Volume Profile 2020

White Paper • Kristian Henrickson • March 2021

INRIX Volume Profile 2020: Presenting the “New Normal” for Traffic Volumes Across the USA

Problem Statement

Each year, INRIX releases a new version of the Volume Profile dataset that represents our best and most current estimate of typical traffic volumes across the entire USA. The dataset provides a cost efficient, readily available resource that enhances decision-making for many purposes, including retail site selection, real estate development, and transportation and road network planning. Improvements to the estimation methodology and coverage are included in each annual release; however, the Covid-19 pandemic brought a new challenge to estimating typical traffic volumes - because nothing about travel was typical in 2020. To respond to the challenge, INRIX data scientists needed to develop new methodologies and solutions for the 2020 dataset to ensure that the many agencies and partners who rely on INRIX Volume Profile have the most accurate and complete estimates possible.

Abstract

INRIX Volume Profile describes the typical on-road vehicle count for each day of week and time of day for over 3 million miles of highways and roads across the USA. A lot has changed in the last year, and it is clear that the magnitude and distribution of traffic volumes in late-2020/early-2021 represent something between pre-pandemic conditions and the “new normal” that will eventually coalesce. This transitory midpoint is what we sought to capture in this year’s update to the Volume Profile product. This white paper describes the methodological enhancements that were made in 2020, not only to provide improved accuracy and consistency over previous years but to also account for the impact that the Covid-19 pandemic had on traffic volumes.

Features of Volume Profile 2020 Include:

Bigger Data - Most of the improvements in the 2020 Volume Profile product are enabled by the addition of several large, higher-frequency new data providers to the INRIX portfolio. This means substantially increased trip volumes over the last year, and more granularity and confidence in trip routing and arrival times.

Better Representation, Less Bias - Consider that each data stream that INRIX ingests comes from a unique combination of products and technologies, and tells us something unique about a particular group of vehicle types and traveler demographics. Thanks to the enhanced INRIX Trips data in 2020, we know more about how different types of travelers are moving, including public transit vehicles, ride sharing, and delivery vehicles.

More Sophisticated Models - Based on the relative trip density in different traveler categories, we can understand and describe local variations in overall trip density attributable to shifts in land use, traveler demographics, and a variety of other factors. As a result, the new modeling framework we use to infer traffic volume is more accurate and more granular than ever before.

Increased Coverage – Volume Profile 2020 covers more than 3.1 million miles of roads, a 21% increase over 2019.

Addressing Pandemic Magnitude and Volume Shifts – The latest volume estimation methodology was built to balance cues from normal, pre-pandemic traffic conditions and the late 2020 incipient recovery. This is done to create a forward-looking view of typical conditions in a time period where the most recent past is not the best predictor of the near-term future.

Contents

Background

Motivation

Traffic volume data is available in some form through many public transportation agencies around the country. One might ask why, then, do we use a complex process involving multiple data sources to devise volume profiles? Why not use publicly available data instead? The answer to these questions is that, despite providing accurate and reliable long-term traffic counts at certain locations, public agency data has limitations that render it unable to meet the coverage and quality requirements of INRIX Volume Profile. The primary limitations of public data for use in devising volume profiles are:

- Lack of coverage – Fixed mechanical traffic counters are only present on a very small minority of road sections. Furthermore, they are most often only present on major highways and, to a lesser extent, major arterials. Ignoring the other limitations of fixed mechanical sensor data, the lack of coverage alone would be sufficient to render it an unsuitable replacement for volume profiles.
- Lack of data standardization – Public agency traffic data comes from a number of data collection technologies and in a variety of data formats. Even if such data were available with sufficient ubiquity to generate volume profiles, collecting data from disparate sources, associating sensors with the road network, and insuring consistency in representation and temporal/spatial resolution is a time consuming and often challenging process.
- Data Quality – Each data collection technology is associated with a unique list of potential data quality challenges. Addressing these different sets of challenges in an appropriate way across all data sources while ensuring consistency is very difficult to achieve at scale.

Document Organization

The remainder of this document is organized as follows:

Chapter 2 describes, at a high level, the major changes and improvements that have been made to Volume Profiles for the 2020 release including new data sources, modeling methods, and treatments to address traffic shifts caused by the 2020 Covid-19 pandemic.

Chapter 3 outlines the data sources that are used in volume estimation.

Chapter 4 details the methodology and model architecture used to estimate traffic volumes and their temporal distribution.

Chapter 5 describes the validation work that was completed and the resulting accuracy measures.

On the final page, we conclude with a brief summary of the work that went into creating Volume Profile 2020 and a glimpse into efforts already underway in preparation for Volume Profile 2021.

Changes and Improvements in Volume Profile 2020

Data Volume

Most of the improvements in the 2020 Volume Profile product are enabled by the addition of several large, higher-frequency new data providers to the INRIX portfolio. This means substantially increased trip volumes over the last year, and more granularity and confidence in the trip routes and road segment crossing times. This represents a major investment in the INRIX product line, and the impact is massive. In many parts of the USA, the increase in trips volume as of early 2020 is more than 10x that of the same time period in 2018, and the mean GPS ping frequency has more than doubled. Figure 0-1 compares trip counts in two polygons in Seattle, WA and Austin, TX. The total number of trips originating in these polygons grew by more than 6.5x in Seattle and 13x in Austin from February, 2018 to the same month in 2020.

The impact of this is twofold. First, with substantially increased sample size and GPS ping frequency throughout the country, Volume Profile can cover more roads with less noise and fewer errors than ever before. Second, this allows us to define new features related to the different probe vehicle characteristics that can help us better predict local variations in probe penetration rate (more on this in the next subsection).

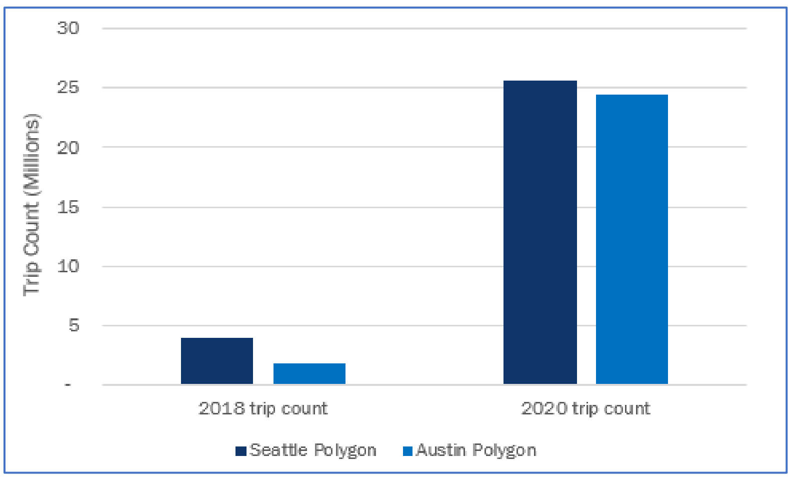

Figure 0-1: INRIX Trip Volume, February 2018 vs. 2020

Modeling Vehicle and Traveler Populations

Consider that each data stream that INRIX ingests comes from a unique combination of products and technologies, and represents a unique distribution of vehicle types and traveler demographics. In aggregate, these streams represent a more or less unbiased sample of the overall vehicle population across large regions. That said, some biases do exist, and the result is that the probe vehicle population is denser in some areas compared to others relative to the overall traffic volume. These variations can be safely ignored for many types of analysis including travel time and performance analysis, but can be quite impactful on the accuracy of volume estimation.

In Volume Profile, we are able to account for local variations in penetration rate by delineating probe data streams into functional categories, each with sufficient trip density to constitute an adequate sample size on all but the smallest roads. Based on the relative trip density in these groups, we can understand and describe local variations in overall trip density attributable to shifts in land use, traveler demographics, and a variety of other factors. As a result, the models we use to infer traffic volume are more accurate and more granular than ever before.

Volume Profile and the 2020 Pandemic

Throughout most of 2020, the world has been in a constant state of flux, and nowhere is this more apparent than in the ways that people are moving around our cities and highway systems. It is clear that some travel and traffic patterns will return to some semblance of pre-pandemic conditions, while others may settle into a post-pandemic "normal" that reflects entirely new trends in the ways we work, spend our leisure time, and interact with each other. Thus, the magnitude and distribution of traffic volumes in late-2020/early-2021 represent something between pre-pandemic conditions and the "new normal" that will eventually coalesce when all is said and done. This is exactly what INRIX has sought to represent in the 2020 Volume Profile release.

By combining probe data from 2019 with probe data from late-2020 and building models to balance cues from both of these time periods, our estimates reflect a slightly forward-looking reality from where we are currently where the near-term future is assumed to look more similar to 2019 than the mid-pandemic traffic volume depression. In a great many places the impact of this modeling approach will be imperceptible, such as in areas where traffic volumes have already fully recovered (or nearly so). The roads most impacted will be those that are still in a transient stage of recovery.

For example, consider the comparison in Figure 0-2, which compares the traffic volume distribution (from permanent traffic recorder data) over days of week / time of day during February and September, 2020. I-80 traffic entering West Sacramento from the west is significantly depressed in September relative to February, and the distribution has shifted from early morning peaks toward relatively strong evening peaks. Conversely, beltline traffic on I-80 in North Sacramento for the months of February and September are nearly identical. This suggests significant shifts in the spatial and temporal distribution of traffic, but somewhat lesser shifts in the overall aggregate volumes.

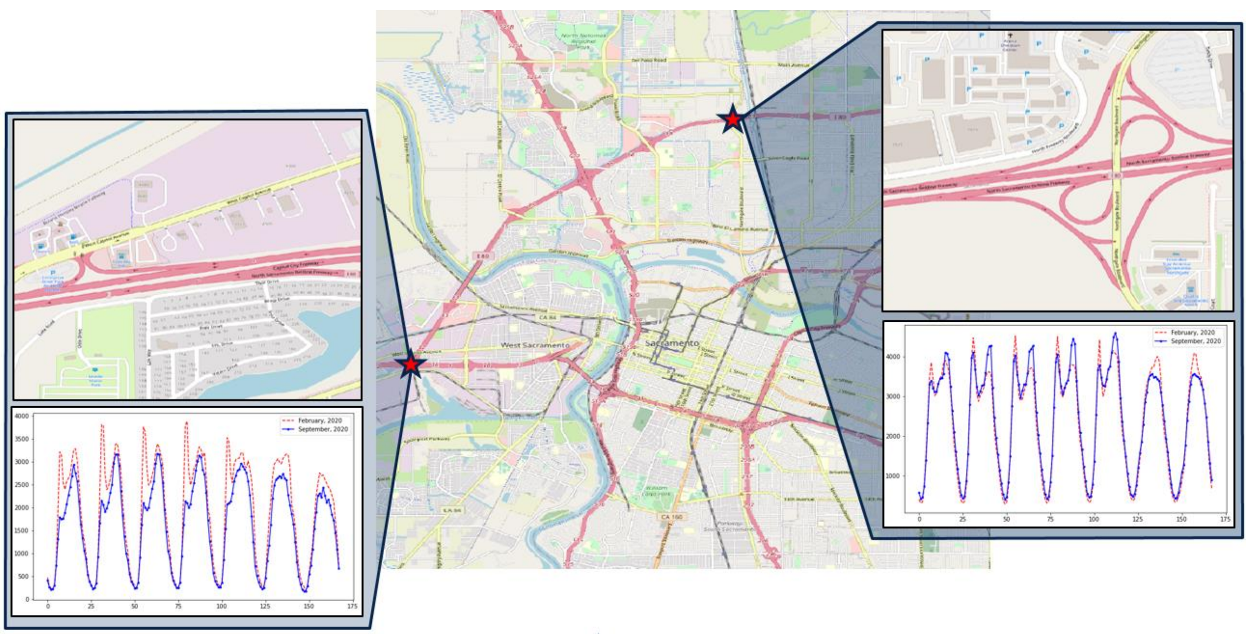

Figure 0-2: Comparison of February and September, 2020 Weekday / Time of Day Volume Distribution for Eastbound Interstate 80 in West Sacramento (left) and Westbound Interstate 80 in North Sacramento (right)

Data Sources

INRIX Trip Paths

INRIX Trip Paths is the most important ingredient in the Volume Profile estimation pipeline. A Trip Path starts with a record of a single vehicle trip including the trip origin, destination, and all GPS points along the driving route. These GPS points are matched to a spatial and topological representation of the road network, and evaluated probabilistically to identify the most likely sequence of road segments traversed on that trip. Measures describing the distance to the assigned road link and the certainty associated with the route choice can be used to filter out outliers and anomalous behavior, resulting in a clean set of segment crossings made by INRIX's massive panel of commercial and personal vehicles. Segment crossing details such as entry/ exit times and speed are the foundation on which Volume Profile is built.

Open Street Maps

Open Street Maps (OSM) data is used both as a source of roadway attribute data and as the basis for the probabilistic map matching that supports INRIX Trip Paths. In order to support Trip Paths, and to provide significantly more granularity in volume estimation, the OSM map is first broken down into a junction-to-junction network. Because this representation breaks at nearly every road access and egress point, it can be safely assumed that volume will be homogeneous throughout the length of every segment. For each segment, a set of attributes is extracted including heading, functional class, and centroid location. Such attributes combined with the trip path crossing details will be the inputs to a series of predictive models used in traffic volume estimation.

Public Traffic Data

Many public agencies around the country maintain a network of permanent, fixed location traffic counters and make the data available publicly. With proper test site selection, quality control, and processing this data may represent the closest thing that exists to a direct measurement of traffic volume at a location. This data is the primary basis for Volume Profiles model training, validation, and quality assessment. In order to avoid large spatial gaps in the training data, supplementary data from public agency short term counts is also used with seasonal adjustment.

As noted previously, mechanical sensor data is not without quality and consistency challenges of its own. Thus, a major component of the work that goes into INRIX Volume Profile is dedicated to addressing the quality and consistency issues present in ground truth data at test locations throughout the country, and applying it assess the fidelity of the devised volume profiles.

Methodology

AADT Estimation

Modeling Approach

A great deal of prior work on probe vehicle based volume estimation has sought to model traffic volume or AADT (Annual Average Daily Traffic) as the outcome variable. The reason for this is simple: most prior work on this subject has not had access to the current INRIX probe vehicle panel, which provides the sample size, stability, and representativeness needed to make unadjusted probe counts a reliable predictor of traffic volume. In volume profiles, we instead predict the volume fraction represented by our probe vehicles or penetration rate and then multiply it by the probe counts to predict volumes.

This arises from a very intuitive interpretation of the modeling task: rather than conflating the prediction of traffic volume and probe sampling idiosyncrasies in a single model, we are simply focusing all efforts on the latter which obviates the need for the former. That is, by adjusting for localized variation in probe penetration rate due to road class, land use, demographic profile, and other factors related to the traveler behavior and driver/vehicle populations, we can make probe vehicle volume the best possible predictor of traffic volume. After all, if the probe data was a perfectly representative sample at a fixed sampling rate across all locations and traveler groups, no model would be needed.

The localized shifts addressed by the modeling process can be observed in, for example:

- More affluent parts of town, where people are more likely to own late model vehicles equipped with in-vehicle navigation systems or vehicles with the ability to interface with mobile phones
- Industrial areas, where freight and contracting vehicles represent a higher-than-average fraction of overall traffic
- Cross-country and scenic travel routes, where navigation apps and services are more likely to be used

Though such regions of higher or lower than average penetration rate are common, the total variation in penetration rate across a city is usually within ±25% of the mean. Traffic volumes, on the other hand, can swing from a few percent of the mean to several hundred percent, which helps to put into perspective the difference in potential estimation error.

Model Architecture

A series of model architectures were investigated, including clustering/nearest neighbors regression, neural network regression, and tree-based methods. Tree-based and clustering methods have several desirable qualities, not the least of which is that outcomes will always be drawn or aggregated from the observed values in the training dataset. Ordinary least squares regression, for example, can predict negative or unreasonably high penetration rates in edge cases.

The robustness and consistent performance advantages provided by random forest regression made it the clear winner in model testing. A random forest regression model is essentially an ensemble or “forest” of decision trees, where the outcome is the average prediction over all trees in the ensemble. Each tree in the ensemble is trained using a subset of the training data and on a subset of the available predictors, which tends to reduce the potential of overfitting and the impact of outliers. Further, random forests (as with many other tree-based methods) are not subject to any assumptions regarding the functional form of the outcome distribution. The internals of a random forest model can be easily parallelized to complete training and prediction efficiently on the massive datasets that support Volume Profile.

Model Inputs

In addition to considering location, heading, functional class, and other roadway attributes, the models used to predict penetration rate utilize features that describe the distinct profile of probe vehicles that tend to travel there. This way, any particular biases that are associated with one or more categories can be addressed in the model training process. Probe data streams are first categorized according the profile of the data source and observed trip characteristics, and probe count associated with each of these "source groups" become inputs to the predictive model. Table 0-1 shows three such probe vehicle source groups or Table 0-1 categories, each of which represents a different distribution of travelers, vehicles, and trip purposes.

Table 0-1: Probe Vehicle Population Groups

	Data Source Group 1	**Data Source Group 2**	**Data Source Group 3**
Vehicle types	Personal vehicle, public transit	personal vehicle, light commercial vehicle	light commercial vehicle, heavy commercial vehicle
Trip purposes	Family shopping / school / activities, school and work commuting, outing / holidays	Services, ride sharing, farm and work, school and work commuters	Long-haul freight, contracting, deliveries
Rural trip density	Low	High	Moderate / High
Urban trip density	Moderate / High	Moderate	Moderate
Suburban trip density	High	Low / Moderate	Moderate
Interstate trip density	Low / moderate	Moderate / high	High
Non-interstate trip density	Moderate / high	Moderate	Low / Moderate

To understand how such a classification scheme can be useful in predicting traffic volume, consider the scenario in Figure 0-3 below, where two nearby highways have significantly different probe vehicle penetration rates. This difference can be explained by several factors, but most notably the fact that the eastbound TX 71 carries a much lower percentage of freight vehicles compared to southbound Interstate 35. These two highway sections are in very close proximity and of similar AADT, functional class, speed limit, etc., which shows that it is not possible to fully describe the differences in vehicle populations present on different road sections using static road attributes alone. When training a model, examples such as these would be problematic because there is no apparent explanation for the difference in outcomes. However, if some representation of the relative volumes of each of the above-described probe vehicle categories are used as predictors, the differences in vehicle populations and associated differences in penetration rate are fully captured.

In this case, the relative volume of Group 3 (in the above classification scheme) on I-35 is more than twice as high as on TX 71. The exact penetration rate of INRIX probe data in each of these categories is not known beforehand, but in a modeling context their relative volume can provide all of the information that is needed to track shifts in the vehicle population and resulting shifts in probe penetration rate.

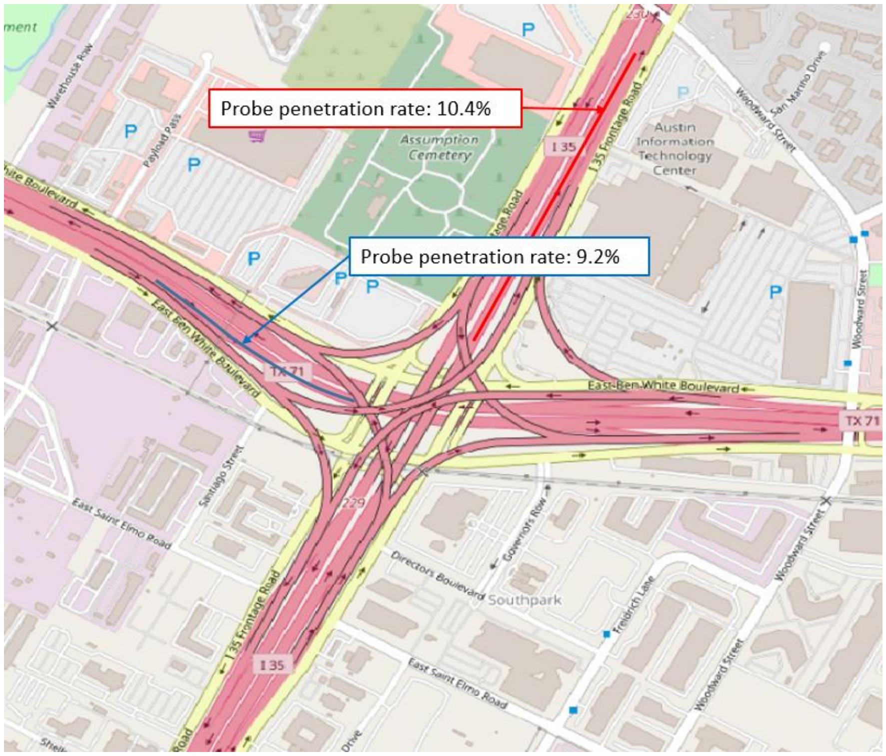

Figure 0-3: Intersection of TX 71 and I-35 near Austin, TX

Of course each group is not homogeneous, and in fact all of the groups overlap to some extent with respect to the vehicles and traveler populations represented. That said, the groups are designed to be a) comprehensive, such that their union captures the full spectrum of traveler profiles, b) informative for modeling, in that each group describes a unique and meaningful dimension of traveler population, and c) interpretable, so that modeling results can be explained and contextualized with respect to real-world phenomena.

Estimating the Temporal Distribution of Volume

Unlike the AADT, the temporal distribution of traffic by day of week and time of day can be directly estimated from the temporal distribution of probe vehicle crossings. This is estimated quite simply as shown below in Equation 0-1. Note that the bin volume fractions are multiplied by 7, so that that the result represents a fraction of AADT. The result of this is a raw distribution, which is then subjected to a compression step to remove noise and, where necessary, infill missing time bins.

Equation 0-1: Volume Time Distribution (as a fraction of AADT)

$$g(h,d) = \frac{7 \times pcount_{h,d}}{\sum_{k\in\{Sun,Mon\ldots,Sat\}}\left[\sum_{m\in\{1,2,\ldots,D\}} pcount_{m,k}\right]}$$

Where:

$pcount_{h,d}$ = total probe vehicle count for weekday d and time of day bin h

D = total number of time bins in a day, 24 for hourly bins and 96 for 15-minute bins

In the compression step, a multi-output regression model is trained using the binned volume fractions (Equation 0-1) as both input and output. The model is trained using the approximately 30%-40% of all road segments where the sample size is greatest and, as a result, the noise is lowest. The result is a near-perfect, slightly de-noised representation of the input raw time distribution features. Further, this allows us to generate high-fidelity profiles for roads with relatively low sample size and even missing time bins.

An example of the result of this step is shown below in Figure 0-4, where the compressed temporal distribution is superimposed on the raw time bin fractions.

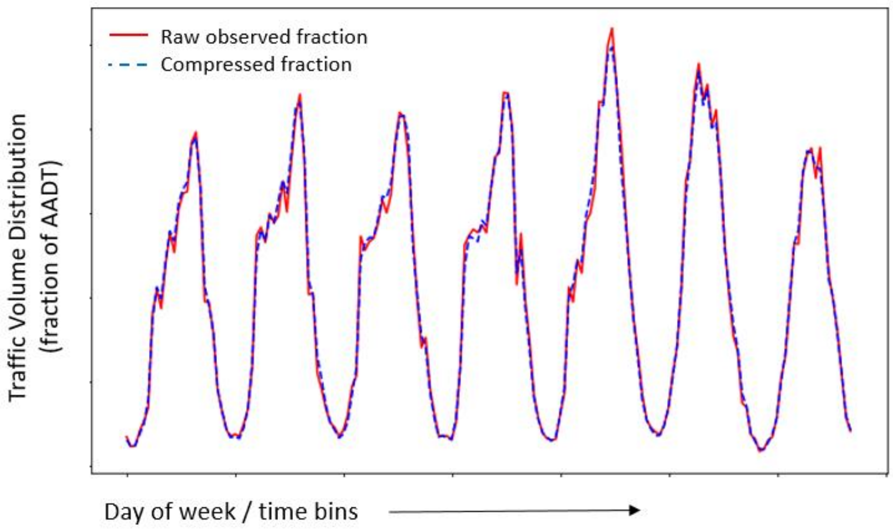

Figure 0-4: Example of a Compressed Volume Time Distribution

On very small roads where the sample size is too low to make the raw temporal distribution features meaningful, a simpler model is used to infer the temporal distribution from nearby roads using heading, functional class, and other road attributes and context features.

Validation

Volume Profile 2020 is not strictly an estimate of the true, on-road traffic volume across any particular time period. Instead, it is designed to represent a mid-point between the still fluctuating conditions in late 2020 and the pre-pandemic normal conditions. Another way to think about this would be volume levels that have returned to near pre-pandemic magnitude in aggregate, but for which the spatial and temporal distributions have shifted toward something in between what was observed in 2019 and late 2020. That is the intent of the validation work described here, by first developing a validation dataset that reflects this near-term future and then applying that dataset in evaluating the accuracy of Volume Profiles.

Data Sources

A validation dataset was developed from permanent traffic recorder stations in Vermont, Washington state, and California. The sites were selected to insure coverage of a wide variety of road types functional classes, as well on the basis of data quality and completeness. Quality control efforts were undertaken to ensure that the traffic counts were complete, free from errors, and observed rather than imputed/infilled by the agency that maintains the sensors.

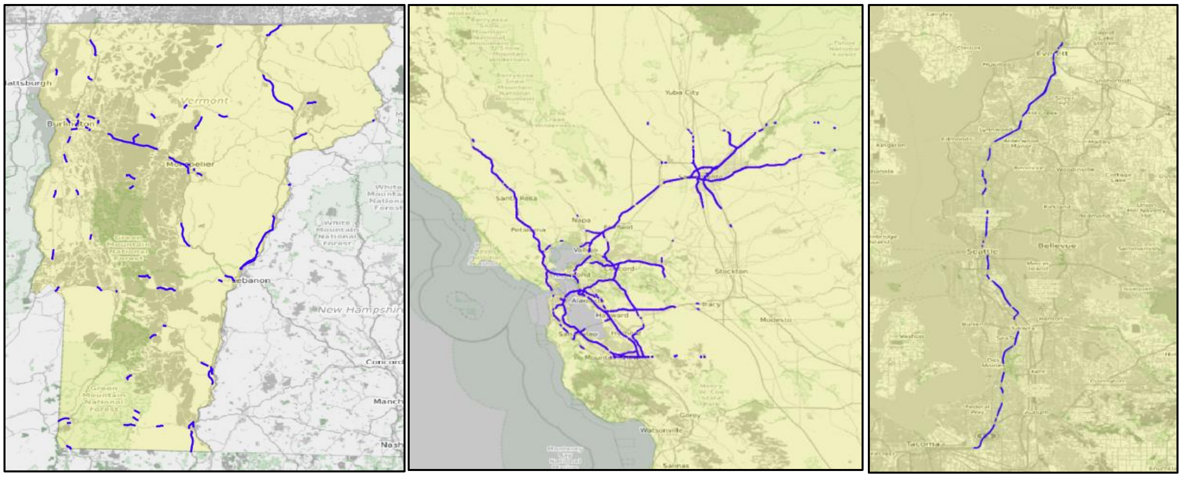

Figure 0-5: Map of Permanent Traffic Counters Located in Vermont (left), the San Francisco/Sacramento Areas in Central California (center), and I-5 in Washington State (right)

Figure 0-5 (above) shows the locations of all traffic recorders as OSM segments, and Table 0-2 (below) shows the number of stations by state. The stations in California represent a wide variety of rural and urban highways, while the stations in Vermont are relatively low volume rural and suburban roads. The stations in Washington State are all along Interstate 5, stretching Tacoma in the South past Everett in the north and passing through downtown Seattle.

Table 0-2: Number of Traffic Count Stations by Region

State	Number of Count Stations
Vermont	80
California	230
Washington	130

Validation Dataset Development

In order to reflect the reality described in the introduction this this section, the ground truth data used for evaluation is collected from early 2020 (pre-covid) and late 2020, seasonally adjusted to reflect annual averages, and then adjusted in aggregate by region to reflect a recovery to early 2020 normal using INRIX Trip Trends. The result is then averaged to get a single volume measure for each road segment and day of week / time of day bin. An example calculation is given below for AADT; the daily/hourly values are adjusted similarly.

$$AADT_i = \frac{1}{|M|} \sum_{m \in M} \frac{ADT_{m,i}}{NVMT_{m,r} \times SF_{m,r}}$$

Where:

$AADT_i$ = Average Daily Traffic for road segment i

$ADT_{m,i}$ = Average Daily Traffic (measured, permanent traffic counter data) for month m and road segment i

M = the set of months used in developing the ground truth dataset

$NVMT_{m,r}$ = Normalized Vehicle Miles Traveled (from Trip Trends) for month m and region r

$SF_{m,r}$ = Seasonal adjustment factor for month m and region r

The regions (r) described in the calculation above are San Francisco, Sacramento, and the states of Washington and Vermont. This means that effectively the same adjustment will be applied to all road segments within a region. $NVMT_m$ is the normalized distance traveled metric used in INRIX Trip Trends, represent the total distance traveled by on-road vehicles in a region divided by the travel distance that would have been typical in early 2020. Practically speaking, it represents the total travel distance as a fraction of what would be expected in the absence of a global pandemic, or the extent to which travel has recovered in aggregate over a region.

Results

The results presented in Table 0-3 are for the directional AADT, meaning the annual average daily traffic by direction of travel (rather than both directions combined, the more common definition of AADT). The mean absolute percent error (MAPE) and root mean squared error (RMSE) are shown by ground truth directional AADT bin, where the RMSE is shown as a percentage of the mean directional AADT for that bin. The estimated and ground truth directional AADTs are plotted in Figure 0-6, where the colors indicate the region where the traffic counters are located. The accuracy of prediction is high and consistent across road classes and traffic volumes, and shows that the estimation methodology performs well even on the most challenging low-volume roads.

Table 0-3: Directional AADT Accuracy Evaluation

AADT Bin	RMSE	MAPE	Station Count
0 - 5k	14.8%	8.9%	36
5k - 10k	11.1%	9.9%	27
10k - 25k	15.9%	14.3%	14
25k - 50k	14.1%	11.1%	44
50k - 100k	10.9%	8.9%	180
> 100k	11.5%	9.2%	90

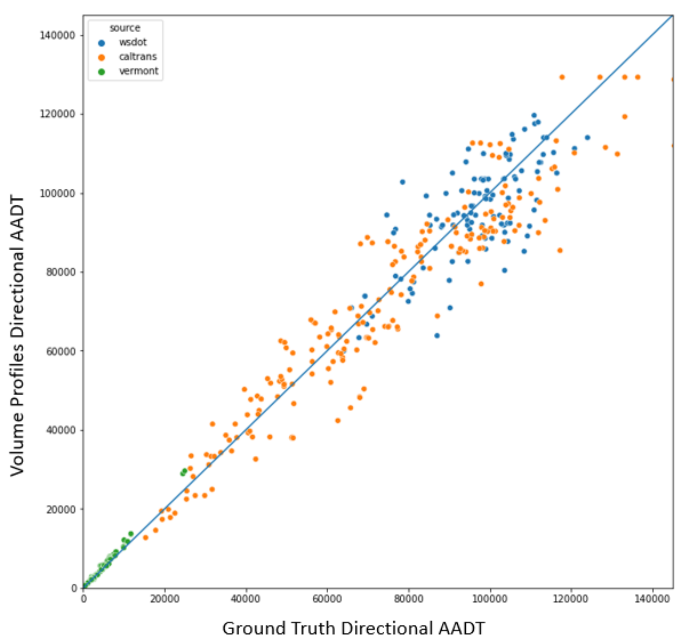

Figure 0-6: Volume Profiles vs. Ground Truth Directional AADT

Table 0-4 shows the MAPE and RMSE by day of week (rather than AADT) for all AADT bins. The accuracy is quite good and fairly uniform across all days of week, which suggests that the probe count distribution is a reliable predictor of the temporal distribution of traffic across weekdays.

Table 0-4: Average Day of Week Traffic Volume MAPE and RMSE

Day of Week	RMSE	MAPE
Sunday	14.8%	9.8%
Monday	13.7%	11.3%
Tuesday	14.0%	11.6%
Wednesday	14.9%	10.0%
Thursday	14.8%	10.4%
Friday	13.9%	12.8%
Saturday	18.2%	12.5%

Table 0-5 below shows the accuracy measures for the most granular hour of day and day of week data, where time of day is filtered to only include the hours from 7:00 AM to 8:00 PM. INRIX volumes are underestimating by approximately 39 vehicles per hour on average, while the mean percent error is positive at 1.3%. This indicates that low volume roads and time periods are overestimated slightly, while higher volume roads and time periods are being underestimated to a small degree. As shown in Table 0-6, the accuracy generally increases with AADT. Figure 0-7 shows a visual comparison of the ground truth and estimated traffic volumes by the same day of week and time of day bins.

Table 0-5: Day of Week / Hour of Day Accuracy Measures (Overall)

Mean Error (veh / hour)	-38.9
MAPE	14.0%
MPE	1.3%
RMSE	18.0%
R^2 of Prediction	0.91

Table 0-6: Day of Week / Hour of Day Accuracy Measures (By Directional AADT Bin)

AADT Bin	RMSE	MAPE	Station Count
0 - 5k	29.1%	18.4%	36
5k - 10k	17.8%	15.2%	14
10k - 25k	22.0%	17.4%	44
25k - 50k	19.3%	15.3%	180
50k - 100k	16.3%	13.2%	27
> 100k	15.3%	12.4%	90

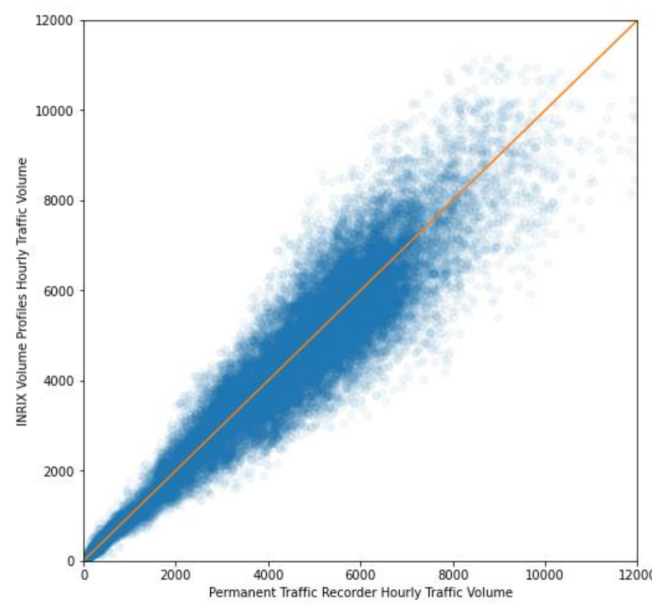

Figure 0-7: Average Traffic Volume by Day of Week / Hour of Day, Estimated vs. Ground Truth

Conclusions & What's Ahead

Each year, a great deal of data science effort goes into updating, refining, and validating the INRIX Volume Profile methodology, as well as incorporating lessons learned over the past year from customer feedback and internal analysis. However, 2020 was a unique year for Volume Profile in a couple different ways. First, the traffic impacts of the Covid-19 pandemic were so large and rapidly evolving as to require a major retooling of our methodology. Second, changes in the INRIX GPS data panel over the previous year provided the massive lift in probe vehicle volumes needed to support improvements in the accuracy and granularity of our models while increasing coverage. The result is a better product, and a unique solution to the challenge of representing "typical" volumes in a year that has been anything but.

Future work in volume profiles is focused on a couple key initiatives. First, efforts are underway to offer monthly volume profiles (rather than an annual average), which will provide a more up-to-date view as well as the ability to look at volume trends over time. Second, we are working to automate and enhance anomaly flagging, quality monitoring, and alerting. This will be crucial to moving to a monthly release cadence, and will improve our responsiveness to issues that arise.