USING PASSIVE DATA TO BUILD AN AGILE TOUR-BASED MODEL: A CASE STUDY IN ASHEVILLE

Josephine D. Kressner, PhD (Corresponding Author)
Transport Foundry
3423 Piedmont Rd NE Suite 481
Atlanta, GA 30305

Gregory S. Macfarlane, PhD
WSP | Parsons Brinckerhoff
434 Fayetteville St, Suite 1500
Raleigh, NC 27601

Leta Huntsinger, PhD, PE
WSP | Parsons Brinckerhoff
434 Fayetteville St, Suite 1500
Raleigh, NC 27601

Rick Donnelly, PhD, AICP
WSP | Parsons Brinckerhoff
6100 Uptown Blvd NE Suite 700
Albuquerque, NM 87110

Submission Date: 15 April 2016

ABSTRACT

To date, the primary sources of data for travel modeling have been comprehensive household travel surveys, which are collected at high cost and yet tend to have measurement error and nonresponse bias. There has been interest in using passive data for travel modeling, but existing trip- and tour-based models require a direct link between demographics and trip-making behavior that is uncommon in passive data. This project demonstrates an attempt to overcome these limitations by building an agile tour-based model with passive data using an innovative person-based discrete event simulation framework. The simulation-based model is compared with a modern trip-based model recently developed for the North Carolina Department of Transportation (NCDOT) and the French Broad River Metropolitan Planning Organization (FBRMPO) covering the Asheville region in North Carolina. Demand trip tables from both models are fed into the region's static assignment model and validated against traffic counts. The assignment results are similar in terms of average link error against traffic count data; this is the case even though the simulation-based passive data model was neither calibrated to local conditions nor adjusted with scaling factors or shadow prices. A discussion of costs, development time, complexity, and usability between NCDOT's recent aggregate trip-based model and the new tour-based model constructed from passive data are provided. We believe the new modeling approach developed through this research will be a good fit for small and medium-sized communities, as it permits tour-based models to be developed with substantially less investment in time, money, and data.

OBJECTIVES AND MOTIVATIONS

To date, the primary sources of data for travel modeling have been comprehensive household travel surveys, which are collected at high cost and yet tend to have measurement error and nonresponse bias (e.g., *1*, *2)*. Many have investigated the use of passive location data including GPS, WiFi positioning, Bluetooth, and triangulated signal data as an alternative or supplement to surveys (e.g., *3*, *4*). However, these types of passive location data typically lack information about the people that make each trip when collected without a coordinating traditional survey. This makes these data ill-suited for building current trip- or tour-based models; instead, the data have primarily been used for survey supplementation (e.g., *5*), model validation (e.g., *6, 7*), or for specific project studies (e.g., *8*). For a comprehensive review of emerging data usage in research and practice to date, see Lee, Sener, and Mullins (*9*).

Meanwhile, consumer data firms have been compiling information about individuals and households for decades, typically selling the inexpensive, up-to-date data to commercial marketers. These readily available data contain the majority of household and individual demographic and socioeconomic fields that are used in travel demand modeling, but these data lack trip-making behavior. This weakness, along with some concerns over data integrity, has led modelers to rely instead on relatively limited and outdated U.S. Census data.

This research attempts to overcome the respective limitations of passive location and consumer data by building an agile tour-based model with passive data using an innovative person-based discrete event simulation framework rather than local household travel survey data. This research brief will present the methodology used to build the model and then will present a comparison of validation results between the passive data tour-based model and the aggregate trip-based model recently developed for the North Carolina Department of Transportation (NCDOT) and the French Broad River Metropolitan Planning Organization (FBRMPO) covering the Asheville region in North Carolina. The paper concludes with a discussion of costs, development time, complexity, and usability.

METHODOLOGY

The difference between modern travel modeling approaches and the approach we use in this project can be explained by examining a queue at a bank. Using modern approaches for modeling, one customer out of every 100-200 customers who wait for service would be surveyed. The surveyor would collect data about the survey participants' experiences in the queue, asking questions and measuring experienced wait times along the way by that customer. To create a model of the system, the survey data would be expanded to represent all of the customers using sophisticated data expansion techniques. In travel modeling specifically, the expansion is done by comparing the demographics of the survey respondents to the demographics of the estimated full population.

In the passive data model, a different class of measurements are used. Rather than detailing a small subsample of individual survey participants as they move through the system, the surveyor would passively measure *events* across the whole system. In our bank queue example, events to be measured would include the time between customer arrivals, service times at the teller window, and overall teller details like number of tellers by time of day. These types of measurements are more easily collected passively, and therefore can be collected with much larger sample sizes over longer periods of time. This measurement approach, when used with a discrete event simulation, results in a statistically robust model that is not unduly affected by outliers. At the same time, an analyst studying the simulated system can still examine individual experiences by tracing the activities of simulated individuals.

In the context of simulating person-level tour-based travel, events are measured passively by many different third-party data providers. For example, a GPS data provider can provide statistical characterizations of the time of day that people travel to work from home, summarized by small geographic home zones. Other types of location data can statistically characterize the length of time spent at a workplace before making another trip, summarized by small geographic workplace zones. Using discrete event simulation, data fusion, and other statistical techniques with these statistical characterizations, person-level travel diaries can be synthesized systematically using the same principles one would use in simulating a bank queue.

We implemented this proposed simulation-based model for the Asheville, North Carolina region using passive data, some of which was kindly shared by NCDOT and FBRMPO. Data sources used in the model include household-level data, firm-level data, origin-destination data, travel time data in traffic, and the National Household Travel Survey. The model outputs a long-formatted, flat file that details the start and end of events (home, work, other, travel) with a latitude/longitude/timestamp. The output can be systematically reformatted into a traditional relational database with person, firm, trip, and tour tables.

For comparison, we took advantage of the fact that a best-practices trip-based model was built for the same region in recent years. To compare the two models, we aggregated the simulation-based model output into demand tables with the same TAZ system and time periods used by the trip-based model. We then assigned both sets of demand tables[1] to the network using an identical static assignment procedure, which was built and calibrated for the trip-based regional model in TransCAD. We compare the two model results using standard measures of model output.

A limitation of this methodology is that it does not currently support multimodal travel. The Asheville region has a sufficiently high auto travel rate (estimated mode split of auto 94.5 percent, walk/bike 5.2 percent, and transit 0.3 percent) that mode designation does not affect the overall model accuracy. However, future work in cities with substantial transit systems requires that multimodal capability be part of our future research agenda.

RESULTS

The passive data model produces effectively equivalent levels of accuracy when compared with the aggregate trip-based model using standard validation measures. The FBRMPO aggregate trip-based model produced a total of 1,292,984 internal household trips (8.8% intrazonal), and the passive data tour-based model produced a total of 1,029,560 trips (8.7% intrazonal). Table 1 and Figure 1 present the validation results. The results reveal that the trip-based model and its static assignment model are well-

TABLE 1 Validation Results

			Aggregate Trip-Based Model			*Tour-Based Model from Passive Data*		
Facility Type	n	Avg Count	VHT	RMSE	PRMSE	VHT	RMSE	PRMSE
Freeways	103	22,935.6	62,051.2	3,321.6	14.5%	78,042.0	5475.1	23.9%
Expressways	6	13,732.2	2,022.9	2,168.2	15.8%	1,965.0	1984.5	14.5%
Boulevards	15	15,965.1	2,081.2	5,695.2	35.7%	2,302.2	5920.4	37.1%
Other Major Thoroughfares	337	10,322.2	31,646.6	4,181.5	40.5%	32,980.0	5113.7	49.5%
Minor Thoroughfares	414	2,938.9	18,360.0	2,026.8	69.0%	18,856.2	2359.8	80.3%
Total	875	8,433.7	116,162.0	3,250.3	38.5%	134,145.4	4106.4	48.7%

[1] The prototype passive data model only handles internal household trips to date; demand tables for commercial vehicles, visitors, and external trips were carried over directly from the trip-based model.

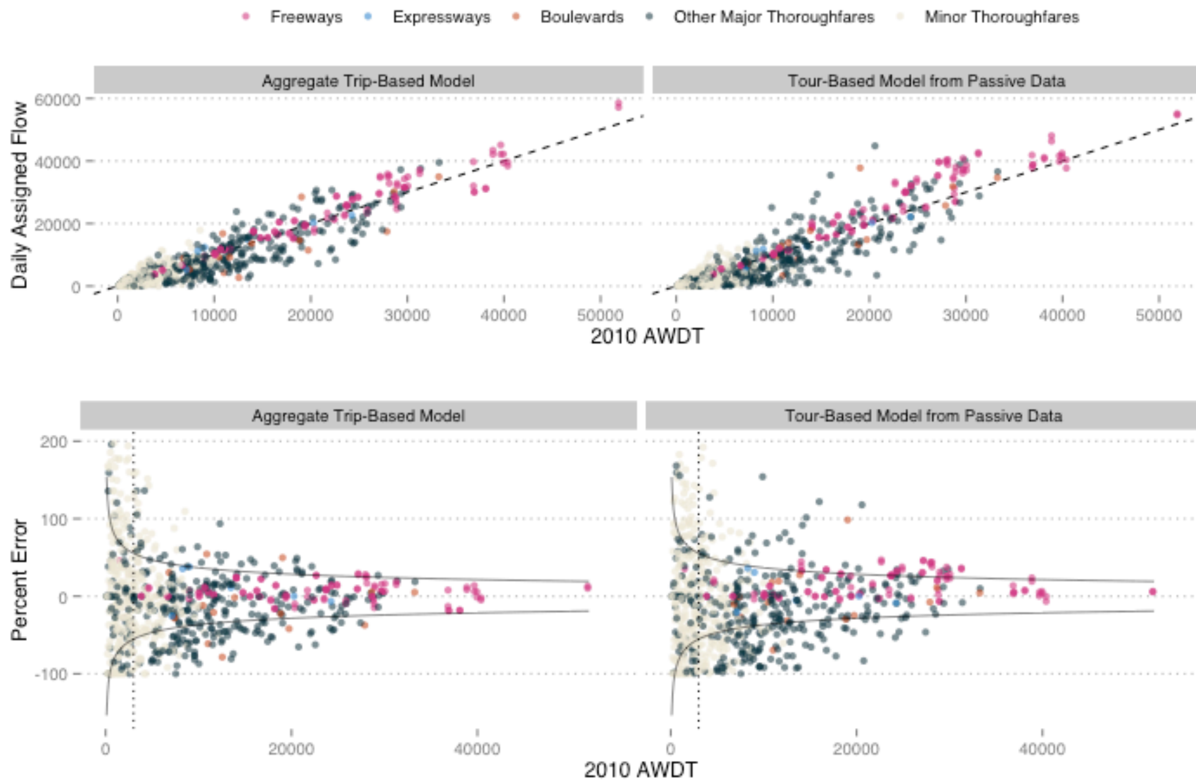

FIGURE 1 Validation results of both models after static assignment.
In the left column, the reference trip-based model validation is depicted. In the right column, the passive data tour-based model validation is depicted. The top row illustrates the count scatterplots, and the bottom row illustrates the maximum desirable deviation tolerances. The points are color coded by facility type. Note that 41 and 47 links in each model, respectively had percent error greater than 200%. However, these were all on Other Major or Minor Thoroughfares with AWDT volumes under 3,000 (shown with a vertical dashed line). For visual clarity, we do not show them.

calibrated to local conditions. At least 90 percent of the counts are less than the maximum desirable deviation, as shown in the bottom row of Figure 1.

The table and figure also show that the tour-based model from passive data produces *un-calibrated* assignment results that are as good as the calibrated trip-based model. Except for over-assigning the freeways, which might be overcome with future development of this prototype model and/or by calibrating the static assignment model, the passive data model statistics look very good. Not surprisingly, many of the same patterns found in the trip-based model are in this model. Both models have a tendency to over-assign low-volume major and minor thoroughfares. And both have one or two outliers that keep the percent RMSE statistics from looking superb.

The costs and time to produce this passive data model, when compared with the trip-based model development, was significantly lower. To make the cost comparison as conservative and equal as possible, we compare just the person-hours. For the trip-based model, we include just the portion of the development time spent on the resident trip model, totaling about 462 person-hours. This includes the processing and analysis of the survey data, estimation, implementation, calibration, and validation. This does *not* include the time and resources spent to collect the local travel survey or to develop the zonal system and zonal data used in the trip-based model (which are not used in the passive data model), the time to process the on-board transit survey or to develop the mode choice model, nor does it include the time spent in meetings coordinating the development and training. The passive data model development in the Asheville region was completed in about 75 person-hours. This includes processing all the passive datasets, running the simulation model, processing the results into a format accepted by the assignment model, and validation. It does not include, just like the trip-based model, time developing a mode choice model nor time in meetings with the local MPO and DOT.

To keep the comparison as equal as possible still, let us assume 50% availability of one worker. The person-hours would result in timelines of about 6 months and 1 month, respectively. Keep in mind that the trip-based model does require a local household travel survey, which adds about 6 months worth of work and, for the Asheville region, a cost of $272,000. The trip-based model also requires that the MPO spend time beforehand developing a zonal system and zonal data for the trip-based model. This would not be required for the new model due to the use of third-party household- and firm-level data with latitude/longitude locations. The cost to purchase third-party data varies according to many different dimensions, but such data collectively would cost much less than the cost of the local household travel survey alone.

For reference, the FBRMPO model began development in March of 2012 and was delivered in July of 2014 (2 years, 4 months), which included the local household travel survey, the resident trip-based model, a commercial vehicle model, an external trip model, and a mode choice model. The passive data model development began in December 2015 and completed in March 2016 (4 months), which included the tour-based model and refactoring the codebase that runs the simulation.

In terms of complexity and usability, the passive data model also excels. The model run-time is much shorter, allowing planners to run the model repeatedly under different scenarios. The demand portion of the trip-based model runs in about 4 hours. Conversely, the new prototype model runs with a full population in just under 1 hour on a personal laptop. Additionally, the model output is much more detailed. With the passive data model, individual travel diaries are available with tour-level information that is linked to households and firms. Future development will include linking the detailed model output with an open-source dynamic traffic assignment model, such as MATSim that can include a mode choice model sensitive to costs derived from the network in simulation-time. This kind of interoperability and detail can facilitate analysis for many more regional questions than an aggregate, trip-based model can.

Lastly, the model codebase itself is simple and robust relative to existing modeling practice. The code is written in a combination of R and Python, relying on other well-built open-source libraries wherever possible. As a result, the simulation engine itself fits in one file that is less than 600 lines of code. These 600 lines of code will not change when running the model in Asheville, North Carolina, in Seattle, Washington, or in Atlanta, Georgia; nor will they change with updated input data in subsequent years due to the fact that the model is not calibrated with any custom factors. Following this method allows developers to push updates and bug fixes to all model users. It also frees up time to invest in novel and transferrable analysis tools.

IMPLICATIONS

For many small- and medium-sized communities, household travel surveys are out of reach due to the high cost of conducting them. For those communities that can afford a household travel survey, cost constraints often limit them to a small sample size resulting in data that are behaviorally rich, but pose significant challenges in the disaggregation of the data for better understanding travel markets. As a result, these communities are often left with more aggregate travel models that may not fully represent the diverse travel behavior within the region. This may further impact the application of these aggregate models with respect to understanding the tradeoffs between various land use and transportation alternatives that a community may wish to evaluate. We believe the agile modeling approach developed through this research is a good fit for small- and medium-sized communities because it will help them overcome the challenges of costly data collection and of aggregate models that may not fully understand travel behavior. It will offer the benefit of person- and tour-based analysis without the costs and development requirements of an activity-based model. The universal applicability of the simulation engine as well as the reduced needs for manual data collection can also save these regions in development costs.

As we all know, model development is only the beginning of the story because that is when the work of forecasting and planning begins, without which we would not need the model. When combined with an assignment model, we foresee this type of simulation-based demand model as an effective tool for both forecasting work and for scenario analysis. With forecasting work, one usually alters the input population to match a future forecast and reruns the base year demand model with the different population. The simulation-based model could function the same way, with population estimates being developed externally just as is typically done today. As an added benefit though, with the lower costs and time requirements of this new passive data model, communities would also be able to update the model much more frequently than current practice allows. This could better inform population forecasts and provide opportunities for more timely before-and-after analysis to inform future forecasting work. During scenario analysis, estimating mode split and network loading would function in much the same way as a traditional model when using the simulation-based model with a traffic assignment model that includes a mode choice model sensitive to network and land-use changes.

Because the simulation-based demand model itself is not calibrated with adjustment factors, the input behavioral data can be tweaked in a straightforward manner. For instance, analysts working with state-of-the-practice models cannot adjust parameters that would create more trip chaining behavior without upsetting the model calibration results in other interrelated but non-obvious steps of the model. In the simulation model by contrast, the user can simply increase the number of activities on a tour and the simulated individuals' daily patterns will adjust automatically, leaving activity durations and paths between them relatively unchanged. Similar conclusions can be drawn with changes in the rate of work trips versus non-work trips or changes in commute departure times. Due to the tight relationship between model input and model output, it may also be easier to capture uncertainty in all of these.

STATEMENT OF INNOVATION

This study contributes to the modeling community by testing out a novel approach for modeling travel with passive data rather than local survey data.

STATEMENT OF FINANCIAL INTEREST

The authors do not have any direct financial interest with regard to the publication of this research. The research was self-funded in an effort to advance practice.

REFERENCES

[1] Roux, Sophie, and Jimmy Armoogum. Calibration Strategies to Correct Nonresponse in a National Travel Survey. In Transportation Research Record: Journal of the Transportation Research Board, No. 2246, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 1-7.

[2] Son, Sanghoon, Asad Khattak, Xin Wang, Paul Agnello, and Ju-Yin Chen. Quantifying Key Errors in Household Travel Surveys: Comparison of Random-Digit-Dial Survey and Address-Based Survey. In Transportation Research Record: Journal of the Transportation Research Board, No. 2354, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 9–18.

[3] Bricka, Stacey, Johanna Zmud, Jean Wolf, and Joel Freedman. Household Travel Surveys with GPS: An Experiment. In Transportation Research Record: Journal of the Transportation Research Board, No. 2105, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 51–56.

[4] Wolf, J. Applications of New Technologies in Travel Surveys. Presented at the International Conference on Transport Survey Quality and Innovation, Costa Rica, 2004.

[5] Wolf, J., W. Bachman, M. S. Oliveira, J. Auld, A. Mohammadian, and P. Vovsha. Applying GPS Data to Understand Travel Behavior, Volume I: Background, Methods, and Tests. Publication NCHRP Report 775. National Cooperative Highway Research Program, Washington, D.C., 2014

[6] Gur, Y. J., S. Bekhor, C. Solomon, and L. Kheifits. Intercity Person Trip Tables for Nationwide Transportation Planning in Israel Obtained from Massive Cell Phone Data. Transportation Research Record: Journal of the Transportation Research Board, No. 2121, 2009, pp. 145–151.

[7] Huntsinger, L. F., and R. Donnelly. Reconciliation of Regional Travel Model and Passive Device Tracking Data. Presented at the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.

[8] Bachmann, C., M. J. Roorda, B. Abdulhai, and B. Moshiri. Fusing a Bluetooth Traffic Monitoring System with Loop Detector Data for Improved Freeway Traffic Speed Estimation. Journal of Intelligent Transportation Systems, Vol. 17, No. 2, 2013, pp. 152– 164.

[9] Lee, Richard J., Ipek N. Sener, and James A. Mullins III. Emerging Data Collection Techniques for Travel Demand Modeling: A Literature Review. State of Texas. Strategic Research Program. http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/161407-2.pdf