INTEGRATION OF THE NATIONAL LONG DISTANCE PASSENGER TRAVEL DEMAND MODEL WITH THE TENNESSEE STATEWIDE MODEL AND CALIBRATION TO BIG DATA

Vincent L. Bernardin, Jr., PhD
Director, RSG
2709 Washington Ave, Ste. 9
Evansville, IN 47714
Vince.Bernardin@RSGinc.com
Ph: 812-200-2351

Nazneen Ferdous, PhD
Travel Demand Modeler, CH2M
2411 Dulles Corner Park, Suite 500
Herndon, VA 20171
Nazneen.Ferdous@ch2m.com
Ph: 703-376-5000

Hadi Sadrsadat, PhD
Consultant, RSG
2200 Wilson Blvd., Suite 205
Arlington, VA 22201
Hadi.Sadrsadat@RSGinc.com
Ph: 888-774-5986

Steven Trevino
Analyst, RSG
2709 Washington Ave, Ste. 9
Evansville, IN 47714
Steven.Trevino@RSGinc.com
Ph: 812-200-2352

Chin-Cheng Chen
Forecasting Office Supervisor, Tennessee Department of Transportation
Suite 900, James K. Polk Building
Nashville, TN 37243-0334
chin-cheng.chen@tn.gov
Ph: 615-253-6301

Submitted: November 14, 2016
Word Count: 6,466 words + 4 tables/figures @ 250 words = 7,466 equivalent words

ABSTRACT

The Tennessee Department of Transportation chose to replace their quick-response-based long distance component in their statewide model by integrating FHWA's new national long distance passenger travel demand model into their new statewide model and calibrating it to long distance trips observed in cell-phone based origin-destination data from AirSage. The new national long distance model is a national scale, tour-based simulation model developed from FHWA research on long distance travel behavior and patterns. The tool allows the evaluation of many different policy scenarios including fare or service changes for various modes including commercial air travel, intercity bus, and Amtrak as well as highway travel. The availability of this new tool represents a new opportunity for state DOTs developing statewide models. Commercial cell-phone based big data on long distance trips also represents a new opportunity and a new data source on long distance travel patterns which have previously been the subject of very limited data collection in the form of surveys. This project is the first to seize on both of these new opportunities by integrating the new national long distance model with the new Tennessee statewide model and by processing big data for use as a calibration target for long distance travel in a statewide model. The paper demonstrates the feasibility of integrating the new national model with statewide models, the ability of the national model to be calibrated to new data sources, the ability to combine multiple big data sources, the value of big data on long distance travel as well as important lessons on its expansion.

Keywords: Long Distance Travel, Statewide Model, Travel Demand Forecasting, AirSage, rJourney

INTRODUCTION

The Tennessee Department of Transportation (TDOT) chose to implement an innovative approach to forecasting long distance passenger travel in their new statewide model. The standard practice for handling long distance passenger travel in statewide travel models is to add one or more special long distance trip purposes in a three- or four-step model structure, often borrowing parameters from studies such as NCHRP 735 (1). Instead, TDOT chose to integrate FHWA's new national long distance passenger travel model, rJourney, into their statewide model and calibrate it to long distance trips observed in cell-phone based big data.

Although long distance trips are much less common than short distance trips, because each trip has the potential to contribute so many vehicle miles of travel (VMT), these trips have a large and disproportionate effect on congestion and traffic on major intercity corridors such as I-40, I-75, I-24, and I-65 in Tennessee. A significant portion of long distance trips related to business travel also have notably higher value of time than most other trips, so reductions in delays for these trips can produce comparatively large economic benefits.

The availability of FHWA's new national rJourney model (2, 3, 4), together with the availability of new big data sources such as cell phone derived big origin-destination (OD) data, presented a new and exciting opportunity to dramatically improve the representation of long distance travel in Tennessee and allow new types of scenario analysis. For instance, the inclusion of a robust long distance mode choice modeling in rJourney allows the evaluation of scenarios such as increased air fares, expanded Amtrak service or new intercity bus services and the impact of such assumptions on highway volumes.

This is the first application of the national long distance model to support statewide modeling and forecasting, and is believed to also be the first use of big cell-phone based OD data to support development of a statewide travel model, although it is known that work to update the Virginia statewide model with similar data began at close to the same time.

The work to incorporate the new national model within TDOT's statewide model was part of a larger update to the model. TDOT originally developed a simple statewide model for Tennessee in 2003. TDOT developed a new, version 2, statewide model in 2014 to support development of their statewide long range plan. Although the version 2 model was also limited in sophistication due to the project schedule, it included three times as many zones (5) and road miles and offered much finer resolution in the representation of projects and their impacts. The version 2 model also included a truck model supported by the purchase of an eight week dataset of truck GPS based OD data from the American Transportation Research Institute (ATRI). The data included information from over 234,000 individual trucks on over 6.5 million truck trips representing roughly 11% of the trucks on the road for 56 days. The version 2 model also used Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics developed by the Census Bureau in cooperation with the Bureau of Labor Statistics which provides big OD data on commuting patterns based on administrative tax records.

This data-driven approach, albeit incomplete and supplemented by traditional synthetic quick response methods in version 2, lead to very good model performance. The model's highway assignment achieved impressive validation statistics versus traffic counts for a statewide model including a 37% root mean squared error (RMSE) and a correlation coefficient of 0.97.

Since the project schedule had precluded the incorporation of all the desired functionality and features in the version 2 model, TDOT embarked upon an update to develop the version 3 Tennessee Statewide Travel Model (TSTM3). The new model incorporates a new commodity flow freight model, an advanced trip-based model for short distance passenger trips including mode and destination choice models with non-home-based trips linked to home-based trips (6), as well as the integration of the new national model for long distance passenger trips. Building on the success of the data driven approach with the ATRI and LEHD data in version 2, TDOT purchased cell-phone based data to support the development of the version 3 model, with special filtering for long distance trips to support the calibration of the national long distance model in particular.

The paper below begins by describing the data and its processing before turning to describe the integration of the national model with the Tennessee statewide model designed to produce reasonable runtimes. The paper then documents the success of the project at calibrating the national model to the big OD data and its support of overall impressive validation statistics before offering some concluding thoughts.

INTEGRATED MODELING METHODOLOGY

The demand forecasting components of the Tennessee Statewide Travel Model (TSTM3) can be grouped into three sets of models: (i) freight demand models, (ii) short distance daily passenger demand models, and (iii) long distance passenger demand models. The latter, long distance models are the focus of this paper and include the national long distance passenger demand model together with several data manipulation and processing steps to achieve integration of and translation between the networks and zone systems for Tennessee and the national model. (See Figure 1.)

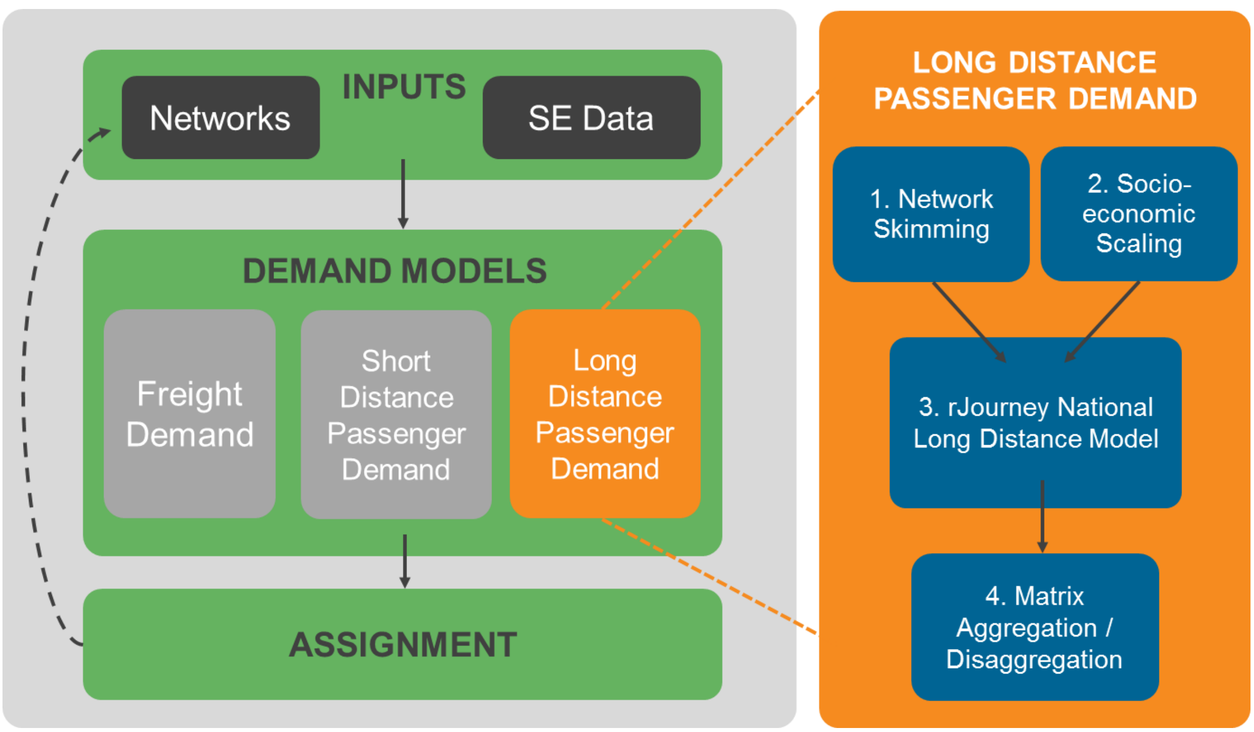

FIGURE 1 Tennessee Statewide Travel Model Integration with the National Long Distance Passenger Travel Demand Model

The first process involved in the integrated model run is the production of the highway skim for the national model using the Tennessee model network. The national model natively uses a highway network developed from the Oak Ridge National Highway Planning Network and a zone system of 4,570 zones or national use modeling areas (NUMAs) comprised of counties in rural areas and Census Public Use Microdata Areas (PUMAs) in urban areas. To facilitate the integration of the TSTM3 with the national model, the national model's highway network was added to the TSTM's model network outside of Tennessee, while the more detailed TSTM network was retained within (and immediately adjacent to) Tennessee. Outside of Tennessee the national model's original centroids and centroid connecters were retained. Within Tennessee, the TSTM centroid nearest the center of each NUMA was designated as the centroid for that NUMA. Highway skims for the national model were thus created using the national model's zone system but the TSTM's more detailed highway network within Tennessee. (A separate highway skim is used for the daily short distance passenger trips using the more detailed zone system within Tennessee, while a third highway skim with yet another zone system is used for the freight demand model.) This approach has the advantage of limiting the size and associated run time of the skim and not requiring processing to convert skim matrices between zone systems. Moreover, it allows the national model to be sensitive to network changes in Tennessee without requiring improvements to be coded in two separate networks (one for short distance and one for long distance models).

The second process required to integrate the TSTM3 with the national model is to ensure that the socioeconomic growth assumptions from the TSTM3 are incorporated into the national model's inputs. There are two parts to achieving this. The first is the scaling of the synthetic population for the national model to reflect population growth assumptions and the second is the update of the national model's destination choice size variables to reflect employment growth assumptions. The national model uses a detailed synthetic population covering the whole country. Control variables for the synthetic population come from the Census's American Community Survey (ACS) in the base year. Creating future year synthetic populations for the whole country poses several challenges. The process is both data and runtime intensive. Detailed demographic control total forecasts are required covering the whole country. Moreover, these control totals must be carefully crafted or checked to ensure that they are internally consistent (e.g., the distribution of households by size and the distribution of households by workers must have the same number of total households for each zone and the number of households with X workers cannot exceed the number of households with at least X people, etc.). Given both the data and computational challenges of synthesizing new populations for alternative future scenarios, an alternative approach was developed to simply rescale the base year synthetic population based on forecast growth in households. This approach has the limitation of not being able to reflect future changes in the characteristics (such as income) of households in an area, but allows household growth forecasts from the TSTM3 zones to be used to automatically and reliably update the national model's inputs with very limited runtime, without requiring the development of detailed socioeconomic forecasts for the whole country or complex data reconciliation.

The other part of the socioeconomic updating is the recalculation of the national model's destination choice size variables. Fortunately, this process is substantially simpler, and only involves the recalculation of formulas using updated employment data forecasts taken from the TSTM3. Thus, the national model reflects the household and employment growth scenarios from the TSTM3 zonal data within Tennessee without requiring the duplication of this data in another dataset. Employment and household growth outside Tennessee is taken from a simple input table with household and employment totals provided at the county level.

After creating the necessary inputs for the national model using the TSTM3's highway network and zonal data, the national model is run. The national model is a household level disaggregate tour-based simulation model. In some regards, it can be considered akin to a simplified activity-based model. (For details of the national model, see 4.) The national model's components begin with tour generation, scheduling, duration, and party-size models by purpose followed by mode and destination choice models similarly segmented by purposes including leisure/vacation, visit friends or relatives, personal business, commute, and employer's business. The national model includes four modes: highway, intercity bus, intercity rail, and commercial air travel. As described above, the national model uses the TSTM3's highway network for its highway travel times. It also uses the TSTM3 highway travel times to update intercity bus travel times. Tennessee implementation uses the national model's original networks for intercity passenger rail and commercial flights, but the user can adjust these to create alternative scenarios such as increased or decreased commercial air service or fares or new intercity rail service.

The final process in the integrated modeling system is matrix manipulation to convert the trip list from the national model, using its NUMA zone system, into a trip table matrix using the TSTM3's assignment zone system. This involves both the disaggregation of national model zones to TSTM3 zones within Tennessee and immediately surrounding areas and the aggregation of national model zones farther away from Tennessee. The demand within Tennessee (and nearby) is disaggregated based on a simple function of the socioeconomic characteristics of the TSTM3 zones within each NUMA, designed to approximate the number of long distance trips produced by and attracted to each zone. Demand farther from Tennessee is simply aggregated into a larger zone system (at the level of states for much of the country far from Tennessee) to keep the number of zones limited for assignment to help manage runtimes.

The resulting integrated system provided an efficient approach, allowing the national model to be run as part of the TSTM3 modeling system, using the information in the TSTM3's highway network and zone system. The long distance components of the TSTM3 including both the national model itself together with the ancillary pre- and post-processing procedures described runs in close to one hour on a machine with 12 physical cores and 32 GB of RAM.

CELL-PHONE DATA

As noted previously, this is believed to be the first use of big cell-phone based OD data to support development of a statewide travel model, although it is known that work to update the Virginia statewide model with similar data began at close to the same time. Large scale, aggregated, anonymous, passively-collected cell-phone OD data such as used in this study has a history of use for various purposes including the estimation of travel times (7) and origin-destination patterns (8) and resulting data has been compared to and incorporated in metropolitan area travel demand models (9, 10, 11, 12, 13, 14).

TDOT acquired origin-destination (OD) data from AirSage, Inc., for the state of Tennessee and a halo area surrounding it. AirSage aggregates and processes information from wireless data providers to provide mobility information such as trip tables. While the exact number of unexpanded observed trips is unknown, an extremely conservative lower bound can be established based on the number of OD pairs reported since at least one unexpanded trip of each type must be observed to be expanded. Using this method, the cell phone data was based on a minimum of 3,355,539 observed trips, although the actual number of observed trips is likely significantly higher. In contrast, combined household travel survey prepared for TDOT from an add-on sample to the 2008-2009 National Household Travel Survey (NHTS) and travel surveys from local metropolitan planning organizations in the state contained a total of 81,065 trips by 10,344 households in 39,782 OD pairs. Thus, the cell-phone data contains at least 84 times as many observations as the household and likely substantially more. The result is that the big cell-phone data provides a much more complete picture of OD patterns compared to the survey.

Another way to understand the difference in the completeness of the new big data versus traditional survey data is to consider the amount of the origin-destination space that the data covers or the percentage of cells within the origin-destination matrix with an observed frequency. TDOT's traditional household survey data included observations of 39,782 origin-destination pairs or 0.3% of the cells in the origin-destination matrix. In contrast, the cell-phone data included observations of 3,355,539 origin-destination pairs or 26.3% of the cells in the origin-destination matrix. This substantially better coverage offered by big data is one major motivation for its use to support travel modeling in general.

There is even further motivation for the use of cell phone or similar passively collected big data for studying and modeling long distance trips in particular. It generally takes significant additional effort in travel surveys to collect an adequate sample of long distance trips. For this reason, it has tended to be expensive and rarely done. For example, TDOT's combined travel survey dataset included only 1,076 long distance trips (over 50 miles in length) out of the 81,065 total trips, and these were clearly skewed towards long distance commute trips and the shorter end of the spectrum of long distance trips. Across the United States over the past twenty years only five useful attempts to collect a representative sample of long distance trips could be identified for the development of the national long distance model. While its anonymous nature precludes it from supplying the same kind of rich detailed information that surveys can, passively collected big data such as the cell-phone based data used in this study provides a cost effective alternative to at least for understanding long distance OD patterns.

In order to use the cell-phone based data to calibrate the national long distance passenger model, it was necessary to first remove commercial travel from the dataset because cell-phone data captures both personal and commercial trips since travelers, including truck drivers, carry their cell phones regardless of their travel purpose. As was noted in the introduction, TDOT had acquired and processed truck GPS data from ATRI. The initial plan was to simply subtract the truck ODs based on the truck GPS data from the total cell-phone based ODs. However, the initial attempt to do so revealed that there were more truck trips than total trips for 11% of the OD pairs observed in the cell-phone data. Although only 0.2% of the total cell-phone trips were involved, given the large number of OD pairs, this was considered problematic.

Upon investigation, it became clear that the primary reason for this was a difference in the way the two datasets were processed relative to the definition of trips and long distance trips in particular. The cell-phone data had been purchased with filtering to remove intermediate stops (such as for fuel, meals, etc.) on long distance trips. Based on AirSage's description of their methodology, if a traveler traveled 50 miles from home the criteria for defining a stop changed and rather than being based on the amount of time the traveler spent in the same place, instead, a stop was coded only when the traveler reached the point furthest from home and began traveling back towards home. In this way intermediate stops between home and the assumed destination at the farthest point from home are removed from the dataset. The truck GPS data was originally not processed in an analogous way, so it included intermediate stops on long distance trips. It was therefore necessary to re-process the truck GPS data, filtering out intermediate stops. However, it is less clear how to define home for trucks and in many cases much more difficult to identify than for most residents of an area who return home most nights.

Moreover, it was deemed important and desirable to allow for multiple destinations on a long distance tour (e.g., a truck carrying one shipment from Nashville to Knoxville may then pick up another shipment and take this to Chattanooga before returning to Nashville). For both these reasons, a slightly different algorithm was used for removing intermediate stops from the truck trips. When a truck traveled more than 50 miles from one origin, A, to a stop, B, based on dwell time, location B was not immediately logged as a stop. Rather, at the next stop C (based on dwell time), the distance between A and B plus the distance between B and C was compared to the direct distance between A and C. If the direct distance between A and C was more than 95% of the sum of the distance between A and B and between B and C, then B was considered an intermediate stop and removed, otherwise it was retained. Thus, this criterion identified intermediate stops based on whether the truck went out of its way to reach the location. This method allowed the removal of many intermediate stops while still allowing multiple “true” stops on a long distance tour.

After re-processing the truck GPS data using this algorithm to remove intermediate stops, the resulting truck ODs were again subtracted from the total cell-phone ODs. The number of OD pairs with more truck trips than total trips was reduced by 87% from 11% of the ODs to only 1.3% involving less than 0.1% of the total trips. Although still not perfect, the output was deemed acceptable because more than 98% of all cells have reasonable auto trips consisting 99.9% of total trips. The remaining OD pairs with more truck trips than total trips were reduced to a fraction of a trip (to retain the information that some sort of trip was observed and allow for expansion given some possibility that a passenger trip may have been observed). This experience points to the importance of a common definition of trips (or stops) when combining multiple big OD datasets.

The expansion of the cell phone data was tested and ultimately adjusted in a two stage process. While the details of AirSage’s data expansion algorithm are proprietary trade secrets, their documentation indicates that they use methods to expand their data based on the ratio of cell-phones to population data at the inferred residence location. This basic approach has been described and studied in academic literature (15, 16, 17, 18) as well as have more advanced methods that make use of traffic assignment and/or optimization methods (19, 20, 21) to expand cell-phone data based on traffic counts. The authors of this paper believe the latter methods to be generally superior because traffic counts provide unbiased information on the total amount of vehicular traffic on the road in various locations and the former methods based only on cell-phone ownership levels fail to allow for any variety of factors which can affect the detection of trips in particular locations, of particular durations, etc.

Initially, the cell-phone data was assigned to the Tennessee model network as a check on its expansion and in order to determine necessary scaling. It is typically necessary to scale cell-phone data to account for the number of cell phones per vehicle (closely related to but slightly different than vehicle occupancy). However, the test revealed relatively poor fit to traffic counts, (somewhat in contrast to experiences with using the data in metropolitan areas where simple scaling usually produces at least reasonable agreement with counts). In particular, when scaled to minimize total loading error, there was substantial underloading in urban areas on the order of -10% versus counts and substantial overloading in rural areas on the order of +15% versus counts. Since it is known that vehicle occupancy is substantially higher on long distance trips than short distance trips, the initial response was to attempt scaling the trips based on distance rather than uniformly as a whole. While this did improve the loading issues, it quickly became apparent that the difference in scaling required to address the loading errors could not be accounted for simply by higher vehicle occupancy on long distance trips. While vehicle occupancy may be two to three times higher on long distance trips than short distance trips, long distance trips appeared to be over-represented by a factor of ten or more. Other hypotheses were therefore also explored, such as that the bias may be related to area type or density rather than trip length. However, none of these explained or corrected the loading errors better than a distance-based correction, and generally they did worse. Therefore, a distance-based scaling was ultimately adopted, but based primarily on the hypothesis of a bias in the cell-phone data rather than on differences in vehicle occupancy. Upon further reflection, the possibility of a bias in cell-phone data toward the detection of long distance trips seems plausible. Since cell-phone mobility data depends on signaling between phones and towers, the likelihood of detection increases with the likelihood of this signaling and this signaling becomes more likely on longer trips for a variety of reasons including that people are more likely to use their phone the longer the trip. The probability of a person using their phone while on a local shopping trip is presumably much lower than the probability that a person uses their phone a trip to another city which is likely to take several hours if not a day or more. This line of reasoning provides at least a plausible explanation for a potential over-representation of long distance trips relative to short distance ones in a cell phone dataset.

While techniques for origin-destination matrix estimation (ODME) from counts without distance-based scaling could have been applied directly to simultaneously address distance bias and other potential issues with the expansion, the authors preferred a two-step process, first scaling trips parametrically based on functions of distance and using non-parametric ODME methods second. This approach helps avoid large adjustments from ODME without a clear understanding of the underlying problem or issue. A good review of ODME techniques, their limits, and effectiveness can be found in the study by Marzano et al. (22).

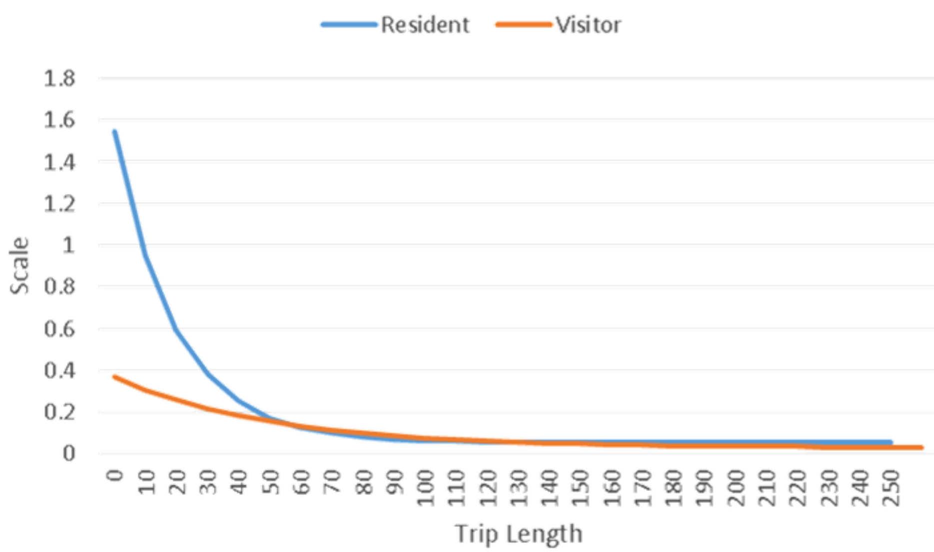

FIGURE 2: Distance-based Scaling Functions for Resident and Visitor Trips

Scaling functions were estimated separately for resident and visitor trips as visitor trips exhibited much more consistent over-representation independent of trip distance. This is consistent with the hypothesis of a bias against short trips since in this context visitors are already by definition (given the structuring of the data) long distance travelers. Parameters were fitted to the function scale = c + a x Exp(b x distance) using least squared errors. For resident trips, c = 0.0612, a = 1.6404, and b = -0.0507. For visitors, c = 0.0292, a = 0.3376, and b = -0.0195). The curves can be seen in Figure 2. The implication of the resident curve is that a 100-mile trip is 12 times as likely to be detected in the cell-phone data as a 10-mile trip. Given that there may be 2 to 3 times as many people on a 100-mile trip as a 10-mile trip, this suggests that a 100-mile trip is 4 to 6 times as likely to be detected than a 10-mile trip for reasons other than vehicle occupancy. The application of these scaling factors did not completely resolve the observed loading errors but significantly improved them reducing urban underloading to roughly -2% and rural overloading to roughly 5%.

After the distance based scaling, ODME techniques were applied to further improve the expansion of the cell-phone data versus counts. Careful consideration was given to setting appropriate bounds on the ODME adjustments. On the one hand, the most limited adjustments capable of producing good agreement with counts are desirable. At the same time, it is important to acknowledge and allow ODME to factor trips to and from certain areas up and down to account for varying degrees of cell coverage and other factors which can cause necessary expansion factors to vary beyond simply the variance in cell-phone market shares by resident areas. After some experimentation, ultimately, a minimum factor of 0.5 and a maximum factor of 5.0 were chosen to limit ODME scaling of any given OD pair. In addition to these limits, the average amount change in the trip matrix from ODME was closely monitored. The average absolute difference between cells in the final adjusted trip matrix and in the scaled matrix was 4.3 trips and the average absolute percentage difference was 1.5%. Together with the limits on minimum and maximum adjustments, these were deemed to be generally reasonable adjustments. The trip length frequency distribution of the adjusted matrix was also compared to the original matrix. The comparison showed that ODME resulted in a modest additional increase in the expansion of short distance trips versus longer trips. This seemed to suggest that the distance based scaling was not excessive but successful in accounting for most of the distance related adjustments. The ODME adjustments improved the fit of the cell-phone based data from 55.5% RMSE to 36.6% RMSE versus over 12,000 traffic counts across the state of Tennessee and given the relatively limited adjustments necessary to achieve this improved fit, this was deemed a successful and helpful improvement to the expansion.

RESULTS

The national model was calibrated to the cell-phone data primarily through the adjustment of constants in its component choice models. In particular, the calibration effort focused on the adjustment of the tour frequency and destination choice models, since the cell-phone data provided information primarily on these dimensions of long distance travel and did not provide information by mode.

For purposes of calibration comparisons, Tennessee zones were grouped into eight districts, each named for their largest/best known urban area(s). The total number of long distance trips bound to or from each district in the model and in the cell-phone data are compared in Table 1. As can be seen, the national model was able to be calibrated to closely reproduce the observed long distance trip generation rates observed from the cell-phone data. This was accomplished through the judicious adjustment of existing constants in the national model and without the addition of any special constants specific to these districts or other districts or zones in Tennessee. Most districts are within about 3,000 trips per day and less than 10% of their total. The Knoxville district is somewhat under-predicted, most likely because the Smoky Mountains and associated tourist areas attract more trips than predicted by the model. Trips to and from the

Tri-Cities district are over-predicted and the reason for this is less clear, but may be due to model not understanding the psychological and/or physical barrier posed by the mountainous topography of this area. Overall, however, the model does an impressive job of reproducing the number of trips observed for each district.

TABLE 1 COMPARISON OF MODELED AND OBSERVED (CELL PHONE DATA) LONG DISTANCE TRIPS BY TENNESSEE DISTRICTS

TN Districts	Observed	Modeled	% Difference
Tri-Cities	17,746	23,531	32.6%
Knoxville	59,149	53,239	-10.0%
Chattanooga	32,455	34,212	5.4%
Cookeville	22,486	21,239	-5.5%
Lynchburg	22,038	19,954	-9.5%
Nashville	88,502	85,622	-3.3%
Jackson	37,264	35,409	-5.0%
Memphis	30,340	31,067	2.4%
Total	**309,980**	**304,272**	**-1.8%**

Calibration of destination choice in the national model was more challenging. The cell-phone data revealed a significant bias against trips crossing the state border with the total number of long distance trips within the state slightly higher than trips to and from the state crossing the state border. The pattern cannot be predicted or explained on the basis of distance alone. For that reason, the gravity models based on NCHRP 735 in the version 2 TSTM could not reproduce the pattern, nor could the original national model. In order to reproduce the observed pattern, a single new term had to be added to the utility function of the national model's destination choice models to account for a psychological bias against crossing the state border. Similar psychological boundary effects associated with rivers, railroads, freeways, and governmental boundaries are commonly observed and incorporated metropolitan destination choice models. The addition of this term allowed the calibration of the national model to reproduce the pattern observed in the cell-phone data. No other district or zone or other special constants were added to the model specification.

TABLE 2 PERCENT DIFFERENCE BETWEEN MODELED AND OBSERVED (CELL PHONE DATA) LONG DISTANCE TRIPS WITHIN TENNESSEE

Origin districts	Destination districts								Total
	Tri-Cities	Knoxville	Chattanooga	Cookeville	Lynchburg	Nashville	Jackson	Memphis	
Tri-Cities	0.8%	0.3%	0.0%	0.0%	0.0%	-0.3%	0.0%	0.0%	0.7%
Knoxville	0.6%	2.0%	0.4%	-0.1%	0.0%	-1.5%	-0.3%	-0.3%	0.9%
Chattanooga	0.0%	0.1%	0.7%	0.3%	0.2%	-0.5%	-0.2%	-0.2%	0.4%
Cookeville	0.0%	-0.1%	0.4%	0.1%	0.0%	-0.6%	-0.1%	-0.1%	-0.3%
Lynchburg	0.0%	-0.1%	0.2%	0.0%	0.1%	-1.3%	0.1%	0.0%	-0.9%
Nashville	-0.3%	-1.4%	-0.2%	-0.9%	-1.5%	6.1%	-0.3%	-1.2%	0.4%
Jackson	0.0%	-0.3%	-0.1%	-0.1%	0.1%	-0.3%	-0.1%	0.8%	0.0%
Memphis	0.0%	-0.3%	-0.2%	-0.1%	0.0%	-1.1%	0.3%	0.1%	-1.2%
Total: All	1.0%	0.3%	1.1%	-0.8%	-1.0%	0.5%	-0.5%	-0.7%	0.0%

The same districts within Tennessee used for trip generation comparisons were also used to help evaluate the distribution of long distance trips within Tennessee. As can be seen in Table 2, the model is able to achieve very good agreement with the observed pattern of long distance trips within Tennessee. The total modeled trips to and from each district are within 1.5% of observed trips for the district, and with the exception of long-distance trips within the Nashville district, all the modeled district level OD flows are within 2% of the observed flows. The distribution of too many long distance trips within the Nashville district may be a result of the fact that the long distance destination choice models are more driven by distance than travel times, so congestion within the Nashville region may not be deterring as many trips as it should. Alternatively, it may simply reflect the inability of the national model to reproduce the complex long distance commuting patterns of this region given the limited spatial resolution of the national model. Despite this particular issue, the overall agreement between the modeled and observed data is quite good.

TABLE 3 PERCENT DIFFERENCE BETWEEN MODELED AND OBSERVED LONG DISTANCE TRIPS TO AND FROM TENNESSEE

Internal districts	External districts							Total
	Northwest	North Atlantic	Northcentral	Carolinas	Alabama-Gulf	Southwest	Georgia-Florida	
Tri-Cities	0.4%	0.1%	0.8%	3.6%	0.0%	0.2%	0.3%	5.3%
Knoxville	0.5%	-2.6%	-1.2%	-1.7%	-0.7%	0.3%	-2.0%	-7.3%
Chattanooga	0.0%	-0.1%	-0.5%	-0.4%	-1.1%	0.1%	2.7%	0.8%
Cookeville	0.0%	-0.2%	0.9%	-0.3%	-0.1%	-0.1%	-0.2%	0.0%
Lynchburg	-0.4%	0.1%	0.4%	0.0%	0.7%	-0.1%	-0.4%	0.2%
Nashville	-0.7%	-0.3%	6.6%	-0.8%	-3.6%	-2.3%	-2.0%	-3.1%
Jackson	0.0%	0.1%	0.6%	0.0%	0.0%	-1.9%	0.0%	-1.2%
Memphis	0.5%	0.3%	0.8%	0.1%	-0.1%	3.4%	0.3%	5.2%
Total	0.3%	-2.6%	8.4%	0.5%	-4.9%	-0.4%	-1.3%	0.0%

Trips to and from the eight internal Tennessee districts and seven external districts covering the rest of the country were also evaluated. As shown in Table 3, as with the internal trips, the national model is able to generally reproduce the observed pattern fairly well. The total modeled trips to and from each district are all within 8.5% of observed trips for the district and most are within about 5%, and with the exception of long-distance trips between the Nashville and Northcentral districts, all the modeled district level OD flows are within 4% of the observed flows. The model under-predicts trips to and from the Knoxville region, most likely underestimating the number of trips attracted to the Smoky Mountains (which is the most visited National Park) and associated tourist areas. The model also over-predicts trips between Tennessee and the Northcentral region, but the reason for this is less clear. Even so, the ability of the national model to reproduce the complex pattern of long distance trips to and from the state is quite good.

CONCLUSIONS

This paper has described the first integration of the national long distance passenger demand model with a statewide travel model and its calibration to cell-phone based OD data for the state of Tennessee, illustrating one of if not the first application of big OD data to statewide modeling. The case demonstrates the ability of the national model to be calibrated to observed data and of an integrated modeling system to produce reasonable runtimes. The case also illustrates the general importance of the processing of cell-phone OD data and hypothesizes an importance bias in cell-phone data towards the detection of long distance trips over shorter ones based on evidence from the Tennessee application. However, the case also illustrates the value of such data through, for instance, its ability to reveal important aspects of long distance travel patterns such as a psychological boundary effect corresponding to the state border in the case of Tennessee. While each state must evaluate the usefulness of various modeling approaches for their own planning and modeling, the case of Tennessee's new statewide model demonstrates that both the new national long distance model and cell-phone OD data can be successfully used and add value to a statewide model.

ACKNOWLEDGEMENTS

This research was sponsored by the Tennessee Department of Transportation in part through the use of state planning and research (SPR) funds.

REFERENCES

1. NCHRP Report 735: Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models. Cambridge Systematics. Transportation Research Board of the National Academies, Washington, DC, 2012.
2. Outwater, M., M. Bradley, N. Ferdous, C. Bhat, R. Pendyala, S. Hess, A. Daly, and J. LaMondia. Tour-Based National Model System to Forecast Long-Distance Passenger Travel in the United States. TRB 94th Annual Meeting Compendium of Papers, 2015.
3. Outwater, M., M. Bradley, N. Ferdous, R. Pendyala, V. Garikapati, C. Bhat, S. Dubey, J. LaMondia, S. Hess, and A. Daly. *Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework*. FHWA, U.S. Department of Transportation, 2015.
4. Outwater, M., M. Bradley, N. Ferdous, S. Trevino, and H. Lin. *Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework: Implementation Report*. FHWA, U.S. Department of Transportation, 2015.

5. Bernardin, V., E. Rentz, and B. Grady. TMIP How-To: Create Travelshed TAZs. Travel Model Improvement Program (TMIP) of the Federal Highway Administration (FHWA), Washington, DC, 2015.
6. Bernardin, V. and J. Chen. New Methods for Improving Non-Home-Based Trips in Trip-Based Models. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C., 2016.
7. Liu, H., A. Dancyk, R. Brewer and R. Starr. Evaluation of Cell Phone Traffic Data in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*. No 2086, 2008, pp. 1-7.
8. Calabrese, F., G. Di Lorenzo, L. Liu and C. Ratti. Estimating Origin-Destination Flows using Mobile Phone Location Data. *IEEE Pervasive Computing*. Vol. 10, No. 4, 2011, pp. 36-44.
9. 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.
10. Lee, R. J., I. N. Sener and J. A. Mullins. An Evaluation of Emerging Data Collection Technologies for Travel Demand Modelling: from Research to Practice. *Transportation Letters: The International Journal of Transportation Research*. Vol. 8, No. 4, 2016, pp. 181-193.
11. Huntsinger, L. F. and R. Donnelly. Reconciliation of Regional Travel Model and Passive Devise Tracking Data. Presented at the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
12. Bindra, S., B. Grady and J. Deshaies. Using Cellphone Origin-Destination Data for Regional Travel Model Validation. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
13. Zhang, W., A. Kuppam, V. Livshits and B. King. Evaluation of Cellular-based Travel Data – Experience from Phoenix Metropolitan Region. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
14. Milone, R. Preliminary Evaluation of Cellular Origin-Destination Data as a Basis for Forecasting Non-Resident Travel. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
15. Calabrese, F., M. Colonna, P. Lovisolo, D. Parata and C. Ratti. Real-time Urban Monitoring using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*. Vol. 12. No. 1, 2011, pp. 141-151.
16. Wang, P., T. Hunter, A. M. Bayen, K. Schechtner and M. C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*. Vol. 2, No. 1001, 2012.
17. Wang, J., D. Wei, K. He, H. Gong and P. Wang. Encapsulating Urban Traffic Rhythms into Road Networks. *Scientific Reports*. Vol. 4, No. 4141, 2014.
18. Alexander, L., S. Jiang, M. Murga and M. Gonzalez. Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 240-250.
19. Ma, J., F. Yuan, C. Joshi, H. Li and T. Bauer. A New Framework for Development of Time-Varying O-D Matrices based on Cellular Phone Data. Presented at the 4th TRB Innovations in Travel Modeling Conference, Tampa, FL, 2012.

20. Iqbal, M. S., C. F. Choudhury, P. Wang, M. Gonzalez. Development of Origin-Destination Matrices using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63-74.
21. Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, M.C. Gonzalez. The Path Most Traveled: Travel Demand Estimation using Big Data Resources. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 162-177.
22. Marzano, V., Papola, A., Simonelli, F. Limits and Perspectives of Effective O-D Matrix Correction using Traffic Counts. *Transportation Research Part C: Emerging Technologies*, Vol. 17, 2009, pp. 120-132.