be 'the vehicle is off and being towed' or 'driver personal use' due to the frequency of the issue.

left_window_to_first_segmented_record	0.0
last_segmented_record_to_right_window	1.2
n_segments	359
n_days_segments	16.4
n_days_no_segments	13.4
total_days	31.0

Many more vehicles / devices in similar scenarios could be automatically identified and an alert system could be set up as well.

Summary

The application of GPS data segmentation is motivated by the presence of unresolved large changes in position in raw GPS data that may have a physical or non-physical origin. GPS data segmentation and the data segment model were described in detail and shown to be useful as a data cleaning step, and as a potential diagnostic for an unstable GPS data stream. GPS data segmentation can substantially impact base usage metrics derived from GPS data, and may be included as a data cleaning step in a current project to provide precise vehicle usage, risk, and environmental metrics to insurance providers.

Appendix A – Population definition

- Company names - A/T Transportation, Accent Moving Storage and Logistics, Ace Intermountain Recycling Center, Antonini Freight Express Inc, Appalachian Freight Carriers, Inc., Apple Towing Co, Arrow Limousine Worldwide, Atlas Disposal Ind LLC, Black Gold Express, Inc., Brady Trucking Inc, California Materials, Inc., Edwards Moving & Rigging Inc, EmpireCLS WW Chauffeured Services, Euless B&B Wrecker Service, Green Lines Transportation - NIIC, Illinois Central School Bus, LLC, Island Transportation Corp. - NIIC, JIT-EX LLC, KeyStops LLC, Knight Brothers, LLC, Livingston Trucking, Inc., Mountain Valley Express Co Inc, Nagle Toledo, Inc., Owen Transport Services, PTG Logistics - NIIC, Pneumatic Trucking Inc, Ridge Ambulance Svc Inc, Suburban Disposal Corp., Super T Transport, Terminal Consolidation Co., Tramcor Corp.
- Device model ER-SF300
- Vehicle Id not ‘00000000-0000-0000-0000-000000000000’
- Time window 8/1/2021 to 9/1/2021

Appendix B – Raw queries

Screenshots of raw queries supporting derived results are here:

```
SELECT TS_SEC, LATITUDE, LONGITUDE
FROM DP_PROD_DB.GPS.GPS_ENRICHED
WHERE TS_SEC BETWEEN 1629479892 AND 1629489916
AND VEHICLE_ID = '1c00ffff-59ae-49c9-4e99-4663f0100000'
ORDER BY TS_SEC
```

_ENRICHED 1

ECT TS_SEC, LATITUDE, LONGITI | Enter a SQL expression to filter results (use C

TS_SEC	LATITUDE	LONGITUDE
1,629,479,892	40.4918251038	-89.0284423828
1,629,479,922	40.5005569458	-89.0284805298
1,629,489,856	40.547039032	-88.9939422607
1,629,489,886	40.546749115	-88.9932098389
1,629,489,916	40.5477180481	-88.9926300049

```
SELECT TS_SEC, LATITUDE, LONGITUDE
FROM DP_PROD_DB.GPS.GPS_ENRICHED
WHERE TS_SEC BETWEEN 1627218793 AND 1627221704
AND VEHICLE_ID = '1c00ffff-59ae-48c9-716b-5d43e7800000'
ORDER BY TS_SEC
```

_ENRICHED 1

ECT TS_SEC, LATITUDE, LONGITI | Enter a SQL expression to filter results (use C

TS_SEC	LATITUDE	LONGITUDE
1,627,218,793	40.3441162109	-111.7393264771
1,627,219,498	40.3440322876	-111.7399368286
1,627,219,790	40.2485771179	-111.7808761597
1,627,219,820	40.3443565369	-111.7397460938
1,627,220,488	40.3437652588	-111.7407531738
1,627,221,704	40.3440246582	-111.739730835