COUNT MULTIVARIATE METRICS: ESTIMATE MODE COUNT AND DISTANCE UNCERTAINTY FROM PHONE SENSORS

For Submission to Travel Survey Methods AEP25

Hui Xian Grace Lim
University of Central Florida
hxgrace@knights.ucf.edu
ORCID: 0000-0002-3720-8281

Michael Allen
National Renewable Energy Laboratory
michael.allen@nrel.gov

K. Shankari, Ph.D., Corresponding Author
National Renewable Energy Laboratory
k.shankari@nrel.gov
ORCID: 0000-0002-7046-5570

Word Count: 6060 words + 5 table(s) $\times$ 250 = 7310 words

Submission Date: August 11, 2023

ABSTRACT

To be informative, travel behavior metrics need to have not only measured values, but also the uncertainty of these values. Smartphone-based travel studies can use phone sensors and manual user labeling to collect trip data, but both of these are susceptible to error.

This work investigates a method for estimating two metrics– the count of trips taken per transportation mode and the distance travelled per mode– and uses variance as a range of uncertainty for those estimates, with the goal that the uncertainty range should capture the actual value from the estimated values. Given a set of phone-based mode predictions and a column-normalized confusion matrix probabilities from an existing mode inference model, we use the closed form solutions for mean and variance of a multinomial distribution to determine the estimated values per mode and their corresponding variances.

We tested this method on three different real world datasets which used phone-sensed data to predict trip modes and prompted users label trip modes, and found that our method works when probability distribution of predicted trips in the computation dataset and the evaluation dataset are very similar. If these distributions are too dissimilar, the range of uncertainty fails to capture actual values. Future work could involve using prior mode distributions to adjust these probabilities to be more similar, or applying a similar method to other metrics. This implies that, in the presence of a similar, labeled dataset, automated sensing outputs can be fully characterized, and can be used in complex decision making.

Keywords: Statistics, Uncertainty, Smartphone App, Travel Survey Methods AEP25

INTRODUCTION

Travel behavior data can inform decision making on many levels, from personal lifestyle choices to large-scale infrastructure and transportation planning. In the past, travel behavior has been collected through telephone surveys (*1*), vehicle GPS surveys (*2*), and web-based surveys (*3*).

Because of their location tracking capabilities, smartphones can create travel diaries. Travel diaries document travel behavior as a set of trips between locations. Smartphones, however, are unable to directly collect information such as the modes of transportation used on a particular trip. As a result, many travel diary studies rely on participants to label trip modes, which is prone to incorrect or completely missing mode labels. To make up for this, travel modes can be inferred using smartphone-based data, but these inferred modes carry some associated uncertainty.

de Jong et al. (*4*) describe the risks that come with overlooking uncertainty margins around a prediction. A decision could be predicted to be very successful, but that prediction could have a large amount of uncertainty, making the decision based on it more risky. Both a range of possible outcomes and the probabilities for those outcomes are needed to properly inform transportation infrastructure projects. The range of outcomes should capture the true value.

This work proposes a process for estimating mode counts– the number of trips taken in each mode of transportation– and the distance travelled per mode, from a set of phone-sensed predictions, as well as the associated uncertainty of each measure, such that the actual values per mode fall within the range of uncertainty from the expected values. This process uses multinomial distributions to represent the spread of ground truth modes for each sensed mode. The closed form solution for the mean of a multinomial is used to find the expected values, and the closed form solution for the variance is used to find the uncertain range.

We validate that multinomial distributions can indeed be used to characterize trip mode data, and discuss the effectiveness of our process, including the conditions required to get accurate resulting estimates and uncertainties. We then apply our methods to three different real-world datasets. We used open-source travel diary creation algorithms from the OpenPATH platform to generate these datasets, but our method works with any set of algorithms, regardless of classification methods.

The main contributions of this paper are:

1. developing a novel method to estimate the uncertainty of mode and distance counts given the output of a travel diary creation algorithm and a confusion matrix representing the quality of the algorithm,
2. demonstrating that these can be computed efficiently using closed form solutions for multinomial distributions, and
3. evaluating on three different real world datasets.

The rest of the paper is structured as follows: the next section reviews related work (Section 3), and the following section outlines and validates our proposed method(Section 4). Then we describe our process for applying our methods to three datasets and discuss the results (Section 5). Finally, we present our conclusions and suggest future work (Section 6).

RELATED WORK

Travel diary platforms

There is well-established literature and several implementations for travel diary data collection using smartphones.

However, we are not aware of prior work that has considered all the components required for fully characterizing travel behavior metrics in complex situations: (i) a data collection platform, (ii) automated mode inference, (iii) rigorous evaluation, and (iv) incorporating the results of that evaluation into the metrics calculations.

There has been substantial work done in automated mode detection of uni-modal trip sections using machine learning. For example, researchers have used decision trees (*5*) (*6*), Hidden Markov Models (*7*) (*8*), and neural networks (*9*) (*10*) for uni-modal mode detection, but most of them have not been incorporated into full-scale platforms.

Data collection platforms such as the open-source Itinerum (*11*), ohmage (*12*), the Statistics Netherlands TABI application (*13*), automated mode inference and rely on users to provide trip mode information. Proprietary platforms such as FMS (*14*) and rMove (*15*) are less well-documented, so we have not been able to find evidence of automated mode inference, or error characteristics of post-processing algorithms in general.

Some open-source platforms that have implemented automated mode detection have evaluated their algorithms rigorously (e.g. MEILI (*16*)) but focused on their accuracy and F-score, and have also focused on data collection without computing downstream metrics.

Finally, projects that have computed metrics from automated mode inferences have not characterised their accuracy (e.g. MatkaHupi (*17*), Peacox (*18*)) or have been rigorously evaluated but do not consider the implications to downstream metrics (e.g. MotionTag (*19*)).

Statistics and uncertainty

Work has been done to understand uncertainty and uncertainty propagation in travel demand models (*20*) which found that uncertainty compounds over the stages of a travel demand model, and that in general, the predictions made by travel demand models are highly uncertain. Lemp et. al (*21*) found that predictions made by traffic forecasts are also highly uncertain and stresses that uncertainty in transportation-related metrics be considered by transportation planners and policy makers.

Most measures of uncertainty for confusion matrices are based on metrics such as accuracy, precision, recall, and F-score (*22*), and multi-class evaluation metrics include balanced accuracy, Matthew's correlation coefficient, and Cohen's Kappa, described by Grandini et al. (*23*). Darling (*24*) discusses the sources of uncertainty in classifiers and their estimates of the probability that a sample belongs in a certain class.

Braga-Neto (*25*) describes how to quantify error for classifiers. There, the expected error is the probability of the predicted class and ground truth class being different. This work uses probabilities from column-normalized confusion matrices, meaning that ground truth classes were conditioned on predicted classes.

Caelen (*26*), Tötsch and Hoffman (*27*), and Barranco-Chamorro (*28*) take a Bayesian approach to the confusion matrix, and use Dirichlet or Beta distributions on each cell of a confusion matrix, allowing for uncertainty in the confusion matrix itself. In particular, Caelen demonstrates how multinomial distributions can be used to describe confusion matrices. We take that concept and apply it in this work.

Previous work by Allen and Shankari (*29*) characterized the uncertainty of smartphone-sensed, travel-based metrics. There, the uncertainties from mode and trip length were used to characterize the uncertainty of energy consumption estimates. Research by Kosmacher and Shankari (*30*) also evaluated the accuracy of trip length computations and mode inference.

METHODS

The goal of this work is to provide expected values for a particular measure, and a range of uncertainty such that the actual values for that measure typically fall within the range of uncertainty from the expected value.

We start by showing that the problem is an instance of the multinomial distribution and illustrate how the related closed form solutions can be used to compute the values. We then verify that the spread of predicted values for each mode can be treated as a multinomial through computation of the probabilities of certain outcomes. Finally, we determine that using variance, rather than standard deviation, as our uncertain range captures the actual measures more reliably.

Using multinomial distributions to estimate mode-specific uncertainty

Intuitively, to compute the true count for a mode m given a set of predicted modes and the probability of misprediction, we need to sum the estimated count from: (i) trips with predicted mode m that were predicted correctly, and (ii) the count from trips with predicted mode $\neq m$ that were the result of mispredictions (Figure 1) The mispredictions from trips predicted as m will contribute to the true counts of other modes.

Considering a set of trips with a predicted mode m, if we conceptually think of "flipping them over" to find the true mode, the result should correspond to the column corresponding to m in the confusion matrix ($P_m(true|pred)$). Thus, the problem of determining true counts maps well to the multinomial distribution.

Concretely, we meet all the criteria for a multinomial distribution:

1. k discrete categories - these are the set of potential modes,
2. fixed probability for each category - this is the $P_m(true|pred)$ from the confusion matrix column for m, and
3. independent trials - mode inference algorithms typically work at the trip level, so the probability of a trip using mode m doesn't depend on any other trips.

Once we have computed the set of true counts for each predicted mode, we can sum them mode-wise to obtain the final true counts (Figure 1).

Formally, for a given confusion matrix, say that θ_p is the vector of column-normalized probabilities for predicted mode p. θ_{tp} then is $P(actual = t|predicted = p)$. N_{tp} is the value of the cell where actual mode = t and predicted mode = p, and n_p is the predicted value for mode p. The closed form solution for mean and variance can then be used to say that for each N_{tp}:

$$E[N_{tp}] = n_i \theta_{tp} \tag{1}$$

and

$$Var[N_{tp}] = n_i \theta_{tp}(1-\theta_{tp}) \tag{2}$$

Then, the actual value for mode j can be estimated by adding $E[N_{tp}]$ for all predicted modes p, and the variance is the sum of $Var[N_{tp}]$ for all predicted modes p:

$$E[N_t] = \sum_{p=1}^{m} E[N_{tp}] = \sum_{p=1}^{m} n_p \theta_{tp} = \sum_{p=1}^{m} n_p P(actual = t|predicted = p) \tag{3}$$

$$Var[N_t] = \sum_{i=1}^{m} Var[N_t p] = \sum_{p=1}^{m} n_p \theta_{tp}(1-\theta_{tp}) \tag{4}$$

$$= n_p P(actual = t|predict = p)(1 - P(actual = t|predict = p)$$

Given the high quality confusion matrix probabilities shown in Table 2 and the correspond-

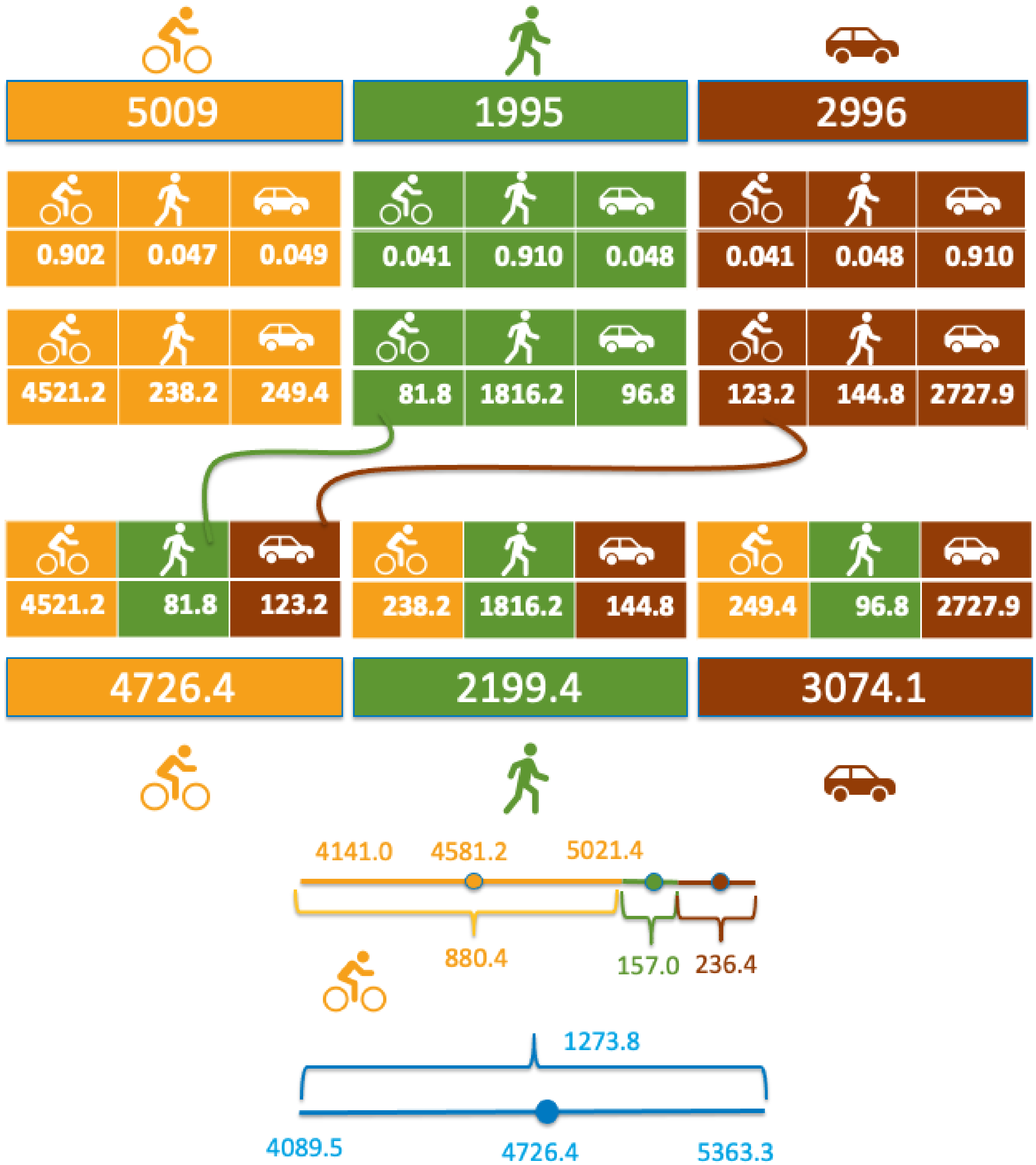

FIGURE 1 Worked example, using mode counts, of the method for computing the expected value (top) and variance (bottom) given: (i) predicted mode counts and (ii) $P(true|pred)$ for each mode. For each mode, we use the probabilities to determine the distribution of true counts given predicted counts. Then we add up the true counts from each predicted mode to get the total true count. For variances, we use a similar method in which we sum up the variances from each predicted mode.

ing predictions shown in Table 3, the expected count per mode and variance per mode is found per the steps in Figure 1.

1. We determine θ_p for each mode
2. We find $E[N_{tp}]$ and $Var[N_{tp}]$ for all predicted modes p using Equation 1 and Equation 2
3. We re-group $E[N_{tp}]$ and $Var[N_{tp}]$ by the true mode t.
4. We find $E[N_t]$ and $Var[N_t]$ by summing all $E[N_{tp}]$ for a true mode t.
5. **Example:** $E[N_{bike}] = 4521.2 + 81.8 + 123.2 = 4726.443$.
6. **Example:** $Var[N_{bike}] = 440.2 + 78.5 + 118.2 = 636.9$

Since multinomial distributions work for discrete values, we converted our distance data from meters to the nearest kilometer, so that the inputs are integers usable with multinomial distributions. Then the same process as outlined in Section 4.1 is used.

Verifying multinomial characteristics of data

To justify using the closed form solutions for the mean and variance of a multinomial distribution, we verify that our data can be represented by multinomial distributions.

Given that five trips were predicted as bike, three trips were predicted as walk, and two trips were predicted as car, and using an arbitrary count-based, column-normalized confusion matrix, we manually calculated the probability of three cases:

1. There were no trips with a ground truth mode of bike
2. There was one trip with a ground truth mode of bike
3. There were ten trips with a ground truth mode of bike

Using the same set of probabilities and predictions, we then found the probability of each of the three cases occurring using multinomial distributions. One multinomial distribution was made for each predicted mode, according to the corresponding column of probabilities. Using the probability mass functions of the respective multinomial distributions, we found the probabilities for all possible instances that would fulfill each case requirement. We then added these probabilities up for each case.

The probabilities calculated manually and the probabilities calculated using multinomial distributions matched, indicating that our data could indeed be represented by multinomial distributions.

Effects of probability on using standard deviation as an uncertain range

Say that we have two multinomials, each with a different probability distribution, which we refer to as PD_1 and PD_2. In the example below, probability distributions are a set of three probabilities that describe the likelihood of a trial falling into each of three categories, which we generally refer to as category A, category B, and category C.

Even if PD_2 is fairly close to PD_1, the difference between the expected values generated from PD_1 and PD_2 falls outside the range of one standard deviation of the data generated from PD_1 if PD_1 and PD_2 are not extremely similar. Table 1 shows an example of the difference in expected values for category A, $Diff(A)$, and the corresponding SD_1 for category A as the difference between PD_1 and PD_2 increase. In this case, 10000 random variables were sampled from two multinomial distributions characterized by PD_1 and PD_2, with 1000 trials.

PD_1 is kept constant, at [0.8, 0.05, 0.15]. PD_2 is varied from PD_1 by a difference of 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, and 0.4. The difference between expected values begins to fall outside of one standard deviation of PD_1 at a difference between probabilities of 0.025. It begins to fall outside of the range of one variance at a difference between probabilities of 0.2.

PD_2	$Diff(A)$	$SD_1(A)$	$Var_1(A)$
[0.8, 0.05, 0.15]	0.0352	12.790718	163.602475
[0.79, 0.05, 0.16]	9.9293	12.607993	158.961494
[0.775, 0.05, 0.175]	25.1165	12.702414	161.351324
[0.75, 0.05, 0.2]	49.6176	12.593497	158.596171
[0.725, 0.05, 0.225]	75.0212	12.742726	162.377065
[0.7, 0.05, 0.25]	99.7538	12.721244	161.830042
[0.6, 0.05, 0.35]	199.9392	12.590705	158.525852
[0.5, 0.05, 0.45]	300.0050	12.658316	160.232969
[0.4, 0.05, 0.55]	399.6070	12.636061	159.670036

TABLE 1 Difference between expected values of multinomials for the specified and their corresponding standard deviations and variances probabilities

This means that if we have two datasets that are extremely similar, or even if we are using two halves of the same dataset, it is unlikely that the actual value is captured within one standard deviation from the expected value. Additionally, if probability distributions vary too greatly, the variance will also fail to capture the actual value.

Variance as a more appropriate measure of uncertainty

Since one standard deviation fails to capture the difference in expected means even when probability distributions are extremely similar, in an ideal case the actual measures still fail to be captured within the range of one standard deviation. To demonstrate this, we use a confusion matrix for a high quality classifier (Table 2) and mode count predictions based on the predicted mode distribution from that high quality confusion matrix (Table 3). Evidently, even in the most optimal scenario, it is possible to have expected values that are over 3 standard deviations from the actual values, even though the percent relative errors between the two never exceed 3% (top of Table 4).

	Bike	Walk	Car
Bike	0.902632	0.041041	0.041150
Walk	0.047571	0.910410	0.048334
Car	0.049798	0.048549	0.910516

	Bike	Walk	Car
Bike	0.800908	0.108108	0.107698
Walk	0.052092	0.206880	0.748015
Car	0.147001	0.685012	0.144287

TABLE 2 High quality confusion matrix probabilities (left) and lower quality confusion matrix probabilities (right) where rows represent ground truths and columns represent predictions

	Bike	Walk	Car
Predicted count	5009	1995	2996

	Bike	Walk	Car
Predicted count	4982	2018	3000

TABLE 3 Predicted counts used with high quality confusion matrix (left) and with lower quality confusion matrix (right)

		Actual Value	Expected Value	SD from actual	Difference	Relative pct error	Variance
High quality CM	Bike	4762	4726.443	-1.408	-35.557	-0.746	636.958
	Walk	2243	2199.361	-1.900	-43.639	-1.945	527.476
	Car	2995	3074.196	3.307	79.196	2.644	573.271
Lower quality CM	Bike	4733	4531.376	-5.641	-201.623	-4.260	1277.276
	Walk	2179	2921.048	21.952	742.049	34.055	1142.581
	Car	3088	2547.574	-14.288	-540.426	-17.501	1430.530

TABLE 4 Results from using the high quality confusion matrix and predictions (top) and lower quality confusion matrix and predictions (bottom)

In the case where a lower quality confusion matrix is used and the predictions are based on a different distribution than specified by the lower quality confusion matrix, the range of multiple standard deviations is inadequate to capture the actual measure. If we vary the probabilities used to find expected values, the actual values can be over even twenty standard deviations away from the expected values. Using a range of multiple standard deviations also fails to be effective, as it changes case-by-case. For example, after looking at the results using the high quality confusion matrix (top of Table 4), the range might be set at four or five standard deviations. However, looking at the results with the lower quality confusion matrix (bottom of Table 4), it appears that a more appropriate range is somewhere over twenty one standard deviations. Variance as our uncertain range, on the other hand, captures the actual values more reliably, even though the relative percent error is much higher than when using the high quality confusion matrix.

EVALUATION AND RESULTS

Data used

Three different datasets are used in this work, referred to as `MobilityNet`, `All_CEO`, and `Durham`.

1. `MobilityNet` is a publicly available dataset of artificial trips that were created by traversing predefined travel routes using predetermined modes of travel (*31*). `MobilityNet` has distance data available that we use to make distance estimates per mode, but not enough trips to use for mode counts.
2. `All_CEO` is a dataset of trips that covers 1.5 years of the Colorado Energy Office's CanBikeCO programs in six locations across Colorado, which provided e-bikes to low income participants (*32*), and
3. `Durham` is a set of trips from a similar program from `Durham`, North Carolina.

Since most current travel diary creation are closed source, we used the open source travel diary creation algorithms from the OpenPATH platform to predict the sensed modes in these datasets. The algorithms were developed before `MobilityNet`, `All_CEO`, or `Durham` datasets were generated using an independent dataset. Note, however, the methods outlined here are independent of the travel diary creation algorithm used.

Assumptions

To apply our methods to a particular dataset, we:

- process our data so that user-provided ground truth modes map to sensed modes (Table 5). Ground truth modes that did not map to sensed modes– such as Scooter share and

Skateboard– were removed.

- assume that mode predictions are independent of each other, which allows us to combine variances per mode by addition.
- assume entries of the column-normalized confusion matrix represent $P(actual|predicted)$

sensed mode	All_CEO and Durham	MobilityNet
bicycling	Bikeshare, Pilot ebike, Regular Bike	BICYCLING, E_BIKE
bus	Bus, Free Shuttle	BUS
car	Gas Car, drove alone, Gas Car, with others, Taxi/Uber/Lyft, E-car, drove alone, E-car, with others	CAR
train	Train	TRAIN, LIGHT_RAIL
walking	Walk	WALKING

TABLE 5 Mode mappings to sensed modes for All_CEO, Durham, and MobilityNet.

Metrics and goals

Now we want to validate whether our method (Section 4.1) provides a range of uncertainty that capture actual values for these real-world datasets. We use the variance (Section 4.4) as our bounds. This means that for each mode, the actual value should be within one variance of the expected value.

Evaluation procedure

The ground truth mode for a trip is the primary mode of a trip; that is, the user-provided mode used to travel the farthest during that trip. Before count or distance confusion matrices are constructed for any dataset, we determine the primary mode for each trip and use it as the ground truth mode. We refer to the dataset used to create confusion matrices for our calculations as the computation dataset. The dataset that we obtain our mode predictions from is referred to as the evaluation dataset.

We use three different methods– evaluating against MobilityNet, self-validation, and cross validation (Figure 2). When evaluating against MobilityNet, we use MobilityNet's artificial trips as our computation dataset and a dataset that uses real trips, such as All_CEO or Durham, as our evaluation dataset. When doing self-validation, we shuffle and split one dataset of real trips into two halves, using one as our computation dataset and the other as our evaulation dataset. Finally, when doing cross-validation, we use two different datasets with real trip data, such as All_CEO or Durham, with one as the computation dataset and the other as the evaluation dataset.

For readability, plots are split between modes with high counts or distances and modes with low counts or distances, and scaled appropriately. This is because some modes were used more frequently than others; for example, there were 40606 ground truth car trips in the All_CEO dataset, but only 169 ground truth train trips.

Results and Discussion

We started with the MobilityNet distance confusion matrix since it was created from artificial trips and was not dependent on the travel patterns in any particular region (Figure 3). However, the actual and expected values differ significantly in each of the All_CEO and Durham datasets. More seriously, given our metrics, none of the actual values are captured by the uncertainty ranges.

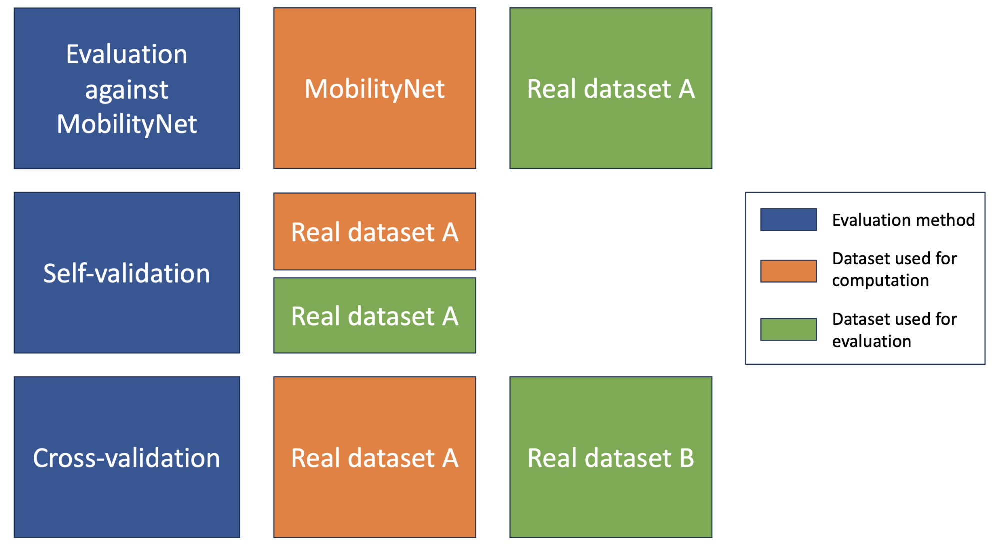

FIGURE 2 The process used for each method.

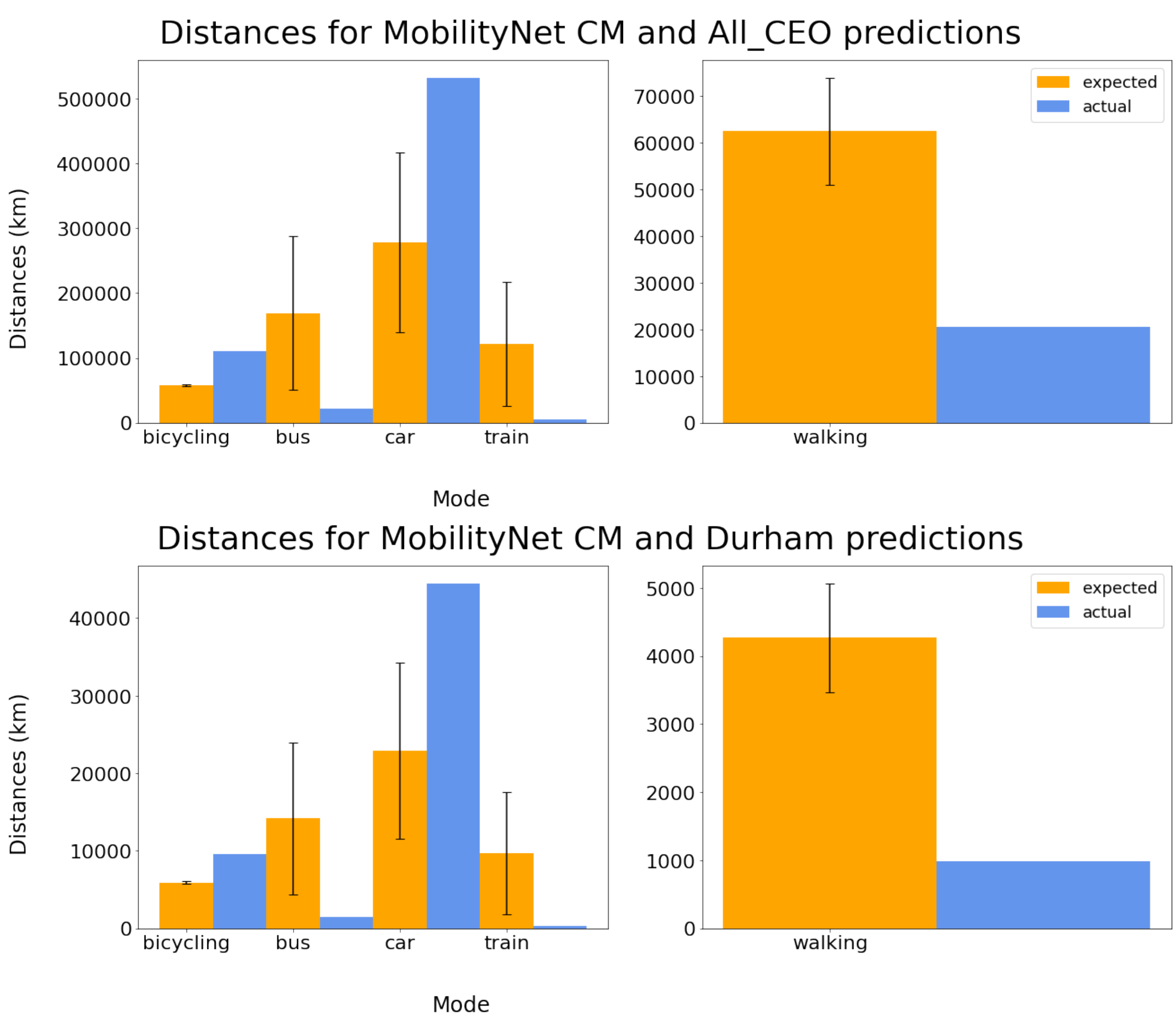

FIGURE 3 Results of estimating distances when using a confusion matrix from `MobilityNet` **and predictions from** `All_CEO` **(top) and** `Durham` **(bottom)**

`All_CEO` and `Durham` have more similar confusion matrix probabilities for distances, while `MobilityNet` has a markedly different distribution of probabilities for each sensed mode. For example, for the probabilities in the predicted walk column (Figure 4), the largest probabilities for `All_CEO` and `Durham` are for ground truth modes bicycling and car. For the predicted walk column of `MobilityNet`, bicycling and car ground truth modes have the lowest probabilities. Three out of the five `MobilityNet` predicted walk probabilities also exceed the corresponding `All_CEO` and `Durham` probabilities by 0.2 or more. As a result, variance calculated from `MobilityNet` does not capture actual values (Section 4.3).

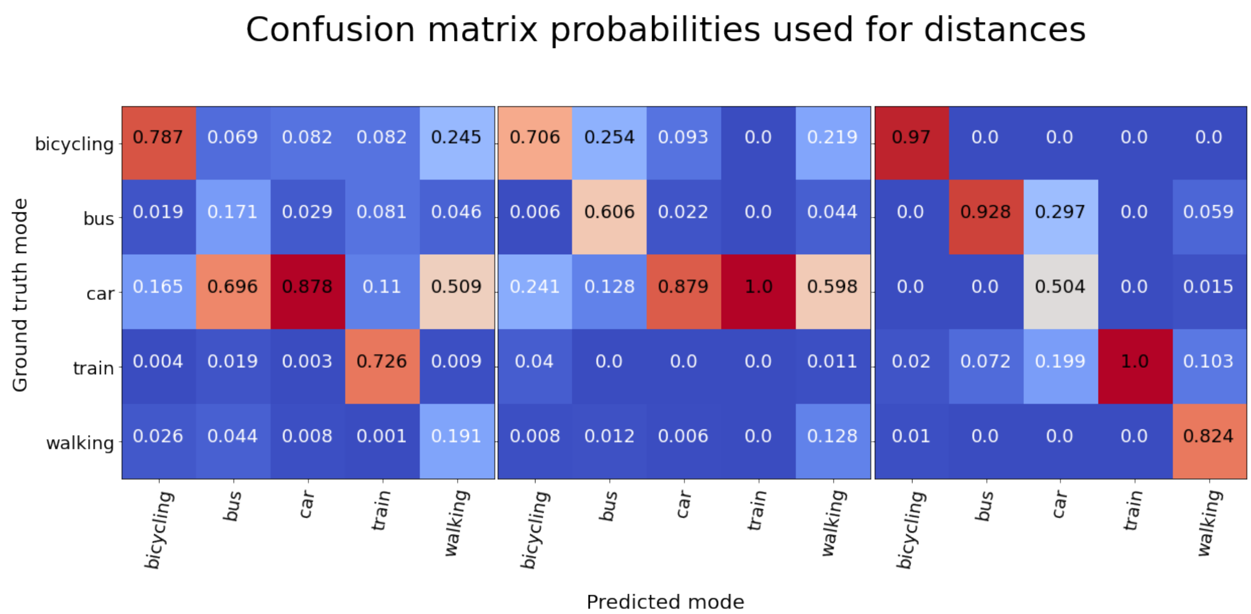

FIGURE 4 Confusion matrix probabilities used for distances for `All_CEO` (left), `Durham` (middle), and `MobilityNet` (right)

Given that the `MobilityNet` results did not match our expectations, we experimented with using labeled, real-world data collection instead. These real-world datasets do not have high quality ground truth, and do not account for segmentation error, but they are larger and may provide more realistic error estimates.

We experimented with two options, self-validation through splitting a single dataset into two parts, and cross-validation through using datasets from two different geographic locations (Figure 2). Self-validation gives us a reference for results in an ideal case, since the confusion matrix probabilities used are extremely close to the probabilities that describe the data used for predictions. Cross-validation allows us to see how our method performs in a realistic scenario, where the confusion matrix is from a different program than the predictions.

These results (Figure 5) are much better, since the actual distances per mode all fall within the range of uncertainty from the expected distances. The proportion and size of the uncertain ranges per mode are also fairly consistent across results.

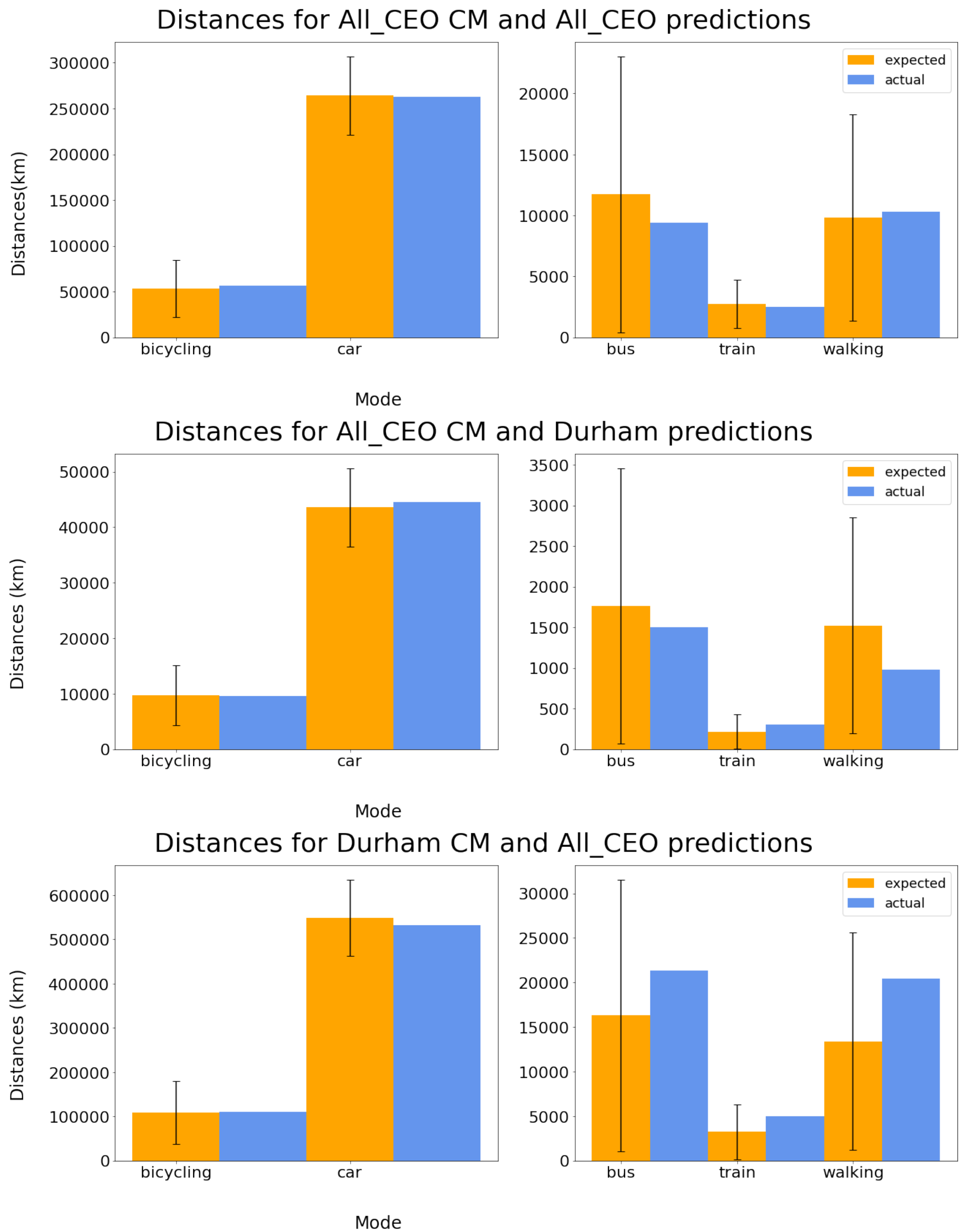

FIGURE 5 Results of estimating distances with confusion matrices from user-labeled trips: `All_CEO` **confusion matrix to** `All_CEO` **predictions (top),** `All_CEO` **confusion matrix to** `Durham` **predictions (middle),** `Durham` **confusion matrix to** `All_CEO` **predictions (bottom)**

Counts

The `MobilityNet` dataset did not have enough trips to be usable for estimating counts. Even if there were enough trips, unlike with distance, no high quality granular ground truth for counts exists from `MobilityNet` because there is no method to generate a count-level confusion matrix that takes segmentation error into account. So, for counts, we only used the real world datasets `Durham` and `All_CEO`.

The results (Figure 6) largely meet our goals, with the actual counts falling within the uncertainty range from the expected counts in almost all cases. The lone exception is the `train` case while using the Durham confusion matrix to predict `All_CEO` trips.

This is because the distribution of probabilities for the ground truth mode train is drastically different between `Durham` and `All_CEO`. In the `Durham` dataset, all 3 ground truth train trips were predicted as bicycling or walking, while in the `All_CEO` dataset, ground truth train trips were predicted as each of the five sensed modes at least once. This means that the variance for the mode train should be much higher in order to properly represent `All_CEO` for the ground truth mode train, hence why `Durham` fails to capture the actual counts within the calculated uncertain range.

Since using `Durham` as our computation dataset and `All_CEO` as our evaluation dataset for distances met our goals, it may seem unusual that doing the same for counts did not. However, the distance travelled by bicycling is more than the distance travelled by walking, which means that even though the number of trips for each mode doesn't change, bicycling contributes more to the variance for distance calculations than in the count calculations.

The calculations for distance and variance are outlined below. For berevity, $Var(actual = i|predicted = j)$ is abbreviated as $Var(i|j)$ and $P(actual = i|predicted = j)$ is abbreviated as $P(i|j)$ for some modes i and j.

For counts, the variance calculation is:

$$\begin{aligned}
&Var(train|bicycling) + Var(train|walking) \\
&= n_{bicycling} \times P(train|bicycling) \times (1 - P(train|bicycling)) \\
&+ n_{walking} \times P(train|walking) \times (1 - P(train|walking)) \\
&= 11388.0 \times 0.000840 \times (1 - 0.000840) + 26626.0 \times 0.001120 \times (1 - 0.001120) \\
&= 9.561706 + 29.782960 \\
&= 39.344667
\end{aligned}$$

For distances, the variance calculation is:

$$\begin{aligned}
&Var(train|bicycling) + Var(train|walking) \\
&= n_{bicycling} \times P(train|bicycling) \times (1 - P(train|bicycling)) \\
&+ n_{walking} \times P(train|walking) \times (1 - P(train|walking)) \\
&= 59503 \times 0.039710 \times (1 - 0.039710) * 75133 \times 0.011357 \times (1 - 0.011357) \\
&= 2269.034886 + 843.591857 \\
&= 3112.626742
\end{aligned}$$

Note how the predicted distance for bicycling is around 80% of the predicted distance for walking, even though the predicted count for bicycling is around 40% of the predicted count for walking, and how $P(train|bicycling)$ for distances is magnitudes larger than $P(train|bicycling)$ for counts. This increase in values leads to a higher variance that captures the actual value for distances.

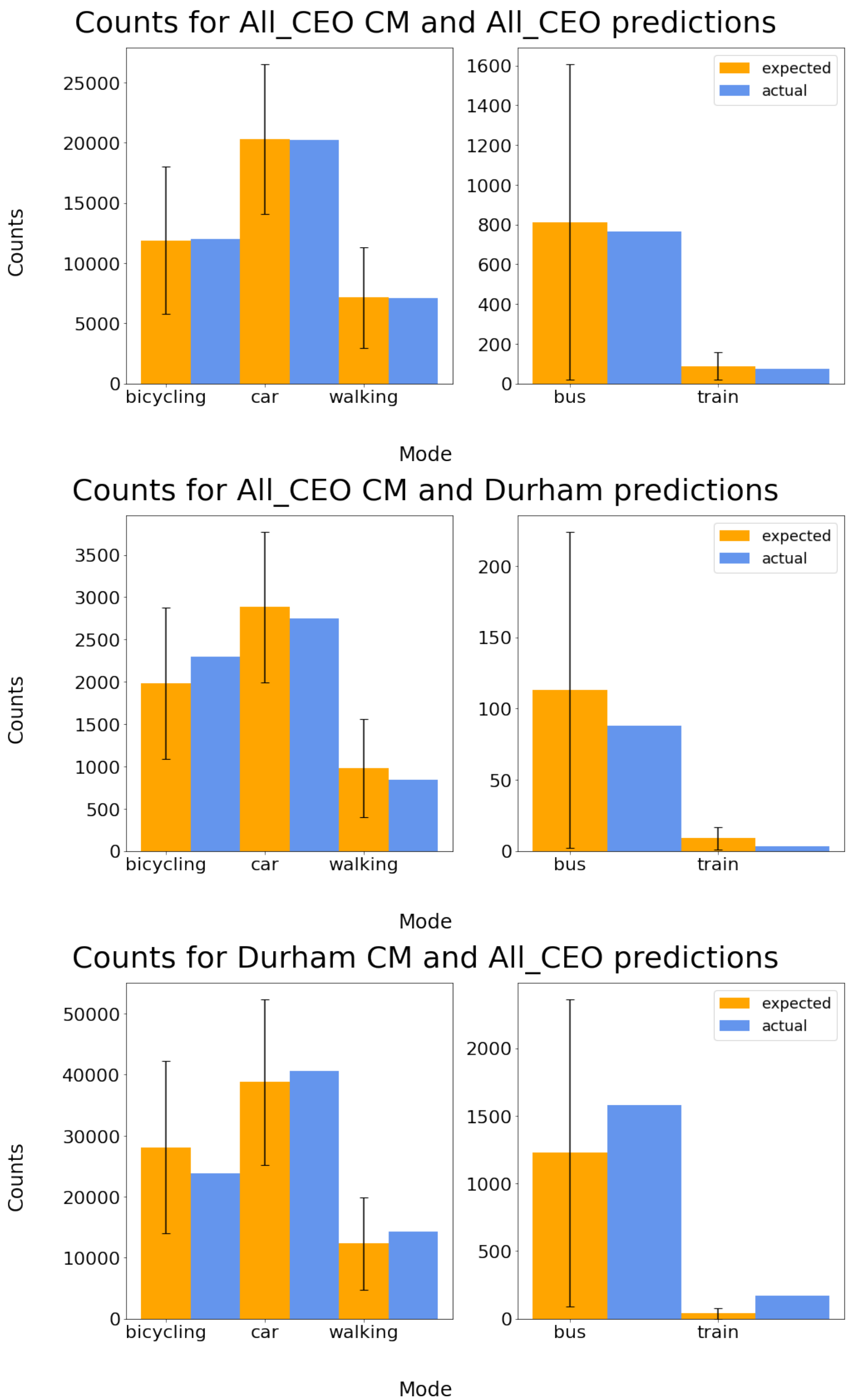

FIGURE 6 Results of estimating counts when using the `All_CEO` confusion matrix on predictions from `All_CEO` (top), while using the `All_CEO` confusion matrix on predictions from `Durham` (middle), and the `Durham` confusion matrix on predictions from `All_CEO`.

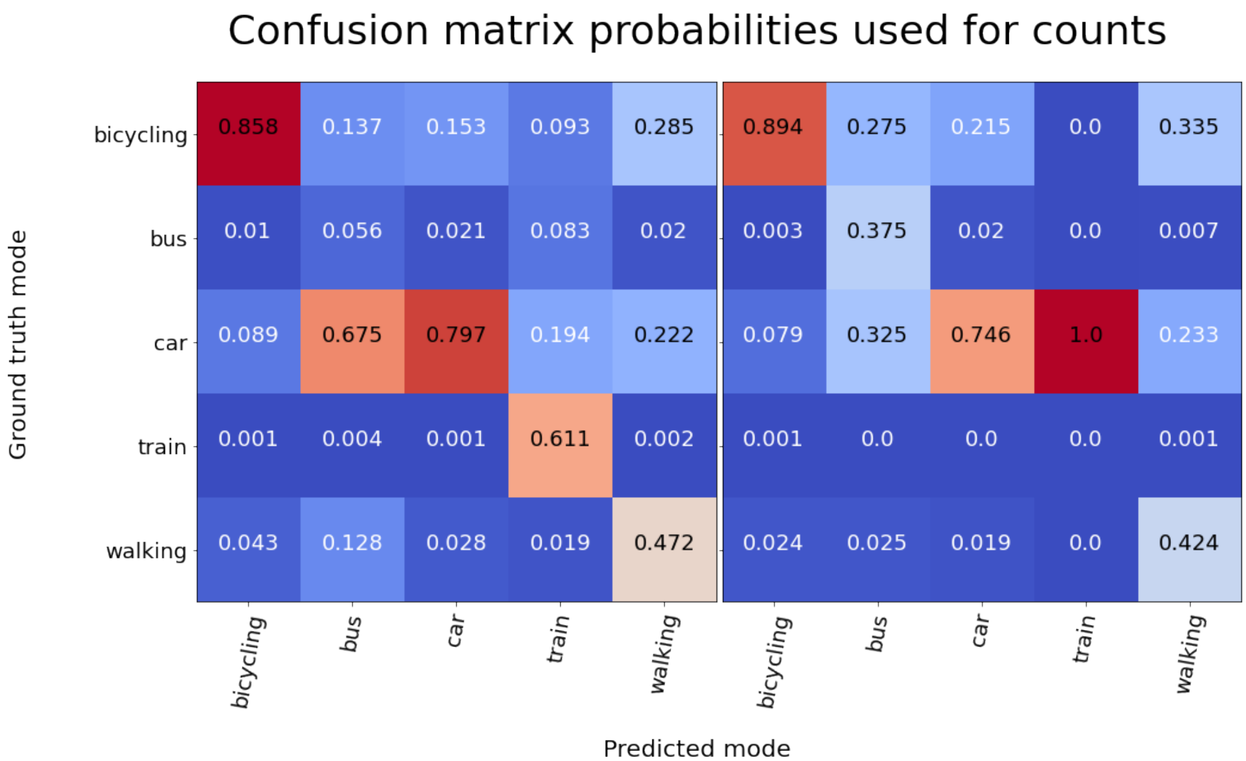

FIGURE 7 Confusion matrix probabilities used for counts for `All_CEO` **(left) and** `Durham` **(right)**

CONCLUSION

The purpose of this work is to provide a measure of uncertainty based on a set of predictions and a confusion matrix from some previous mode inference model, regardless of the actual mode inference model used. Our method uses column-normalized probabilities from a classifier's confusion matrix and multinomial closed form solutions for mean and variance to give an expected value and uncertain range for each sensed mode.

We tested our method for measures of both counts and distances with three different datasets, and found that this method works if the confusion matrix probabilities used are similar to probabilities from the classifier used to find predictions per mode. Disclaimer: Since our model of uncertainty did not capture the true variation, we used the value of the variance calculated from that model to add to our estimated mean and get a wider range of uncertainty. The variance we add to the mean should not be interpreted as the variance to use for calculations derived from this data. This is an area open to improvement; in the future, confusion matrix probabilities could be adjusted using some prior mode distribution to more closely resemble the data that the predictions come from. This method could also be extended to apply to metrics other than mode count and distance travelled per mode.

When fully implemented, phone-based travel surveys can reduce mode labeling requirements to in turn reduce user burden, and mode usage can be calculated over multiple users and over a span of time. The metrics of count and distance can also give insight into the potential impact of large-scale infrastructure and transportation decisions, and inform individuals on their travel behaviors.

ACKNOWLEDGEMENTS

NREL is operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported in part by the DOE Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. This material is also based upon work supported by the DOE Office of Energy Efficiency and Renewable Energy Vehicle Technologies Office, as part of the Energy Efficient Mobility Systems program's Core Tools activity area. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting this work for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

AUTHOR CONTRIBUTION

The authors confirm contribution to the paper as follows: study, conception, and design: K. Shankari, M. Allen, data collection: H. Lim, analysis and interpretation of results: K. Shankari, M. Allen, H. Lim; draft manuscript preparation: H. Lim, K. Shankari. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Santos, A., N. McGuckin, H. Y. Nakamoto, D. Gray, and S. Liss, Summary of Travel Trends: 2009 National Household Travel Survey. *Federal Highway Administration*, 2011.
2. Stopher, P. R., P. Bullock, and F. Horst, Exploring the Use of Passive GPS Devices to Measure Travel, 2002.
3. Zhou, J., Sustainable commute in a car-dominant city: Factors affecting alternative mode choices among university students. *Transportation Research Part A: Policy and Practice*, Vol. 46, No. 7, 2012.
4. de Jong, G., M. P. A. Daly, S. Miller, R. Plasmeijer, and F. Hofman, Uncertainty in traffic forecasts: literature review and new results for The Netherlands. *Transportation*, 2007.
5. Reddy, S., M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, Vol. 6, 2010, p. 1–27.
6. Zheng, Y., Y. Chen, Q. Li, X. Xie, and W.-Y. Ma., Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web*, Vol. 4, 2010.
7. Zhong, M., J. Wen, P. Hu, and J. Indulska, Advancing Android activity recognition service with Markov smoother, 2015, pp. 38–43.
8. Shah, R., C.-y. Wan, H. Lu, and L. Nachman, Classifying the mode of transportation on mobile phones using GIS information, 2014, pp. 225–229.
9. Yang, F., Z. Yao, and P. Jin, GPS and Acceleration Data in Multimode Trip Data Recognition Based on Wavelet Transform Modulus Maximum Algorithm. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2526, 2015, pp. 90–98.
10. Gonzalez, P., J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi, and R. Perez, Automating mode detection for travel behaviour analysis by using global positioning

systemsenabled mobile phones and neural networks. *Intelligent Transport Systems, IET*, Vol. 4, 2010, pp. 37 – 49.

11. Patterson, Z., K. Fitzsimmons, S. Jackson, and T. Mukai, Itinerum: The open smartphone travel survey platform. *SoftwareX*, Vol. 10, 2019, p. 100230.
12. Tangmunarunkit, H., J. Kang, Z. Khalapyan, J. Ooms, R. Nithya, D. Estrin, A. Hsieh, B. Longstaff, S. Nolen, J. Jenkins, C. Ketcham, J. Selsky, F. Alquaddoomi, and D. George, Ohmage: A General and Extensible End-to-End Participatory Sensing Platform. *ACM Transactions on Intelligent Systems and Technology*, Vol. 6, 2015, pp. 1–21.
13. McCool, D., P. Lugtig, O. Mussmann, and B. Schouten, An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. *Journal of Official Statistics*, Vol. 37, 2021, pp. 149–170.
14. Zhao, F., A. Ghorpade, F. Câmara, P. Singapore, C. Zegras, M. Ben-Akiva, and F. Zhao, Stop Detection in Smartphone-based Travel Surveys, 2014, Vol. 11.
15. *rMove*. `https://rmove.rsginc.com`, 2022, accessed: 2023-08-01.
16. Prelipcean, A., G. Gidófalvi, and Y. Susilo, MEILI: A travel diary collection, annotation and automation system. *Computers, Environment and Urban Systems*, Vol. 70, 2018.
17. Jylhä, A., P. Nurmi, M. Sirén, S. Hemminki, and G. Jacucci, MatkaHupi: a persuasive mobile application for sustainable mobility, 2013, pp. 227–230.
18. Schrammel, J., M. Busch, and M. Tscheligi, Peacox - Persuasive Advisor for CO2-Reducing Cross-modal Trip Planning, 2013.
19. Molloy, J., A. Castro, T. Götschi, B. Schoeman, C. Tchervenkov, U. Tomic, B. Hintermann, and K. Axhausen, The MOBIS dataset: a large GPS dataset of mobility behaviour in Switzerland. *Transportation*, 2022, pp. 1–25.
20. Zhao, Y. and K. Kockelman, The Propagation Of Uncertainty Through Travel Demand Models: An Exploratory Analysis. *The Annals of Regional Science*, Vol. 36, 2001.
21. Lemp, J. and K. Kockelman, Understanding and Accommodating Risk and Uncertainty in Toll Road Projects: A Review of the Literature. *Transportation Research Record*, Vol. 2132, 2009.
22. M, H. and S. M.N., A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, Vol. 5, 2015.
23. Grandini, M., E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: an Overview*, 2020.
24. Darling, M. C., Using Uncertainty To Interpret Supervised Machine Learning Predictions, 2019.
25. Braga-Neto, U., *Fundamentals of Pattern Recognition and Machine Learning*. Springer International Publishing, 2020.
26. Caelen, O., A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, Vol. 81, 2017.
27. Tötsch, N. and D. Hoffmann, Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer Science*, Vol. 7, 2021.
28. Barranco-Chamorro, I. and R. Carrillo-García, Techniques to Deal with Off-Diagonal Elements in Confusion Matrices. *Mathematics*, Vol. 9, 2021, p. 3233.
29. Allen, M. and K. Shankari, Estimating Travel Energy Consumption Uncertainty Based on Inferred Travel Mode and Sensed Travel Length. *Transportation Review Board Annual Meeting*, 2023, under Review.

30. Kosmacher, G. and K. Shankari, I Know I'm Right, But Does my Phone? Evaluating the Interplay between Trajectory Segmentation and Mode Inference Errors. *Transportation Review Board Annual Meeting*, 2022, under Review.
31. Shankari, K., J. Fuerst, M. F. Argerich, E. Avramidis, and J. Zhang, MOBILITYNET: TOWARDS A PUBLIC DATASET FOR MULTI- MODAL MOBILITY RESEARCH. *ICLR 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020.
32. Akcicek, C., Z. Aemmer, K. Shankari, and A. Duvall, Freewheeling: What Six Locations, 61,000 Trips, and 242,000 Miles in Colorado Reveal About How E-Bikes Improve Mobility Options, 2023.