STREETLIGHT INSIGHT

All-Vehicles Volume Methodology and Validation White Paper

Canada

Updated February 2023

Table of Contents

Introduction

StreetLight Data, Inc. ("StreetLight") Vehicle Volume provides a quick, easy, and cost-effective way to measure traffic at the yearly, monthly, daily, and even hourly levels. Volume estimates can be derived for any location, such as a road, park, Dissemination Area, or user-defined area. It can also be used to gauge zone-to-zone traffic, providing comprehensive estimates for projects like turning movement studies and travel demand models. StreetLight Vehicle Volume is available for analyses in the U.S. and Canada.

The StreetLight team is continually updating the data sources and improving the algorithms that power its Metrics. This white paper provides technical detail about the methodology, algorithm development, validation, and data sources used in StreetLight's Vehicle Volume output for All Vehicles in Canada. This white paper was first published in October 2019 and is updated periodically to reflect the most recent methodology and validation.

What's New in the February 2023 Update

Starting in February 2023, Canada Volume Metrics in the StreetLight InSight® platform will reflect updates to the methodology outlined in this paper. The updates apply to any StreetLight analysis for the All Vehicles mode that uses "Volume" as the selected output type for the months of January 2019 through the present.

- **Expanded training and validation data:** We obtained data from more than 1,900 additional traffic counters to use in training and validating the Volume model. Many of the new counters are in downtown areas, where our counter data was previously sparse. Monthly traffic counts now come from four provinces instead of just two, expanding coverage in the eastern part of Canada. This new training and validation data has lessened spatial bias in the Volume model and greatly reduced errors at counter locations, especially in downtown areas.
- **Improved monthly adjustment model:** The monthly model, which applies a seasonal adjustment to annual average daily traffic estimates, now uses a more sophisticated machine learning algorithm. This improvement results in lower errors and better month-over-month trends.

Methodology

Model Overview

We employ multiple machine-learning models and pull from several data sources to generate StreetLight All Vehicles Volume estimates. These models have gone through several iterations over the years as we have learned from experience (including validations conducted by our customers) and gained access to new data sources.

In the following sections, we use the terms Volume, MADT, and AADT. Volume refers to traffic volume generally and can apply to any period of time. MADT (Monthly Average Daily Traffic) is a more specific measure that refers to the average daily traffic volume in a given month. The term AADT (Annual Average Daily Traffic) refers to the average daily traffic volume over a specific year.

Data Sources

INPUT FEATURES

The StreetLight All Vehicles Volume models are built to use the following sources of input data.

- **Location-based services (LBS) trips:** Our processing algorithms link pings from LBS devices into trips, then classify these trips into travel modes. The Volume models use trips that we have identified to be from vehicles.
- **Navigation GPS trips:** The Canada models use trips derived from commercial and personal GPS data.
- **Census data:** We use demographic data like population, number of housing units, and whether or not a location is in an urban area.
- **OpenStreetMap (OSM) data:** OSM provides information on road classification, roadway characteristics, such as number of lanes, and speed limit, among others.
- **Weather data:** Temperature, precipitation, and visibility affect travel and help us better predict seasonal changes.

GROUND-TRUTH DATA

To provide ground-truth data on traffic volume, we rely on traffic counters in a variety of locations. We have acquired and processed counter data from 3,487 sites across Canada, representing a mix of small and large roads and urban and rural locations.

A mixed-model approach is used to estimate MADT. This approach starts with an estimate of AADT for a given location in 2019. As shown in Table 1 and Figure 1, the 2019 AADT model is built from 3,487 permanent counters in nine provinces across Canada: Alberta, British Columbia, Manitoba, Nova Scotia, New Brunswick, Ontario, Prince Edward Island, Quebec, and Saskatchewan.

Model	Year of Counter Data	Feb 2023 Model Update		Previous Model Version	
		Number of Provinces	Number of Locations	Number of Provinces	Number of Locations
Canada AADT 2019 (base model)	2019–2020	9	3,487	9	1,832
Canada MADT (monthly models)	2019	4	1,160	2	419
	2020	4	1,334	2	420
	2021	4	1,372	2	414
	2022	3	416	2	396

Table 1: Number of counter locations used in training the Canada AADT and MADT models, by year. Some locations may have counters in more than one travel direction. Data for 2022 includes only January–April.

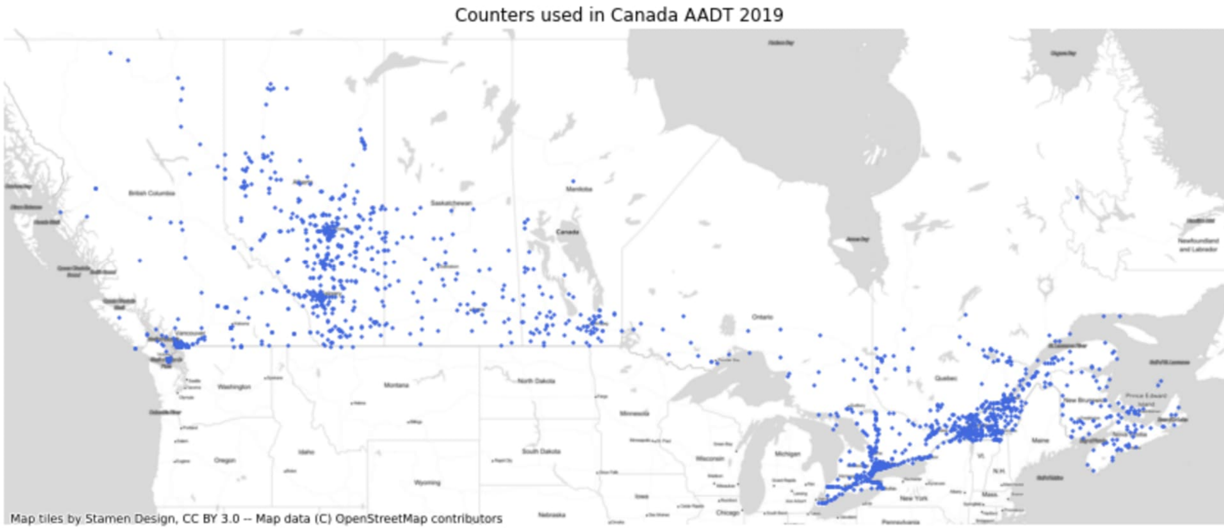

Figure 1: Locations of permanent counters used in the Canada AADT 2019, which is the base model for Canada MADT.

To generate MADT estimates, a second set of models estimates a seasonal adjustment. These models rely on counters that report MADT on a rolling monthly basis. For the February 2023 update to monthly models, we obtained data from hundreds of new counters, increasing our total number of monthly counter locations from 422 to 1,469. (Some of these locations have counters in more than one travel direction, bringing the total number of counters to 2,313.) Counters represent all road types and sizes, with the majority of counters on roads of 5,000 to 35,000 in average daily traffic.

The new counters have improved the spatial coverage of our training dataset. We previously used monthly data only from Alberta and the city of Surrey, BC due to limited data availability. For the February 2023 update, we've added data from permanent counters in Calgary, Winnipeg, Vancouver, and Toronto. Figure 2 shows counter locations used in the monthly Volume models. By adding counters in Toronto and Winnipeg, coverage is no longer limited to just the western part of Canada. Also, whereas the previous model version relied heavily on highway counters and had limited representation of local urban streets, we now have many counters on roads in urban cores, which reflect unique traffic conditions and patterns. This expanded coverage is expected to reduce spatial bias in the Volume Metric. At the same time, since the majority of counters are still in Alberta and British Columbia, the model likely remains biased toward those locations.

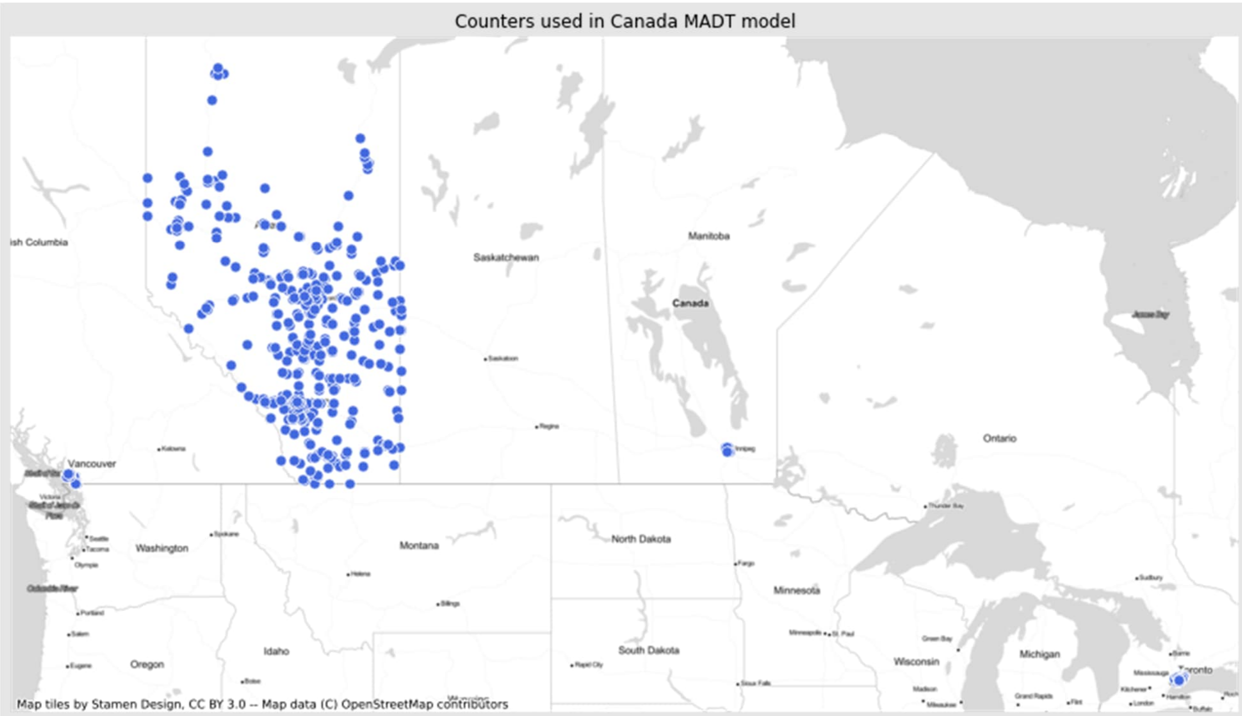

Figure 2: Locations of permanent counters used in MADT models in Canada. The February 2023 updates added counters in Vancouver, Winnipeg, and Toronto. Previous versions relied on counters in Alberta and Surrey, BC, only.

Figure 3 below shows the distribution of counter data availability over time. Most of the newly added counters provide data for January 2019 through June 2021. Many of the new counters located in British Columbia and Alberta did not have data available in July 2021 onwards, while those in Toronto do not have updated data for 2022. As a result, the MADT models may not perform as well in those locations in late 2021 and 2022.

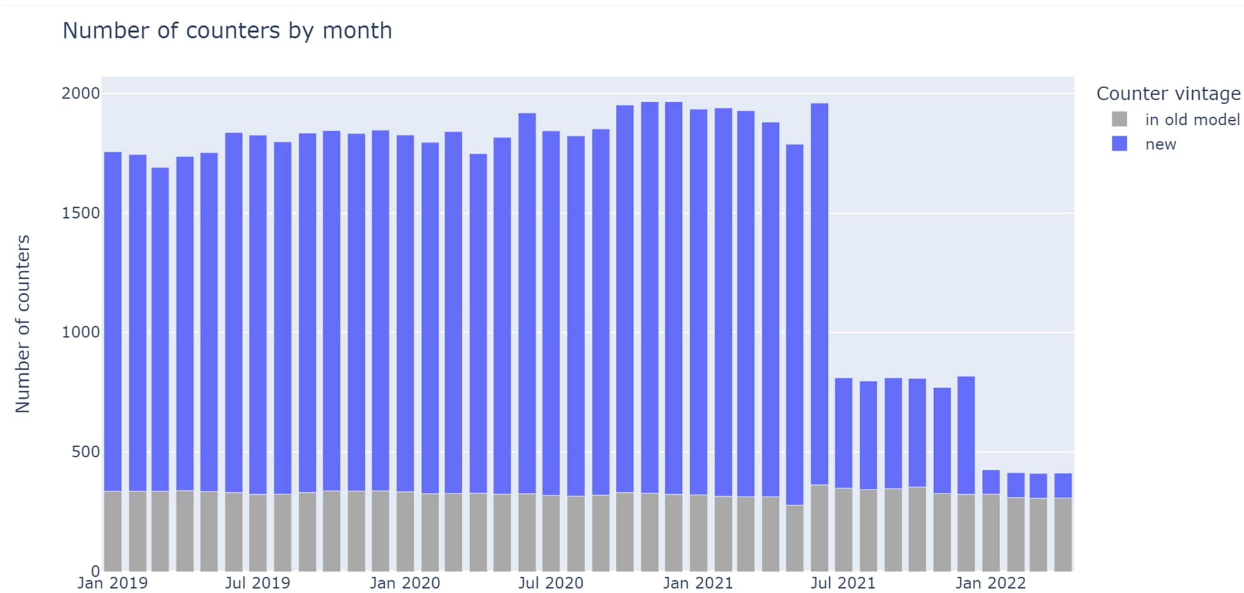

Figure 3: Number of counters by month used for training and validation in the updated model. The number of counters represents unique counters and travel directions.

Algorithm Details

The Canada Volume algorithm uses 2019 AADT as a base estimate of traffic on an average day in 2019. We chose to use 2019 as the base year because it was unaffected by the COVID-19 pandemic and because availability of ground-truth data is good. The AADT base models use a combination of 43 input features, including vehicle trips processed by our multimode algorithm. These features are entered into a pair of models: a model for low-volume roads and a main model for all other roads. The models are fit using the Extreme Gradient Boosting (XGBoost) algorithm.

Once we have the base estimate of average daily traffic in 2019 for any given month, we need to apply a seasonal adjustment factor to estimate monthly average daily traffic. To generate this seasonal adjustment factor for each new month from 2019 to the present, we train a set of three models using vehicle trip data and other features as inputs. Roads with very high volume are modeled separately using XGBoost. For all other roads, a pair of XGBoost and Random Forest models are trained. Because each of these models has different strengths — the XGBoost excels at seasonal patterns while the Random Forest guarantees estimated volume are not negative — we blend their outputs in a way that optimizes accuracy and comparability. We then apply that seasonal adjustment factor to the base AADT estimate to produce an estimated MADT for the given month.

Prediction Intervals

Our Volume estimates also include prediction intervals that represent the likely upper and lower bounds of the true value. The 95% prediction interval is an estimate of the range within which the true value is expected to lie 95% of the time. To construct these intervals, we fit a quantile regression model to the MADT errors. The prediction intervals included with StreetLight's Metrics are only available for All Day, Average Day estimates (not for specific day types or day parts) and only when there are at least 25 days in the input period. It should be noted that prediction intervals are only reported when analyzing pass-through zones and are not available when analyzing area zones or conducting Origin-Destination analyses.

Estimating Volume from the MADT Model

Volume for Specific Time Intervals

The Volume models described in the previous section estimate monthly traffic volumes (MADT). To generate Volume estimates for other time intervals, such as for specific days of the week or times of day, we use MADT as the base, then apply a time-specific scaling factor. The scaling factor is equal to the ratio of LBS trips for that time interval to LBS trips for that month. Our assumption here is that the ratio of StreetLight-estimated trips to actual trips for a location is stable throughout the month. Possible trip sample bias by time of day or day of week is not presently accounted for in our Volume Metric.

Volume for Areas

The MADT models are designed to estimate Volume for roads, because ground-truth data from permanent counters is available for road segments. We use an additional algorithm to develop Volume estimates for areas. In order to estimate Volume for a specified area, the algorithm selects a subset of roads with trips that start, end, or pass through the zone area. See Figure 4 for an example in which, a specified area (shaded blue) is accompanied by a subset of randomly sampled roads (yellow polygons) in the surrounding area. The number of sampled roads will depend on the size and location of the area zone.

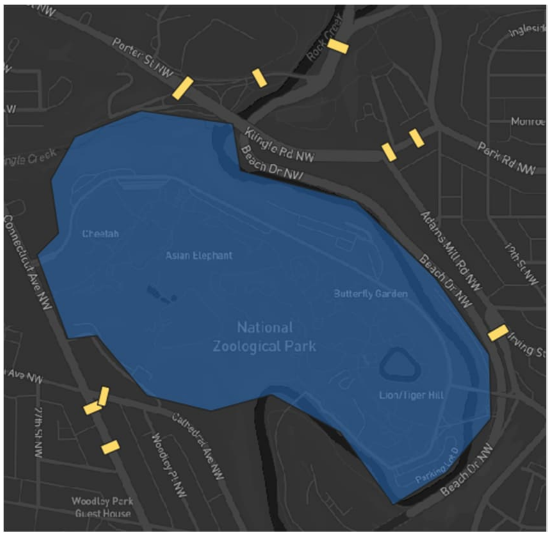

Figure 4: Example area zone with selected gates (yellow polygons) used to calculate MADT for trip starts and stops in the area shaded blue.

For each sampled road, our algorithm does the following:

1. Generates an estimate of MADT from each sampled road.
2. Obtains an estimated trip sample penetration at each sampled road. This is based on the assumption that the sampled roads provide an estimate of the StreetLight trip sample penetration for the area zone.
3. Calculates a weighted average trip penetration from the sub-sampled roads and applies that to the LBS trip count of the area zone to derive a final estimated StreetLight Volume.

This results in an estimated Volume for the defined area based on trip starts and ends.

Volume for Origin-Destination Analyses

After estimating Volume outputs for individual zones (both pass-through and area zones), we apply them to Origin-Destination analyses, which allows for evaluating the number of trips that span locations. The goal is to generate an O-D Volume that provides a number representing a reasonable estimate of the real-world number of trips.

This approach includes the following steps:

1. Calculate the total Zone Activity Volume for each Origin and Destination zone (described in the previous sections).
2. Return the LBS trip counts between each Origin-Destination zone pair.
3. Scale the LBS Origin-Destination sample trip counts to Volume based on the estimated total Volume at each Origin and Destination.

Validation

Validation Methodology

The Canada AADT and MADT models were validated using leave-one-out (LOO) cross-validation. In this technique, one data point (in our case, one counter) is held out and the model is trained on the remaining data points, then error is calculated from the held-out data point. This process is repeated for each data point, then the errors are averaged. We chose LOO cross-validation because it makes the most of the limited training data points available for Canada.

We use the following set of error metrics to evaluate estimation accuracy:

- **Mean absolute percentage error (MAPE)** is a standard measure of error for predictive machine-learning models. It describes errors well on small roads and treats errors of all sizes equally.
- **Normalized root mean square error (NRMSE),** another standard error metric in machine learning, penalizes large errors more, making it more sensitive to the accuracy of AADT estimation on high-volume roads.
- **68th percentile absolute error** represents one standard deviation from the mean within a standard bell curve and thus describes the typical error across the road segments. The 68th percentile absolute error can be interpreted as the expected typical error.
- **95th percentile absolute error** represents the upper limit of the expected absolute error.
- **Median percentage error** is a measure of bias in our model, which also indicates how accurate an aggregate value (such as the sum of volume on all roads in this bin) will be. Values close to zero suggest that our model has low bias, while positive values would indicate more overestimation and negative values indicate more underestimation.

Validation Results

This section presents cross-validation results for estimated MADT in the months from January 2019 to April 2022. For AADT results, please refer to the *StreetLight AADT 2019 and 2020 Methodology and Validation White Paper for Canada.*

It is, useful to look at the correlation between StreetLight estimated MADT and MADT as reported by counters. A strong correlation is a sign of low errors and helps us visually identify places where estimates may be too high or too low. Figure 5 shows the correlation between StreetLight estimated MADT and MADT measured by the 2,313 permanent counters used in the monthly models, for each year. The correlations are all high, with few points deviating far from the best-fit line. R-squared values for the best-fit lines are 0.95 for 2019, 2020, and 2022; for 2021 it is 0.96.

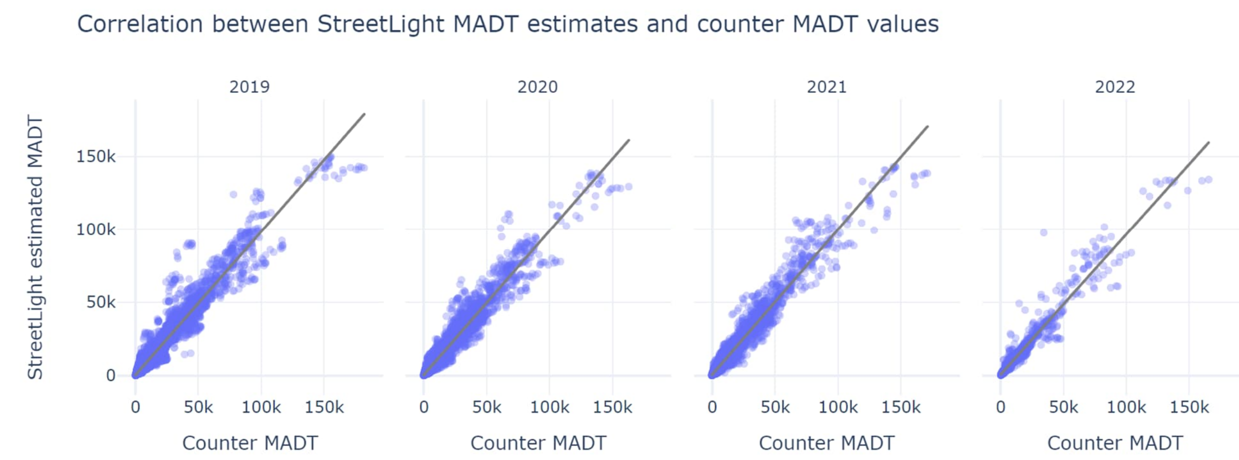

Figure 5: Correlation between StreetLight MADT and MADT as reported by counters, by year. The 2022 data includes January through April only. The best-fit lines have R2 values of 0.95 for 2019, 2020, and 2022; for 2021, it is 0.96.

Table 2 shows error metrics for MADT as compared with counter data, by road size. We report metrics by road size, measured by AADT, because lower-volume roads inherently have higher percent errors. As we can see in Table 2, bias is relatively low across road sizes. For roads over 500 AADT, bias is slightly positive, indicating the models tend to overestimate volume by a small amount. For roads of less than 500 AADT, the models tend to underestimate, by 13.4% on average. The values for 68th percentile absolute error show a typical error is about 15–24% on roads of over 500 AADT, while the 95th percentile absolute error values suggest errors may be as high as 60–70% in rare cases. All error metrics tend to decrease with road size, as expected.

Road Size (AADT)	**Count**	**Median Percent Error (Bias) (%)**	**68th Percentile Abs Error (%)**	**95th Percentile Abs Error (%)**	**MAPE (%)**	**NRMSE (%)**
A: 0–499	1,446	-13.4	33.5	67.9	28.2	34.0
B: 500–1,999	12,413	0.3	23.5	67.8	22.7	28.7
C: 2,000–4,999	14,539	1.8	23.0	58.6	21.0	26.9
D: 5,000–9,999	11,359	3.3	20.3	63.9	20.7	22.2
E: 10,000-19,99–	10,659	0.6	20.0	57.3	19.6	23.1
F: 20,000–34,999	6,360	3.6	18.5	68.1	19.5	23.3

G: 35,000–54,999	2,585	2.3	16.4	46.4	16.3	21.0
H: 55,000–84,999	1,484	5.3	22.1	50.7	19.5	22.7
I: 85,000–124,999	400	5.8	14.7	24.1	11.4	16.4
J: 125,000+	518	3.0	15.8	58.7	16.6	18.0

Table 2: Cross-validation results for StreetLight estimated MADT, by road size, for January 2019–April 2022. Road size is defined by the counter's AADT in 2019. "Count" is the number of data points; each counter may have data for multiple months.

As shown in Figures 6 and 7, the validation results demonstrate the updated models have markedly improved error rates over the previous version. These charts show errors for the new models, compared with errors that are calculated using the old model to generate estimates at both old and new counter locations. (Note that the old model errors differ from those reported in previous white papers because ground-truth data for these counters was not available in the previous version.) As Figure 6 shows, the updates have reduced errors across nearly all road sizes and years. Roads of 5,000 to 54,999 AADT, where most permanent counters are located, have the most substantial improvements.

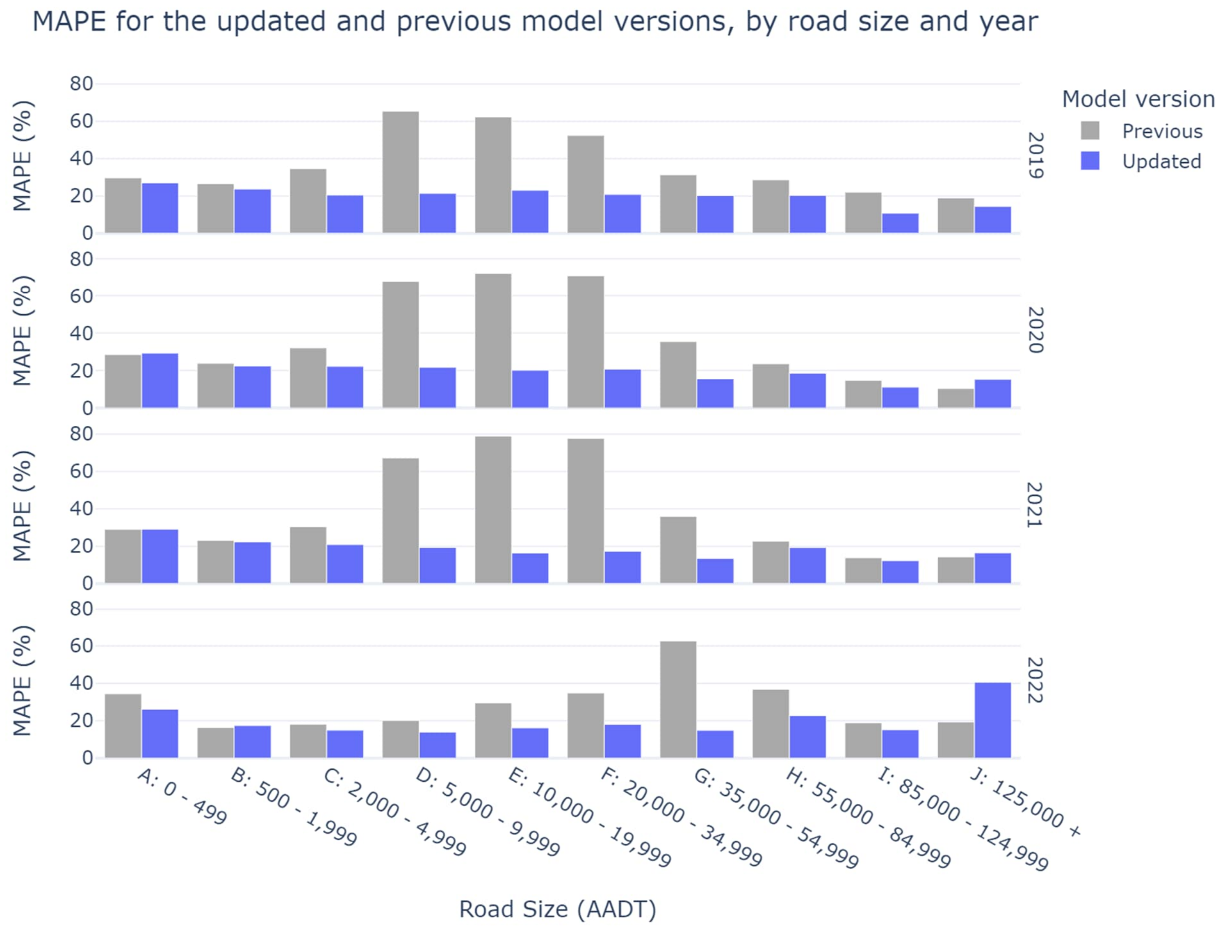

Figure 6: MAPE for the updated model version, compared with the previous model version, by road size and year. Previous model values include errors for counters used in the previous model, as well as new counters.

The model updates have also dramatically reduced bias at counter locations, as shown in Figure 7. Similarly, the largest improvements are in medium-sized roads, while the largest and smallest road size categories sometimes have slightly higher errors than before.

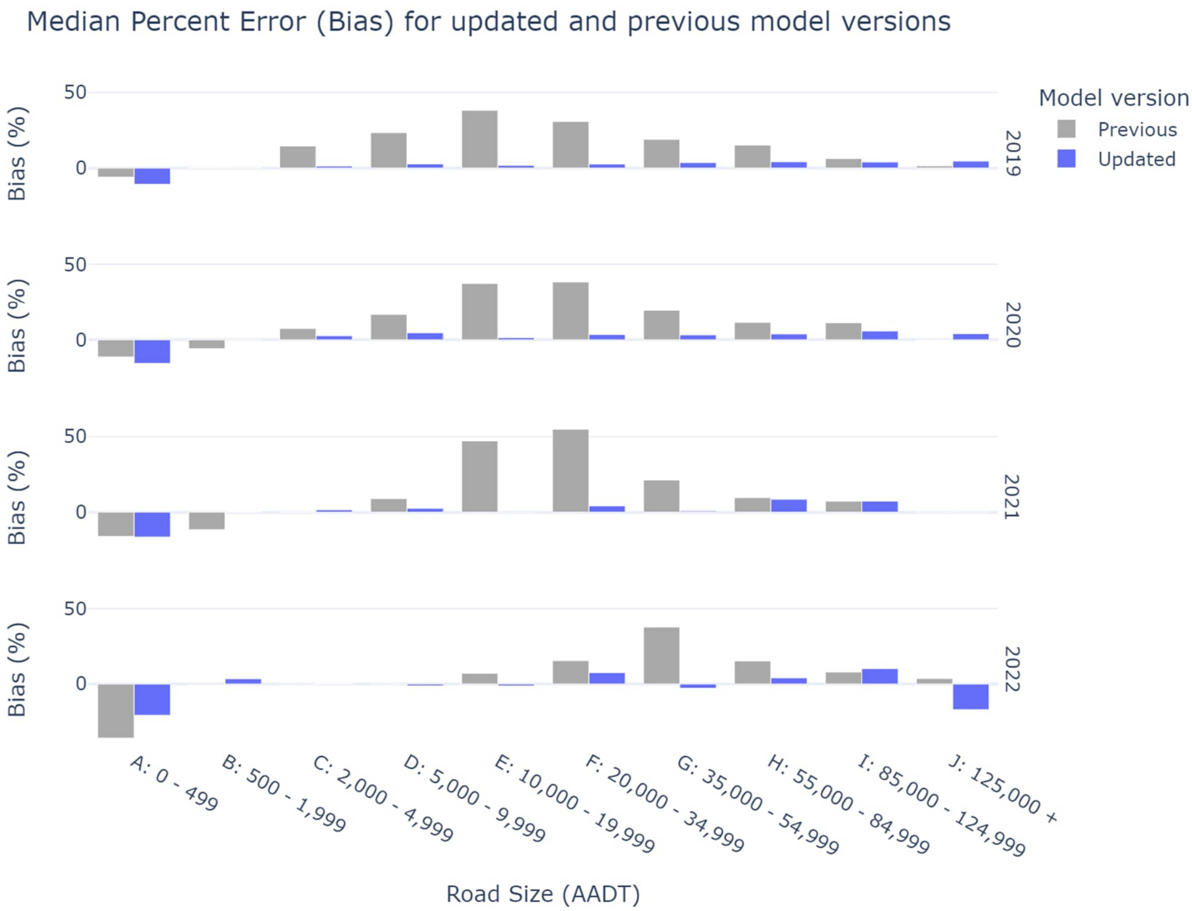

Figure 7: Median percent error (bias) for the updated model version, compared with the previous model version, by road size and year. Previous model values include errors for counters used in the previous model, as well as new counters.

The largest reductions in errors occur in densely populated areas, thanks to the addition of training data from counters on streets in urban cores (see Figure 8). At counter locations with populations of more than 5,000 per square kilometer, the updated models perform much better than the previous version.

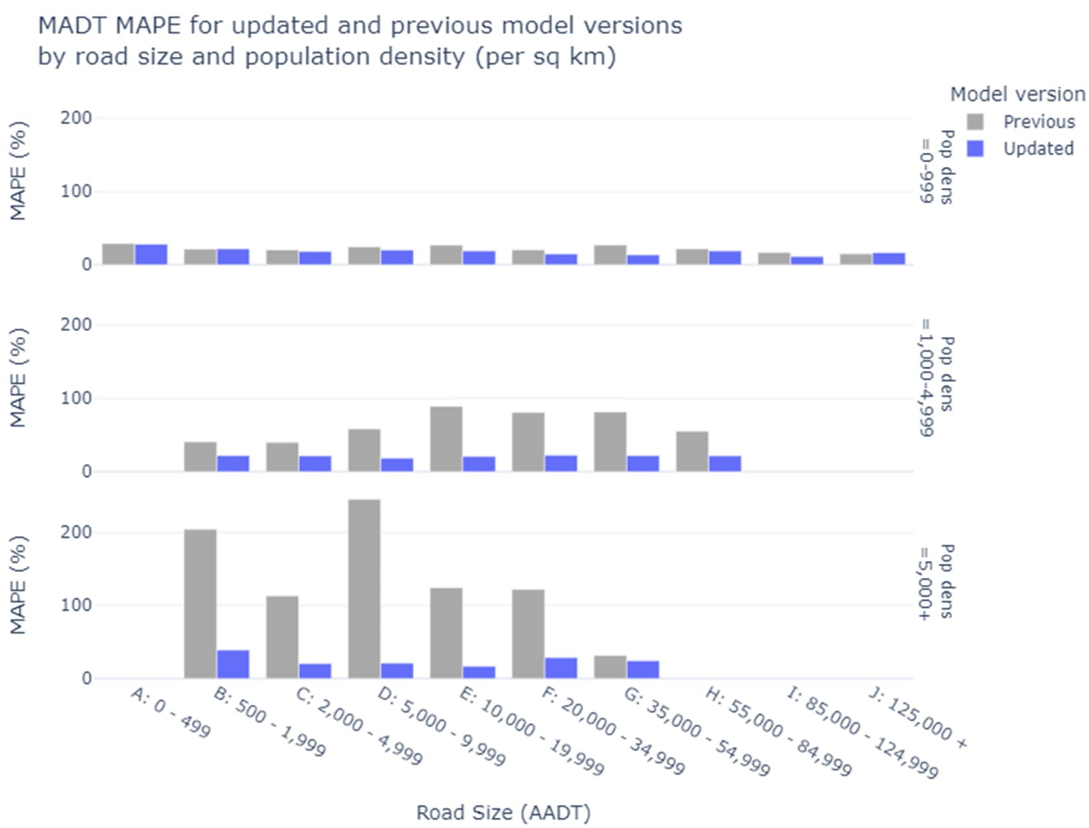

Figure 8: MAPE for the updated model version, compared with the previous model version, by road size and population density. Previous model values include errors for counters used in the previous model, as well as new counters.

As shown in Figure 9, errors drop the most in provinces with new counter data: British Columbia, Manitoba, and Ontario. In Ontario, MAPE at counter locations with less than 55,000 AADT dropped from about 150% with the old model to under 25% with the updates. As a reminder, the previous version of the Volume model had no monthly counter data in Ontario, resulting in higher errors.

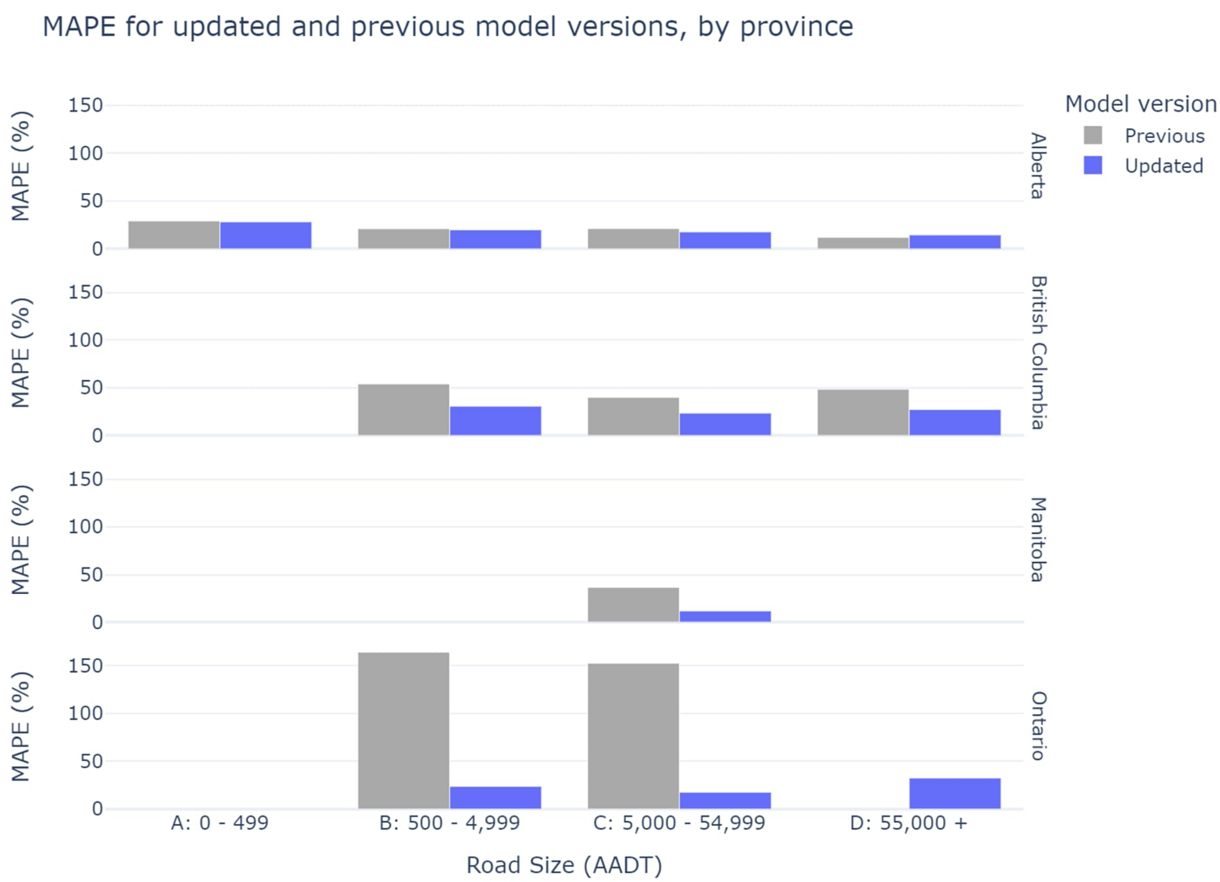

Figure 9: MAPE for the updated model version, compared with the previous model version, by road size and province. Previous model values include errors for counters used in the previous model, as well as new counters.

The comparisons between the updated and previous model versions highlight substantial improvements gained by expanding counter data coverage. In particular, the new models appear to more closely represent road volumes in eastern provinces and densely populated areas. The above comparisons are also reminders that cross-validation results alone do not provide visibility into accuracy in places without counter data. Since data availability is limited in much of the country, our Volume models may still have weaknesses in these locations.

Model Accuracy and Comparability Across Time

Since the monthly models span January 2019 to April 2022, it's important to check that error rates are relatively consistent across time. Figure 10 shows MAPE by month for different road sizes. Road sizes A and D, the largest and smallest roads, have a fair amount of variability, which is expected given that relatively few counters are on these roads. Road sizes B and C, which represent most roads and most counter locations, have stable errors across 2019 and 2020, with lower errors beginning in mid-2021. The 2021 drop is likely due to fewer counters in those months. Most of our new counters have data available for January 2019 through June 2021, but fewer report volumes for July 2021 onwards. As a result, even though the model's validation errors are low for July 2021-April 2022, it may be less accurate at locations where validation data was not available in those months.

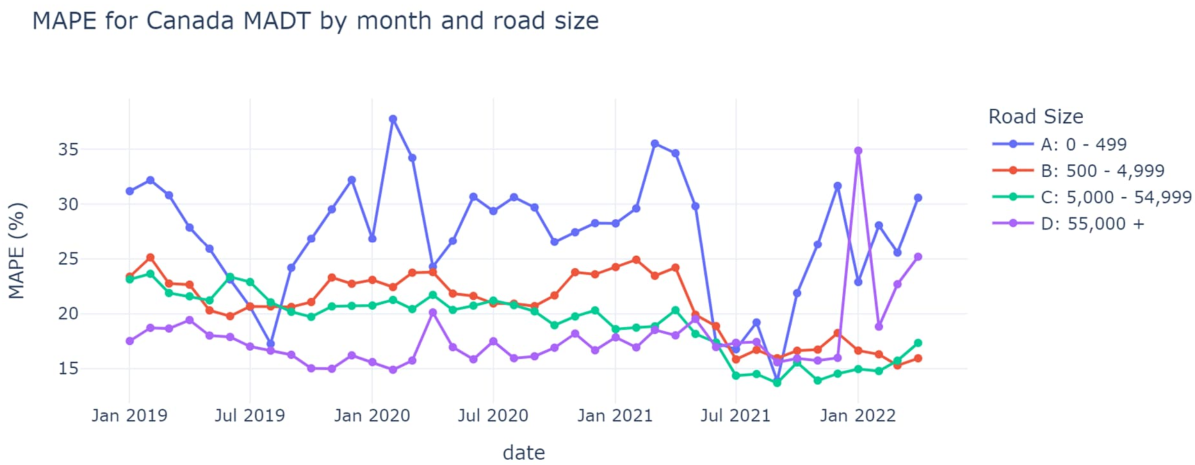

Figure 10: MAPE by month for StreetLight estimated MADT, from January 2019–April 2022. Road sizes A and D have relatively few data points; most roads are in categories B and C.

The models also appear to capture time trends well. Figure 11 depicts estimated Volume compared with counter MADT by month for a selection of permanent counter locations. The locations are randomly selected so that all provinces and road sizes are represented. In all cases, the estimated Volumes match the counter values relatively closely. The results reflect trends like the dip during the COVID-19 pandemic and seasonal peaks in the summer.

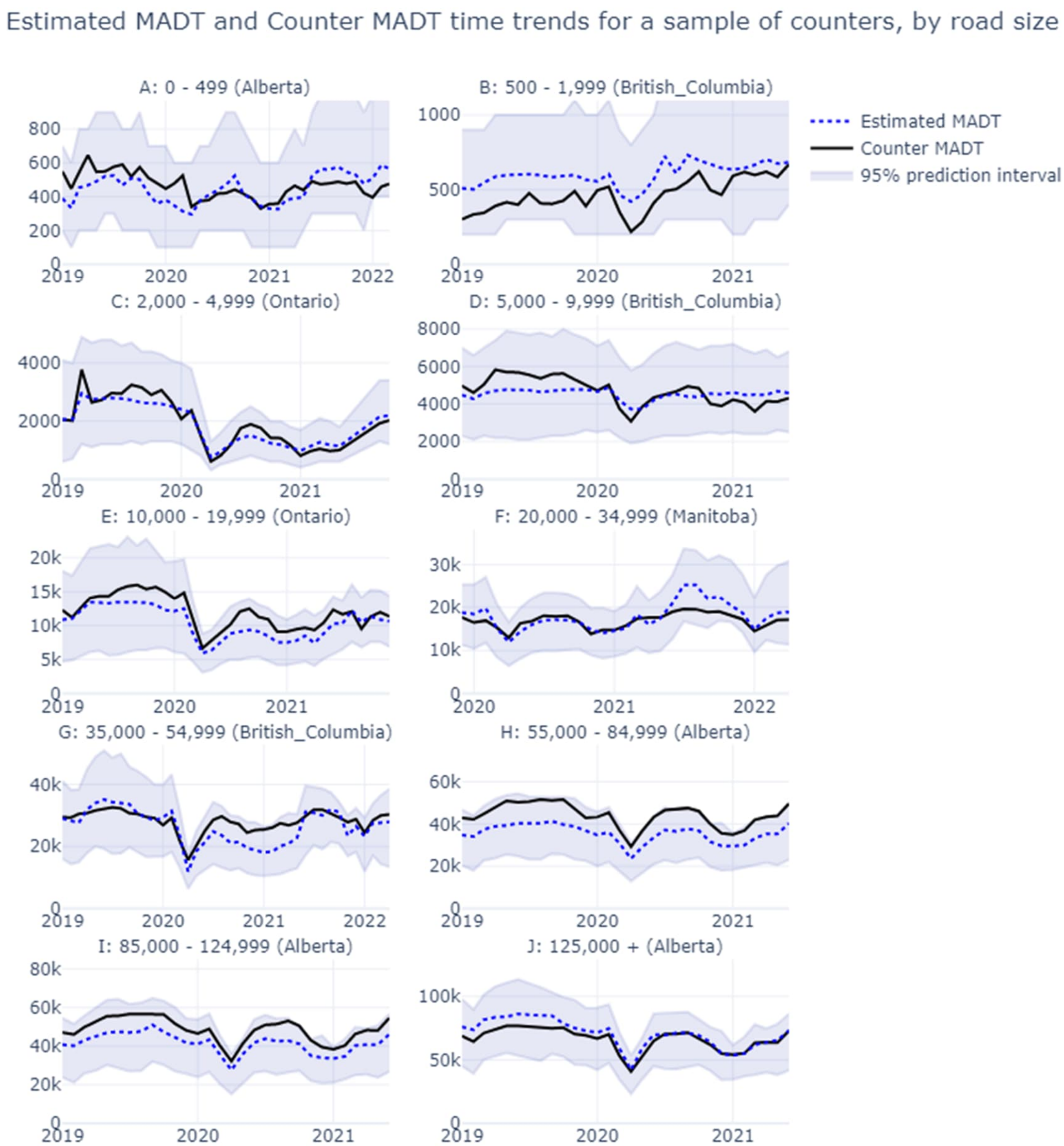

Figure 11: Estimated MADT compared with counter MADT by month for a selection of counter locations. The shaded area represents the 95% prediction interval. Note that not all counters have the full 40 months of data.

Limitations and Our Next Steps

At StreetLight, we are continually updating our data sources and algorithms to enhance the accuracy and coverage of our Volume Metrics. The most recent version of Canada All Vehicles Volume performs well in a variety of contexts. Errors have been reduced in locations where we've obtained new counter data; specifically, urban Vancouver, Winnipeg and Toronto. These reductions in error can be expected to generalize to other urban cores, although it is not possible to quantify the amount of expected improvement. Where permanent counter data is still limited, such as rural and suburban contexts in eastern provinces, accuracy is less certain. The models also perform less well in late 2021 and 2022 in locations with less training data for those months, namely British Columbia, Alberta, and Ontario. Even in these cases, our Volume Metrics will still often provide better insight into travel activity than alternative methods available. We plan to update and improve our models as we expand our sources of trip and validation data.

In order to continue to improve our All Vehicles Volume Metrics in Canada, it’s critical to assemble comprehensive and representative ground-truth data. If you have access to count data that you’d like to be included in future iterations of this work, please contact your StreetLight representative.

About StreetLight

StreetLight Data, Inc. (“StreetLight”) pioneered the use of Big Data analytics to shed light on how people, goods, and services move, empowering smarter, data-driven transportation decisions. StreetLight's proprietary data processing engine, Route Science® algorithmically transforms its vast data resources to measure travel patterns of vehicles, bicycles, and pedestrians, accessible as analytics on the StreetLight InSight® SaaS platform. Acquired by Jacobs as a subsidiary in February 2022, StreetLight provides innovative digital solutions to help communities reduce congestion, improve safe and equitable transportation, and maximize the positive impact of infrastructure investment.

STREETLIGHT DATA