GEORGIA DOT RESEARCH PROJECT 14-31

FINAL REPORT

MACRO-MODELING AND MICRO-MODELING TOOLS FOR HOV-TO-HOT LANE ANALYSIS

Georgia Department of Transportation

OFFICE OF RESEARCH

15 Kennedy Drive

Forest Park, GA 30297-2534

TECHNICAL REPORT STANDARD TITLE PAGE

1.Report No.: FHWA-GA-16-1431	2. Government Accession No.:	3. Recipient's Catalog No.: N/A
4. Title and Subtitle: **Macro-Modeling and Micro-Modeling Tools for HOV-to-HOT Lane Analysis**		5. Report Date: June, 2016 6. Performing Organization Code:
7. Author(s): Randall Guensler, Ph.D., Sara Khoeini, Ph.D., Adnan Sheikh, Ph.D.		8. Performing Organ. Report No.: 14-31
9. Performing Organization Name and Address: Georgia Transportation Institute Georgia Institute of Technology 790 Atlantic Drive Atlanta, GA 30332-0355		10. Work Unit No.: N/A 11. Contract or Grant No.:
12. Sponsoring Agency Name and Address: Georgia Department of Transportation Office of Research 15 Kennedy Drive Forest Park, GA 30297-2534		13. Type of Report and Period Covered: Final; May 2014 – June 2016 14. Sponsoring Agency Code: N/A

15. Supplementary Notes:
Prepared in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract: This report summarizes the analysis of observed commuting changes after conversion of an existing carpool lane into a high-occupancy toll lane, on 15.5 miles of Atlanta I-85. The team explored the correlations between observed changes in travel behavior and the socio-spatial characteristics of the users using 1.5 million license plate observations coupled with census demographic data. Commuter response was evident in the observed changes in commuter choice to continue to use the new managed lane, or move to the general purpose lane after conversion. Aggregate-level socio-spatial analysis of the impacts of the Atlanta I-85 HOV to HOT conversion across demographic groups and socio-economic attributes are presented in this report. This study enhances the ability of modelers to integrate managed lanes into travel demand models, with respect to travel demand response across user characteristics. The report also introduces a comprehensive modeling framework for socioeconomic analysis of managed lanes. The methods developed through this can inform future Traffic and Revenue Studies and help to better predict the socio-spatial characteristics of the target market.

17. Key Words: Managed Lanes, HOT Lane, Carpool Lane Conversion, Tolling and Revenue,		18. Distribution Statement: No Restriction	
19. Security Classification (of this report): Unclassified	20. Security classification (of this page): Unclassified`	21. Number of Pages: 52	22. Price: N/A

Form DOT 1700.7 (8-69)

[This Page Intentionally Left Blank]

GDOT Research Project No. RP 14-31

Project Report (Final)

Macro and Micro Modeling Tools for HOV-to-HOT Lane Analysis

By:

Randall Guensler, Ph.D.
Sara Khoeini, Ph.D.
Adnan Sheikh, Ph.D.

School of Civil and Environmental Engineering
Georgia Institute of Technology

Contract with

Georgia Department of Transportation

June 2016

Table of Contents

Executive Summary

The goal of this report was to develop a simplified planning tool that could be used to assess the changes in commute patterns likely to results from the implementation of future HOT lanes in the Atlanta metropolitan area. The contents of this report, and the spreadsheet models derived for this report, are derived directly from the doctoral dissertations of Dr. Sara Khoeini and Dr. Adnan Sheikh, who completed and defended their dissertation work in 2014 and 2015, respectively (Khoeini, 2014; Sheikh, 2015). Modelers are encouraged to read both of these dissertations to better understand the extent of the data and complexity of the various modeling methods applied to the corridor usage data.

This report summarizes a case study analysis of the conversion of a high-occupancy vehicle (HOV) carpool lane to a high-occupancy toll (HOT) lane, implemented along 15.5 miles of Atlanta I-85 on Oct, 1 2011. The researchers used 1.5 million license plate observations, collected over two-year study period before and after HOV-to-HOT conversion, to identify the observed commutershed, or catchment area of commuters, for this facility. The revealed responses of commuters to the HOT conversion were evident in the changes of their choice to move into the general purpose lane from the HOT lane, move from the carpool lane into the general purpose lane, or continue to use the managed lane or general purpose lane after conversion. The license plate data also revealed changes in use patterns across the spatial domain before and after the lane conversion.

In the first part of this report, the researchers explore the correlations between changes in travel behavior and the socio-spatial characteristics of the commuters. The team then matched license plate data to aggregate Census demographic data to implement an aggregate-level socio-spatial analysis of the impacts of the Atlanta I-85 HOV to HOT conversion across demographic groups and socio-economic attributes. This report presents a spreadsheet-based implementation of the Khoeini (2014) modeling results that can be used for planning purposes to assess future managed lane implementation. The model predicts the observed change in managed lane use over the four-hour morning and evening peak periods. The model predicts a decrease in overall managed lane use over the four-hour peak, in part because fewer individuals will pay to use the HOT lane on the shoulders of the peak (there is already enough capacity on the general purpose lanes). The higher-level of use during the peak-of-the-peak, when congestion is highest, coupled with a decrease in lane use overall for the four-hour period, can be used in tolling and revenue studies. The results of the Khoeini (2014) dissertation will also enhance the ability of modelers to integrate managed lanes into travel demand models, with respect to travel demand response to user characteristics. Khoeini's (2014) dissertation also introduces a comprehensive modeling framework for socioeconomic analysis of managed lanes. The methods developed through her work can inform future Traffic and Revenue Studies and help to better predict the socio-spatial characteristics of the target market once transferability of the models are confirmed.

List of Tables

List of Figures

1 Background Literature Review

The literature review for this report is derived from the doctoral dissertations of Dr. Sara Khoeini and Dr. Adnan Sheikh, who completed their dissertation work in 2014 and 2015, respectively (Khoeini, 2014; Sheikh, 2015). The literature review elements are broken into background discussions on macro-modeling of corridor-level commuter response and micro-modeling of individual-level commuter response. The macro-modeling methods primarily focus on identification of the commutershed, or catchment area, of commuters that will respond to changes in managed lane operation. Macro-level analyses are used in tolling and revenue studies to assess the likely aggregate responses of commuters choosing to use the managed lane or general purpose lanes as a function of socio-economic variables. The micro-modeling literature is more focused on the derivation of individual commuter response to changes in price and congestion levels. Such micro-level models can be useful in setting real-time toll prices designed to maximize vehicle throughput, maximize revenues, minimize emissions, or implement any other or combined optimization function.

1.1 Macro-Modeling Literature

To date, most studies about HOT lanes socio-economic impact have focused on stated preferences toward the use of HOT lanes (Ross et al., 2008, Burris et al., 2007; Dill and Weinstein, 2007; King et al., 2007; Douma, et al., 2005; Munnich and Loveland, 2005; Hultgren and Kawada, 1999; Sullivan, 1998; Supernak, and Golob, 2002). For example, the first comprehensive study conducted by Sullivan and associates (Sullivan, 1998) primarily analyzed the impacts of the HOT lanes on travelers' choice, and assessed public opinions on value pricing.

Although stated-preference surveys have served as the primary basis for socioeconomic analysis of pricing systems, these studies have mainly been conducted before the project implementation, and in some cases did not match with the revealed-preferences (Hultgren and Kawada, 1999; Munnich and Loveland, 2005). Moreover, the sample sizes are very small and, given the basic methods for collecting stated-preference data, may be biased if respondents provide answers they believe researchers expect to hear. The other problem with previous stated-preference studies is the lack of solid statistical analysis in modeling decision making in response to the pricing.

A few recent studies have touched upon the factors that are associated with HOT lane usage, using more advanced statistical methods. However, almost all of the studies conducted statistical modeling based upon the stated-preference/travel diary surveys from a small sample of the population. For example, Li, et al., (2007) examined the determinants of HOT lane use with the first comprehensive survey data (sample size = 759) on the State Route 91 Express Lanes in California using multivariate logistic regression.

As with socioeconomic impact assessment studies, lane use usage studies are usually based on small samples. Model goodness of fit is low, and study results are not consistent across different studies. For example, while income was consistently significant in predicting managed lane usage across all of the studies, the magnitude of the predicted impacts differed considerably. Moreover, age and gender were not significant across all of the studies (Burris, et al., 2012; Devarasetty, et al., 2012; Li, 2007). Other critical socioeconomic variables, such as ethnicity, have never been identified as significant in previous studies. More importantly, because these studies did not have data to study changes in users' choice in response to pricing over a long enough time period, they lack the power to respond to the main environmental justice question which is looking at the disproportionately adverse impact across demographic groups.

The main data source for the development and calibration of the Atlanta travel demand model used in the traffic and revenue study for Atlanta was a household travel survey of eight thousand households, conducted for the Atlanta Regional Commission (ARC) from April 2001 through April 2002 (Jacobs, 2009). However, the HOT lane usage patterns are likely to be significantly different from general corridor usage patterns. Moreover, the socioeconomic analysis was conducted at the county level by simply assessing county-level socio-economic characteristics and their trends over time, without any link back to previous projects. Such limited methods do not provide adequate arguments for the potential market share of a managed lane. Use of the standard travel demand modeling approach to forecast demand for the HOT lane under operating conditions that included pricing, and market sector response to pricing, has not been inadequate. To date, HOT lane demand still exceeds capacity under the maximum toll price on the Atlanta I-85 HOT corridor. Unfortunately, not enough research has been conducted to link the previous projects performance analytical results to the future projects traffic and revenue studies, especially in terms of socioeconomic impacts and targeting market.

Over the past decade, the survey sample sizes have dropped considerably, and are more often now in the range of 2,500 - 10,000 households, representing less than 1% of households in a metropolitan area. Furthermore, two-day travel surveys have been reduced to one-day travel surveys. This is especially problematic because household travel patterns do not stabilize for more than 20-days (Schönfelder, et al., 2006), meaning that researchers must rely on very large representative cross-sectional samples to reduce the potential bias introduced by short-duration sampling (Xu, 2010). The process of conducting surveys is very labor intensive and surveys have become very expensive, and more sensitive with respect to public privacy (Stopher and Metcalf, 1996). For example, Atlanta Household Travel survey conducted in 2011 cost two million dollars for collection of about 10,278 households socioeconomic and trip data, which represents less than 0.5% of the metro area population (ARC, 2011). A household travel survey by the Durham-Chapel Hill-Carrboro Metropolitan Planning Organization required $208 per completed survey (AirSage, 2013).

Another trend in household travel surveys is increasing non-response rates. Furthermore, many of the households that are non-respondents travel more than the average, or are larger households (Kriger, et al., 2006), which potentially creates bias in the collected data. The fact that recent technology advances such as smart phones or high speed Internet may not be equally distributed among the population (income groups, age groups, etc.) could also introduce sample bias.

A recent major enhancement in travel data collection has been the use of passive location data by applying Global Positioning System (GPS) data loggers either in vehicle or hand-held devices. The latest advances technologies are GPS-enabled smart phones and RFID tag reads (Doherty, 2009). Electronic toll collectors identify the user by reading the user's RFID toll tag. Active/interactive technologies such as computer user interfaces and cell-phone apps have initiated the collection of socioeconomic attributes as well as detailed trip characteristics.

A joint project by researchers at IBM and MIT (Lorenzo, et al., 2012) concluded that fine-grained, extensive data from mobile phone networks "is providing us with a more comprehensive view of activity and mobility at the urban scale than travel diaries can possibly do on their own. It also enables us to shed light on hitherto invisible intra-personal variation in travel activity." Compared to data gathered from household travel surveys, cellular technology provided researchers with information about individual mobility with a lower collection cost, larger sample size, higher update frequency, and broader spatial and temporal coverage (Wang, et al., 2012). The cost of collecting cell phone data is relatively low. A recent study by the town of Sierra Vista, Arizona, measured travel across 80 districts, covering 16,000 square miles for 12 weekdays, and collected cell-phone data on more than six million trips for $10,000 (AirSage, 2013).

Lastly, by using high-resolution cameras with manual transcription or automatic license plate readers (ALPRs), researchers can identify license plate numbers and then compare those numbers to state motor vehicle registration databases to identify the household represented by the observation (Colberg, 2013; Guensler, et al., 2013). Accordingly, the high rate of license plate observation on each lane, route, and time period can provide valuable information about users travel behavior (Khoeini, et al., 2012). Changes in vehicle activity are readily discernable over extended periods of time.

While recent information technology advances and technological innovations produce accurate and large samples of trip information, they lack the valuable socioeconomic piece of information. Marketing data have recently been introduced in airport trip generation studies (Kressner and Garrow, 2012) as a potential household-level socioeconomic data source (Khoeini, 2014). Marketing companies collect household/individual-level data using credit reports and other self-report data and assign associated attributes to individual household addresses. The marketing companies also utilize imputation models for missing variables or households. The average cost of purchasing marketing data is less than 10 cents per household, which is significantly lower than household travel survey costs, even with the cost of procuring supplemented travel data (license plate data, cell phone data, etc.).

Targeted sub-regional household stated-preference surveys in a corridor can be used to develop models to predict how users may respond to the implementation of HOT lanes (or other managed lane strategies). However, stated-preference surveys are expensive, data have been sparse, and resulting models thus far have achieved mixed results at predicting responses. Existing travel demand models that are based upon sparse regional surveys may also useful in predicting corridor response, but the stated preference data lack the underlying resolution that are really needed to define the likely commutershed catchment area (based upon managed lane user characteristics) and potential household response to managed lane implementation. On the other hand, license plate data are relatively easy and cheap to collect in very large numbers (millions). When license plate data are combined with demographic data, it becomes possible to model the observed corridor use responses as a function of these socioeconomic and pricing variables. These responses would be physically observed in before-and-after studies. These statistically-derived models should, in theory, be transferrable to new sites and new projects.

1.2 Micro-Modeling Literature

Studies that combine both stated-preference (survey statements of what a respondent would hypothetically do under certain conditions) and revealed-preference data (field observations of what a respondent actually did under specific conditions) are typically described as the most valuable for assessment of consumer response to changes in price or operating conditions. The combination of methods can capture the benefits of both types of data and make up for the shortcomings of each method. For revealed-preference data, these shortcomings include the fact that only the 'most preferred' option is reported, there may be correlation among the different variables, there may be a lack of variation in the data, and important factors may be excluded (Sheikh, 2015). In the case of a managed lane, variables representing traffic conditions may be correlated with toll amounts and time savings, for example. With revealed preference data, the underlying behavioral causes may not be discernable from the available data (we cannot ask the respondent why they changed their behavior) and often researchers want to extrapolate to conditions for which data have not been observed (e.g., how will a traveler respond if gasoline prices reach $10/gallon). Combined studies allow researchers to reliably observe change and apply stated preference data to predict future responses.

Börjesson (2008) estimated mixed logit models for departure time choice using both stated-preference and a form of revealed-preference data. In this case, the revealed-preference data was extracted from a model of the CONTRAM network in Stockholm network. While the travel times from the model were simulated, they were described by Börjesson (2008) as "actual mean travel times." Börjesson (2008) describes the benefits of combined revealed-preference and stated-preference models, as in this case the revealed data are highly correlated. The author modeled departure time choice as a function of travel time variability, but "high correlation of mean travel time and travel time uncertainty in revealed-preference data [made] accurate estimation of the trade-offs unfeasible." Börjesson (2008) later cites this "high correlation between mean travel time and travel time uncertainty in revealed-preference data" as a primary factor in the lack of travel time uncertainty studies using revealed-preference-only data. The paper does not discuss whether perceptions of travel time uncertainty affect the departure time decision, or whether actual uncertainty is the contributing factor. A related issue appears in the research proposed here, as values of toll amounts, travel time, and volumes are very likely to be correlated. However, the article ultimately concludes that stated-preference data are "less trustworthy for trip timing analysis and forecasting," the goals of the paper.

Two of the most frequently cited revealed-preference studies are those by Lam and Small (2001) and Small (2005). Even these studies, however, used both revealed-preference data and stated-preference data. Lam and Small (2001) used surveys asking for vehicle occupancy, job characteristics, and other information. In this highly-cited study, the average travel times for the models were estimated using a "standard engineering algorithm" and volume and vehicle density from loop detectors. The resulting travel time savings for the California State Route 91 lanes under examination were 5.9 minutes in 1998, a value that Lam and Small describe as small in magnitude and which "makes [their] results vulnerable to measurement error." This value is within the same order of magnitude as the median travel time savings of the I-85 Express Lanes, relative to the entire corridor. An important note in this study, which also relates to other loop-detector based studies, is that there are "many assumptions required to convert loop detector data into speeds estimates" (Lam and Small, 2001).

The methodologies presented throughout this report do not have to rely on estimated travel times, as the data include actual travel times in both HOT and General Purpose lanes. The travel time data are direct observations from the toll payment and vehicle monitoring system retained in the State Road and Tollway Authority (SRTA) tolling database.

The study by Lam and Small (2001) used binomial logit models for the route choice models, and included various measures of variability. The Lam and Small paper includes a discussion of the endogeneity in the models, namely in the option to switch to another route. Lam and Small (2001) included time-of-day choice in their next models, but these supplemental data came from surveys, which cannot be repeated in the modeling work presented in our report. Lam and Small (2001) also address the issue that their data only cover a portion of the actual trip length in two ways, by ignoring the limitation (so that the effects are embedded in alternative-specific constants) and by estimating missing travel times. These two methods were also explored by Sheikh (2015) for the I-85 HOT corridor. Finally, Lam and Small (2001) examine transponder choice and find that "transponder installation has its own determinants, distinct from those of the daily decision of whether or not to use the transponder." As for the route choice results, the authors report that "work-hour flexibility [provided by surveys] and total trip distance seem to influence the daily decision of which route to take" (Lam and Small, 2001).

The Small (2005) paper estimated mixed logit models based on both revealed and stated-preference data, with some important points for this dissertation. Small (2005) notes that revealed-preference studies "have been hampered by collinearity among cost and travel-time variables" and that "they have not accounted for heterogeneity in cost or travel time elasticities." An interesting point is that the author does not name any of these revealed-preference studies. Similarly, the revealed-preference data used in the study is self-reported and comes from telephone surveys (Small, 2005).

A number of other studies contained relevant findings for the I-85 HOT lane analytical work. Liu (2007) was a rare study that used revealed-preference data, in the form of loop-detector data, to estimate mixed logit models of route choice. The main determinants in this study were "travel time, reliability, and cost." The Liu (2007) study was also unique in that it examined values of travel time and reliability as they differed with departure time; that is, it did not assume them to be constant across the hours under study. Liu (2007) did not include demographics in that work, which were explored by Sheikh (2015) for the I-85 HOT corridor. Hess (2005) discussed mixed logit models with positive coefficients for travel time; these models indicate that users gain more utility from longer trips. Hess (2005) notes that this utility "gain" is typically seen as the result of model misspecifications or the lack of explanatory power in the data, and proposes other interpretations (Hess, 2005). Goodall and Smith (2010) wrote a paper with some worthwhile methodological variations, such as separating "daily users" of the MnPASS HOT lanes from less frequent users in their models to achieve a much better fit. The paper concluded that "pricing has a negligible influence" on lane use because almost 90% of the facility users were daily users and drivers may "use the HOT lanes as insurance against unanticipated congestion." On the I-85 Express Lanes, however, only 3.5% (4231 out of 120582) of transponders used the priced facility more than 200 times in 2012, or an average of four times a week for 50 weeks of the year (Sheikh, 2015). Goodall and Smith (2010) also raise questions about whether current conditions or previous experience have a greater impact on lane use decisions.

Because the I-85 corridor is the most heavily-monitored in the nation, the I-85 HOT lane use data are comprehensive. Vehicles equipped with SRTA transponders are identified by multiple transponder tag readers along the corridor (about every 1/3 mile). The data contain date/time stamps for each transponder reading, providing information about entry and egress points for each trip. The data can be linked by SRTA back to specific accounts and repeat usage of the facility can be tracked, as well as the toll amount paid per transaction. The data stream also provides travel times through the corridor on the HOT lane. Unlike other managed lane corridors around the nation, the I-85 corridor also includes transponder tag readers in the general purpose lanes (at four or five points, depending upon direction). This means that accurate travel times are simultaneously available for the HOT lane and the adjacent general purpose lanes, yielding travel time savings and reliability data. Furthermore, because the transponder tag reads are available in both the HOT lane and general purpose lanes, the data stream allows analysts to identify use, and non-use of the facility by individual vehicles as a function of congestion levels and price. The dissertation by Sheikh (2015) explores facility use characteristics as a function of price, congestion, and demographic parameters at the micro-modeling level, using techniques similar to those employed by Khoeini (2014) at the macro-modeling level.

1.3 Literature Review Summary

Targeted sub-regional household stated-preference surveys in a corridor can be used to develop models to predict how users are likely respond to the implementation of HOT lanes (or other managed lane strategies). However, stated-preference surveys are expensive, data have been sparse, and resulting models thus far have achieved mixed results in reliably predicting observed response data collected after implementation. Existing travel demand models that are based upon sparse regional surveys are also useful in predicting corridor response, but the data used to develop the regional model lack the underlying resolution needed to define the likely commutershed catchment area (based upon managed lane user characteristics) and their potential response to managed lane implementation. On the other hand, license plate data are relatively easy and cheap to collect in very large numbers (millions). Presuming that license plate data can be effectively combined with demographic data, it is possible to model the observed corridor use responses as a function of these socioeconomic and pricing variables (Khoeini, 2014). The models are based upon physically observed data collected in before-and-after studies. These statistically-derived models should, in theory, be transferrable to new sites and new projects. Similarly, when use and non-use data are available from a managed lane system, refined models can be developed from observational data to assess how commuters respond to changes in price and congestion level at the micro-modeling, day-to-day operational response level (Sheikh, 2015). Macro-level models for commutershed assessment and micro-level models for individual consumer response to congestion and pricing can be used in tolling and revenue studies, designed to assess aggregate-level response to implementation of managed lanes, and then to assess potential pricing structures, designed to set tolls for efficient operation of these lanes. These types of models provide two significant benefits: 1) the models are based upon very large samples of observed license plates, rather than small samples surveys, 2) the models are based upon revealed preference data (actual observation) rather than stated preference (opinion) data. If the models prove accurate, demonstrated through applications in future corridors, license plate observational and model development methods should be implemented on a widespread basis. These models would likely be further improved by linking the observational data with stated preference survey data collected through household surveys of corridor users.

2 HOV-to-HOT Commutershed Macro-Modeling Tool

The goal of the macro-modeling tool for HOV-to-HOT conversion analysis is to predict consumer use of the general purpose lanes and HOT lanes after conversion, as a function of observed general purpose and HOV lane use before conversion, and as a function of the socioeconomic attributes at the Census-tract-level along the corridor. In essence, the tool is designed to assess aggregate response that can be employed in tolling and revenue studies. The tool is developed through the analysis of before-and-after license plate observational data collected over a two-year period on the I-85 HOT corridor (Guensler, et al., 2013).

Household-level survey data are expensive to collect, and often infeasible for many transportation projects. Therefore the macro-modeling tool is designed to employ census-tract-level American Community Survey data, which are free and publicly-available. Census and other public data sources, such as the American Community Survey, do not provide demographic details at the household-level. Nevertheless, as demonstrated in Khoeini (2014) these data are still useful in assessing potential aggregate consumer response at higher-level resolution (i.e. at the corridor level). However, when reliable spatial information about individual corridor users becomes available, more refined data (at the household level) have the potential to significantly improve analytical work in travel behavior and socioeconomic studies (Khoeini, 2014, Sheikh, 2015). Hence, higher spatial resolution data should be used whenever the data are available and are demonstrated to be accurate. Although the modeling tool presented in this report is based upon aggregate Census data, Khoeini’s (2014) work demonstrates that purchased demographic data can also be used in these modeling approaches, using the same modeling tool, but applying different modeling coefficients derived through analysis of the higher-resolution data.

For the research and model development in this report, license plates were observed and linked to neighborhood Census data at the Census-tract-level (see Sheikh, 2015). License plates and household addresses are assigned unique identification numbers within the spreadsheet. To address potential privacy concerns, actual license plate numbers, addresses, and data collector names are not retained in the working spreadsheet and are not presented in this report. In the modeling tool, the salmon colored columns in the spreadsheet contain the final text values that substitute for the original license plates collected in the field, the addresses, and the names of data entry staff members (copied from adjacent columns, e.g. plateIDText is copied from plateID). These columns were copied from the calculated values so that the final unique values could remain in the spreadsheet after plate and addresses are removed. Similarly, high-resolution latitude and longitude values for households are perturbed (random changes to the third decimal place) and then retained only to the third decimal place in the final spreadsheet so that retained position data cannot be used to identify household locations. Researchers that desire to implement the model via the spreadsheet, will need to re-activate these columns in the modeling tool worksheets so that new field-collected license plate and address data can be used.

The report subsections that follow describe the Excel-based modeling tool derived from Khoeini's (2014) dissertation. Individual steps conducted in the model development include:

- 2.1 - Collect license plate data (field observation data)
- 2.2 - Identify and recode plates (manage duplicate observations/anonymize)
- 2.3 - Manage data collector IDs (anonymize and retain for quality assurance)
- 2.4 - Obtain and geocode registration data for plates to link with Census data
- 2.5 - Identify and recode addresses (manage duplicate observations/anonymize)
- 2.6 - Summarize lane use by unique household ID (managed vs. general purpose)
- 2.7 - Obtain and store Atlanta Census tract data for use in data joins
- 2.8 - Join lane use data with Census tract ID and data using geocoded address
- 2.9 - Derive model from observational data and demographics
- 2.10 - Apply model to observational data to predict census tract results

The remainder of this chapter describes each process above and the associated worksheet in the spreadsheet model. Additional chapter sections provide visualizations of the derived model results, including commutershed observation density plots, heat maps of observed changes in catchment response, and changes in distributional ellipses that define the commutershed.

2.1 License Plate Data Worksheet (plateObservationData)

The most accurate and cost-effective method for obtaining transportation facility use data is the collection and processing of license plate data (Khoeini, 2014). Although travel surveys can be employed to collect facility usage data, mail-out-mail-back surveys are relatively expensive (given the postage cost in both directions) and survey response rates are generally in the 5-10% range. Survey self-selection bias, non-response bias, and individual question response bias are problematic in such surveys. Surveys rely on stated-preference data, whereas license plate observations provide revealed-preference data with multiple observations in time and space, providing accurate lane use data. Accuracy and statistical robustness are significantly enhanced when license plate observation data are employed, versus using survey data or regional travel demand model outputs (Khoeini, 2014).

A license plate data collection methodology was developed by the Georgia Tech research team for a before-and-after monitoring study for the Atlanta I-85 HOV-to-HOT conversion (Guensler, et al., 2013). In this study, the research team visited five data collection sites each quarter for two years to assess changes in fleet composition over time (one year before the lane opened, and one year after the lane opened). Each site was located at an overpass with a good line of sight to the corridor. Field teams deployed cameras to record HD videos of the traffic stream, with the cameras focused on the rear of the vehicles (see Figure 1) so that license plates can be read from two lanes per camera view (high definition video is required to capture two lanes in one view).

Figure 1: License Plate Camera View and Data Entry Interface

In the HOV-to-HOT assessment project, the research team collected video data in both the AM and PM peak periods (Guensler, et al., 2013). Each peak session collected data for two hours: 7am-9am for the AM-peak and 4:30pm-6:30pm for the PM-peak. Because traffic around the metro area generally enters Atlanta in the morning and exits Atlanta in the afternoon, AM-peak sessions observed traffic in-bound to the city, while PM-peak sessions observed traffic out-bound from the city.

After returning to the lab, the high-resolution video was processed to record vehicle license plates. The high-definition video files collected during the field deployments were run through a freeware program called Free Video to JPG Converter available from DVDVideoSoft (http://www.dvdvideosoft.com). This program reduces the video files into a series of screen shots every 30th frame, equivalent to approximately two frames per second. The images were then fed into a video-processing program with a user interface that allows a data processor to enter the license plate number for each vehicle seen in the images (right panel in Figure 1 above). Because the video-processing program uses frame grabs, rather than the rolling video, data entry staff can tab through the images rather than having to pause and re-start a video file. The image processing interface results in faster processing times.

As an undergraduate research assistant types the plate entry for each screen grab into the software interface, the software links each plate number to its corresponding video frame number, date/time stamp, lane number, location, and data collector ID (Guensler, et al., 2013). When a license plate is unreadable, the processor records the vehicle as “missed” to allow for a reliable vehicle count. Several factors can result in a missed record when the video is processed. Low light levels, blurred video, tailgating, towing, and lane changes were the most common reasons for missed license plates. Using this methodology, license plate identification ranges from about 50%, under poor lighting conditions, to 95% under ideal conditions. During data collection periods with reasonable lighting, typical capture rates were on the order of 70% to 80%. Details on license plate data processing can be found in Katherine D’Ambrosio’s Master’s Thesis (2011) and a recent work zone research report prepared for the Georgia Department of Transportation (Suh, et al., 2013).

Practical implementation of license plate data collection can employ the method used by the research team, or any reasonable alternative method, such as automated license plate recognition (ALPR) software (Colberg, 2013). For example, the ELSAG Mobile Plate Hunter-900 (MPH-900) is a fixed ALPR system that can be mounted permanently to structures, such as bridges or overpasses, or can be mounted in a mobile configuration, typically on police vehicles (see Figure 2). According to the manufacturer, the MPH-900 ALPR system can read up to 1,800 plates-per-minute at 99% accuracy, and used by hundreds of law enforcement agencies across all fifty states per ELSAG North America (2013).

Figure 2: ELSAG ALPR Cameras Mounted to Tripods

As discussed above, the final output file from license plate data collection and processing steps includes the manually-entered (or electronically-collected) license plate characters, lane, date, time, location, data processor identification number, and a variety of related data (Table 1). The collectedPlateData worksheet in the HOV-to-HOT Macro-Modeling spreadsheet contains the license plate data file collected in the before-and-after study (Guensler, et al., 2013) for the third quarter of 2011, at one site, in which more than 36,000 unique license plates were processed. Each row in the worksheet represents a license plate record with the associated attributes. Actual plate numbers are replaced by uniqueIDs in the demonstration spreadsheet, so that the observed plate numbers can be omitted to address any potential privacy concerns. Model development by Khoeini (2014) actually proceeded with millions of license plates (multiple dates and sites), which were too numerous to include in the spreadsheet.

The keyID variable value is generated in the spreadsheet during data processing. The first digit represents the data collection season (8 quarters), the second indicates location (5 sites), the third represents day of week (Monday, Tuesday, Wednesday, and Thursday), the fourth represents morning or evening session (am, pm), the fifth indicates lane choice (0 corresponds to HOV/HOT, and 1-5 represent the general purpose lanes, left to right in the travel direction), and the remainder are random digits for privacy considerations.

The second step in the macro-modeling process is to match license plate data to household addresses for later use in joining observation data to demographic characteristics that will ultimately be used in the modeling processes.

Table 1: License Plate Data Worksheet Variables (plateObservationData)

Variable Name	Description	Notes
keyID	Unique record identified assigned in video processing	Generated in the spreadsheet as described above
date	Data collection date	Video start and date time are entered by the user into data entry software
timeStamp	Elapsed time since frame one of video at standard frame rate	Captured automatically by data entry software
frameNumber	Elapsed frames from start of video	Captured automatically by data entry software as elapsed frames
plateNumber	Plate number entered by research assistant	Removed from the demonstration spreadsheet
plateNumberNoSpace	Plate number with white space removed for consistency	Removed from the demonstration spreadsheet
plateID	Comes from a VLOOKUP equation using plateNumber in the uniquePlateCoding table	Replaces the plate number in spreadsheet join functions so that the plate number can be deleted from retained files
plateIDText	Text copy of calculated plateID	Used in all remaining join and count functions because calculated values (plateID) are no longer valid once license plates are deleted
vehicleClass	Vehicle class (if used for fleet composition research)	Not included in the demonstration spreadsheet
state	Plate state entered by analyst	Only Georgia plates are employed in analyses
lane	Lane number entered by analyst	Two lanes are processed per video, data entry software captures lane number
comments	Comments entered by analyst	User indicates the type of problem associated with single plate identification
userName	Research assistant user name	Removed from the demonstration spreadsheet
userID	Comes from a VLOOKUP equation using userName variable in the uniqueUserCoding table	Replaces actual userName in the spreadsheet join functions so that userName can be deleted from retained files
userIDText	Text copy of calculated addressLookup	Used in all remaining join and count functions because calculated values (userID) are no longer useable once names are deleted
dateTime	Date and time stamp derived from video frame number	Format: 6/10/2011 11:23:00 AM
quarterCode	Quarter during which data were collected	1 = Jan-Mar, 2 = Apr-Jun 3 = Jul-Sep, 4 = Oct-Dec

Variable Name	Description	Notes
site	Site location number	Coded for each site (spreadsheet contains data from the third quarter of 2011)
amPmCode	Morning vs evening peak	Calculated by spreadsheet for analysis
timeCode	Binary time code (AM=0, PM=1)	Calculated by spreadsheet for use in statistical analysis
dayCode	Day of week code	Calculated by spreadsheet for use in statistical analysis

2.2 Plate Coding Worksheet (uniquePlateCoding)

The uniquePlateCoding worksheet is employed to anonymize the license plate data contained in the plateObservationData worksheet. Every unique plate number in the observation data set is transferred to this worksheet so that a unique plateID can be assigned. For example, plate AAA1234 (if first in the unique plate list) would be assigned the unique plate ID “plate_1000001.” This way, every time the plate is observed, the unique plateID value can be substituted for the plate number and retained for public use. Table 2 contains the variable descriptions for the uniquePlateCoding worksheet.

Table 2: License Plate Data Worksheet Variables (uniquePlateCoding)

Variable Name	Description	Notes
uniquePlateNumber	Unique values for all plate numbers observed during data collection (eliminates multiple occurrences that appear in collectedPlateDataNoSpace)	Removed from the demonstration spreadsheet
ID	Sequential number from 1000001 to number of unique plates observed	
plateID	Concatenated value of “plate_” and sequential number (e.g., plate_1000001)	Replaces plate number in spreadsheet join functions so that the plate number can be deleted from retained files
plateIDText	Text copy of calculated plateID	Used in all remaining join and count functions because calculated values of uniquePlateID are not useable once license plate numbers are deleted

2.3 User Coding Worksheet (uniqueUserCoding)

The uniqueUserCoding worksheet is employed to anonymize the names of the data entry personnel (students are the "users") in the plateObservationData worksheet. Every unique user name in the observation data set is transferred to this worksheet so that a userID can be assigned. For example, Aaron Alton (if first in the unique user list) would be assigned the unique user ID "user_1001." This way, every time the user appears in the worksheet, the unique userID can be substituted for the identifiable name and retained for public use. Table 3contains the variable descriptions for the uniquePlateCoding worksheet.

Table 3: License Plate Data Worksheet Variables (uniqueUserCoding)

Variable Name	Description	Notes
uniqueUserName	Research assistant user name	Removed from the demonstration spreadsheet
ID	Sequential number from 1001 to number of unique plates observed	
userID	Concatenated value of "user_" and ID number (e.g., user_1001)	Replaces actual userName in the spreadsheet join functions so that user names can be deleted from retained files
userIDText	Text copy of calculated userID	Used in all remaining join and count functions because calculated values for userID are no longer useable once names are deleted

2.4 Registration Data Linkage Worksheet (registrationAddressesGeocoded)

The next step is matching observed license plates with addresses in the Department of Motor Vehicles Database so that the team can identify the applicable census tract and census block group for the observed vehicles for use in statistical analysis. To address privacy concerns, matching to the registration database is performed on a remote machine, by a third party. The research team sends a plate ID and a unique key ID for each record. The remote process returns the key ID and address. These records are also mixed in amongst tens of thousands of extra records to ensure that the file recipient must have the proper keys to conduct any matching to observed plates. In addition, this process ensures that none of the spreadsheets that are transferred contain both license plate numbers and the addresses at the same time.

Because the license plate numbers collected are not 100% accurate, due to environmental conditions and human error, only license plates that can be matched to addresses in the Georgia registration database are considered for further analysis. Given the very large numbers of plate data processed for such analyses (tens of thousands to hundreds of thousands of records), small random errors associated with errant plate recording or geocoding simply do not affect outcomes. Because the study only covers vehicles registered in Georgia, out of state vehicles are excluded from data processing (approximately 5% of the observed data are from out-of-state vehicles).

Once a set of household addresses are obtained from the registration database, the addresses are geocoded using an ArcGIS software geocoding routine to provide a latitude and longitude for each address. The latitude and longitude variables are carried to the 6th decimal place from the ArcGIS program for tracking purposes only. The accuracy of derived address location is not this accurate (six decimal places corresponds to sub-meter accuracy). Based upon previous experience, address data correspond to about 10 meter accuracy given the accuracy of ArcGIS address position data.

The registrationAddressesGeocoded worksheet also carries the lane number in which the plate was observed, for use in trip summary calculations and later modeling work. The lane number is encoded as the 5th character in the keyID during plate data entry. This value is pulled back out of the keyID using the Excel Mid function and placed in the lane column. This value will be used again later in the laneUseByHousehold worksheet to summarize the number of observations of vehicles from this household by lane number.

The registrationAddressesGeocoded worksheet illustrates the results of the geocoding process. Each row represents a household in the registration database address and associated latitude and longitude. Because data are collected over multiple days, license plates are often observed more than once in the data set. In addition, it is possible that multiple vehicles from the same household will be observed amongst the tens of thousands of plates. Hence, multiple observations in the plateObservationData often correspond to a unique address in the registrationAddressesGeocoded worksheet. Table 4 contains the description of the variables in this worksheet.

Table 4: Registration Data Linkage Worksheet Variables (registrationAddressesGeocoded)

Variable Name	Description	Notes
keyID	Unique record identified assigned in video processing	Assigned automatically by data entry software. Carried into the offsite matching process
address1	Street address	Returned from the address match process
address2	Suite number	Returned from the address match process
city	City	Returned from the address match process
state	State	Returned from the address match process
zip	9-digit zip code	Returned from the address match process (no-hyphen used)

Variable Name	Description	Notes
latitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
longitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
latitudeRandTrimmed	Latitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
longitudeRandTrimmed	Longitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
latitudeRandTrimmedText	Text copy of calculated latitudeRandTrimmed	Carried for instructional purposes because latitude data are removed from working spreadsheet
longitudeRandTrimmedText	Text copy of calculated longitudeRandTrimmed	Carried for instructional purposes because longitude data are removed from working spreadsheet
lane	Managed Lane = 0 General Purpose Lanes = 1 - 5 (pulled from keyID)	Lane number is encoded in the keyID as the 5th digit by the software during data entry
addressLookup	Concatenated address (formula)	Used in address lookup functions
householdID	householdID comes from a VLOOKUP equation using addressLookup in the uniqueAdressCoding table	Replaces physical address in spreadsheet join functions so that the address can be deleted from retained files
householdIDText	Text copy of calculated addressLookup	Used in all remaining join and count functions because calculated values (householdID) are no longer useable once address data are deleted

2.5 Address Coding Worksheet (uniqueAddressCoding)

The uniqueAddressCoding worksheet is employed to anonymize the physical addresses that are derived from the registration data worksheet (registrationAddressesGeocoded). Every unique address in the registrationAddressesGeocoded worksheet is transferred to this worksheet so that a unique household ID can be assigned to each address. For example, 1201 Ash Place (if first in the unique address list) would be assigned the unique household ID “hh_100001.” This way, every time the user appears in a worksheet, the unique householdID can be substituted for the identifiable address and retained for public use. Table 5 contains the variable descriptions for the uniquePlateCoding worksheet.

Table 5: Household Data Worksheet Variables (uniqueAddressCoding)

Variable Name	Description	Notes
uniqueAddress	Unique values for all addresses received from registration data	Removed from the demonstration spreadsheet
ID	Sequential number from 100001 to number of unique plates observed	
householdID	Concatenated value of “hh_” and ID number (e.g., hh_100001)	Replaces actual address in spreadsheet join functions so that addresses can be deleted from retained files
householdIDText	Text copy of calculated householdID	Used in all remaining join and count functions because calculated values for householdID are no longer useable once names are deleted

2.6 Household-Level Lane Use Analysis Worksheet (laneUseByHousehold)

The laneUseByHousehold worksheet summarizes the number of observations for each household in the HOV lane and in the general purpose lanes.

Table 6 summarizes the variables carried in this worksheet. All of the initial variables were derived previously and are the same as employed in the previous worksheets. These variables are carried into this worksheet so that they can be referenced in analytical work. Physical address data are removed from the final files. The counts in the HOV and general purpose lanes are developed using the COUNTIFS function, to count occurrences in the registrationAddressesGeocoded worksheet (which contains final address and lane use information for each observation). The total number of observations in the registrationAddressesGeocoded worksheet is the count of all rows for which address, city, and zip code match. The total number of HOV observations in the registrationAddressesGeocoded worksheet is the count of all rows where address, city, and zip code match, and the lane value is zero. The total number of general purpose lane observations is total observations minus HOV observations.

Table 6: Lane Use by Household Worksheet Variables (laneUseByHousehold)

Variable Name	Description	Notes
ID	Sequential ID assigned to each record	
address1	Street address	Returned from the address match process
address2	Suite number	Returned from the address match process
Address	Concatenated address from address1 and address2	

Variable Name	Description	Notes
householdID	householdID comes from a VLOOKUP equation using addressLookup in the uniqueAdressCoding table (calculates as “#N/A” in the spreadsheet because addresses were deleted for privacy)	Replaces physical address in spreadsheet join functions (address is deleted from retained files)
householdIDText	Text copy of calculated addressLookup	Used in all remaining join and count functions because calculated values (householdID) are no longer useable once address data are deleted
city	City	Returned from the address match process
state	State	Returned from the address match process
zip	9-digit zip code	Returned from the address match process (no-hyphen used)
latitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
longitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
latitudeRand Trimmed	Latitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
longitudeRand Trimmed	Longitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
latitudeRand TrimmedText	Text copy of calculated latitudeRandTrimmed	Carried for instructional purposes because latitude data are removed from working spreadsheet
longitudeRand TrimmedText	longitudeRandTrimmed	Carried for instructional purposes because longitude data are removed from working spreadsheet
gp	Number of household vehicle observations in the general purpose lanes (lanes 1-5)	Total observations minus HOV observations
hov	Number of household vehicle observations in the HOV lane (lane 0)	Count of rows in registrationAddressesGeocoded with address, city, and zip code match, and lane = 0
total	Total household vehicle observations in the HOV and GP Lanes	Count of rows in registrationAddressesGeocoded with address, city, and zip code match

2.7 Census Tract Data Worksheet (atlantaTractData_ACS2009to13)

The "atlantaTractData_ACS2009to13" worksheet contains socioeconomic data for all of the Census tracts in the Atlanta metropolitan area. The data source is the American Community Survey five-year summary file (2009 to 2013). These data were the most accurate and up-to-date publicly-available socioeconomic data at the time the analyses were prepared.

The US Census Bureau provides publicly-available household socioeconomic data aggregated by geographic boundaries (block groups, tracts, counties, etc.). Until 2000, household socioeconomic data were collected through decennial census long-form surveys, from about one in every six households. Long-form data were not collected in 2010. Starting in 2005, the American Community Survey (ACS) has been annually collecting household socioeconomic data from a small, geographically-representative subset of American households. ACS is a part of the U.S. Census Bureau's Decennial Census Program and is designed to provide more current demographic, social, economic, and housing estimates during the decade between Census data collection and to compensate for the discontinuation in long-form data collection.

Each year, the ACS randomly samples around 3.5 million addresses (1% of total US addresses) and produces statistics that cover 1-year, 3-year, and 5-year periods for geographic areas in the United States and Puerto Rico. The 5-year estimates are available for distinct geographies including the nation, all 50 states, DC, Puerto Rico, counties, places, Census tracts, and Census block groups. The ACS Summary File data cover demographic, social, economic, and housing variables. The ACS 5-year estimates contain additional summary levels, such as census tracts and block groups that are not published in the ACS 1-year and 3-year estimates.

In the atlantaTractData_ACS2009to13 worksheet, the various socioeconomic variables are represented as either absolute or percentage values for each Census tract in the Atlanta Metro area. Detailed descriptions of the variables can be found in ACS manuals and in the associated macro-level assessment dissertation by Khoeini (Khoeini, 2014). It should be noted that even though 36,000 households are present in the spreadsheet, and 1.5 million plate observations in total, there are Census tracts for which zero vehicles were observed. Most of the observations come from the communities and census tracts along the monitored corridor (i.e. the catchment area, or commutershed). The modeling process is designed to use only the data from the tracts that produce trips on the monitored corridor.

The worksheet contains more than 75 demographic variables associated with each Census Tract. Most of the demographic variables are presented in percentages, related to head-of-household or the entire household. For example, gender, age, and race of head of household are included as columns. Household structure (e.g. married with children, single with children, etc.) work status, and education levels are also represented. Household income and income groups are expressed in $5k increments and larger upper bins. Commute travel time bins are also provided. More detailed descriptions of the data can be found in Khoeini (2014).

2.8 Census Data Linkage Worksheet (spatialJoinCensusTract)

The 39,000+ observed trips are represented by individual records (keyID) in the spatialJoinCensusTract worksheet. The worksheet contains the address information for each observation as well as the identification of the lane in which the plate was observed (lane column). Address observations from the previous geocoding process (household locations by latitude/longitude pair) were mapped as a point layer in ArcGIS and overlaid on census tract polygons. In this analytical step, the individual households were linked with their Census tract IDs using a spatial join function in ArcGIS. This ultimately allows the publicly-available Census data for their tract to be linked to the unique household ID for model development and application. Household-level data, rather than census tract data, can be linked to households via a similar process, when household data are available from surveys or commercial sources such as implemented by Khoeini (2014) and Sheikh (2015).

For the purpose of census tract-level modeling, each household (point in the GIS layer) is joined to the Census Tract ID to which it belongs. To create the spatial join, the 2010 Census Tract polygon shapefile (publicly available at the TIGER website) is used. The "Spatial Join" tool in ArcGIS is used to create a spatially-joined table between the household point layer and Census Tracts polygon layer. The spatialJoinCensusTract worksheet contains the data from the spatial join output table. For each household, the GEOID10 column represents the Census tract ID of the joined tract. Table 7 provides the description of the variables in this worksheet.

The next step in the process is to link the applicable census data for each tract in the Census tract worksheet described earlier (atlantaTractData_ACS2009to13) to each record for use in the model development work (discussed in the next report section).

Table 7: Census ID Spatial Join Worksheet (spatialJoinCensusTract)

Variable Name	Description	Notes
FID, objectID, join_ count, target_FID	Numeric tracking values for the join process	Not used in any analyses
keyID	Unique record identified assigned in video processing	Assigned automatically by data entry software
address1	Street address	Returned from the address match process
address2	Suite number	Returned from the address match process
Address	Concatenated address from address1 and address2	
householdID	householdID comes from a VLOOKUP equation using addressLookup in the uniqueAdressCoding table (calculates as "0" in the spreadsheet because addresses were deleted for privacy)	Replaces physical address in spreadsheet join functions so that the address can be deleted from retained files

Variable Name	Description	Notes
householdIDText	Text copy of calculated addressLookup	Used in all remaining join and count functions because calculated values (householdID) are no longer useable once address data are deleted
city	City	Returned from the address match process
state	State	Returned from the address match process
zip	9-digit zip code	Returned from the address match process (no-hyphen used)
latitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
longitude	Accuracy to only four or fewer decimal places is warranted	Output of the ArcGIS process (eight decimal places)
latitudeRand Trimmed	Latitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
longitudeRand Trimmed	Longitude value randomly perturbed and trimmed to 3 decimal places	Perturbs location by plus or minus 300 meters
latitudeRand TrimmedText	Text copy of calculated latitudeRandTrimmed	Carried for instructional purposes because latitude data are removed from working spreadsheet
longitudeRand TrimmedText	Text copy of calculated longitudeRandTrimmed	Carried for instructional purposes because longitude data are removed from working spreadsheet
lane	Managed Lane = 0 General Purpose Lanes = 1 - 5 (pulled from keyID)	Lane number is encoded in the keyID as the 5th digit by the software during data entry
STATEFP10	Census code for state	Linked in the spatial join
COUNTYFP10	Census code for county	Linked in the spatial join
TRACTCE10	Census code for tract	Linked in the spatial join
GEOID10	Census geographic ID	Linked in the spatial join and later used to link census tract data to each record

2.9 Census-Tract-Level Lane Use Analysis Worksheet (laneUseByTract)

The laneUseByTract worksheet contains the summary of pre-conversion observational data for each of the 1514 Census tracts (1514 rows). The first few columns of the laneUseByTract worksheet contain the pre-conversion observational data aggregated by census tract for use in final model application. The calculation aggregates total peak period trips observed for each Census tract. Note, however, that the working spreadsheet provided with this report only contains data that were collected by the research team at one data collection location during Am and PM sessions. The system is designed to use as much data as are collected in the field. As discussed earlier, the full data set was used by Khoeini (2014) to develop the final model parameters. The 39,000+ observed trips in the spreadsheet are tracked by lane in the spatialJoinCensusTract worksheet were aggregated by Census tract ID in the laneUseByTract worksheet by using the Excel COUNTIFS function. The total number of observations in the spatialJoinCensusTract worksheet is the count of all rows where the census tract ID matches the Census ID of the row in laneUseByTract. The total number of HOV observations in the spatialJoinCensusTract worksheet is the count of all rows where the Census IDs match the Census ID of the row in laneUseByTract, and the lane value is zero. The total number of general purpose lane observations in the spatialJoinCensusTract worksheet is the count of all rows where Census IDs match the Census ID of the row in laneUseByTract, and the lane value is greater than zero. After aggregation, each Census tract GEOID row in the laneUseByTract worksheet contains the total number of times that vehicles from households in that Census tract were observed using the HOV lane (column hov)and general purpose lanes (column gp) prior to the managed lane conversion. Lane use data aggregation could also be conducted in Access or any statistical software such as SPSS or SAS if desired, with the results returned to the spreadsheet for further analysis.

The laneUseByTract worksheet also contains the applied results from the final Census-tract-based modeling work that will be discussed in the report sections that follow. The model parameter XB, identified in the subsequent worksheets, is applied to the Census tract observational data to predict the expected number of observations in HOT lanes and general purpose lanes after the carpool lane is converted to a HOT lane (in columns hot_modeled and gp_modeled, respectively). The methodologies for developing the predictions are discussed in the report sections that follow and in more detail by Khoeini (2014).

Table 8: Corridor Use and Model-Predicted HOT Usage by Census Tract (laneUseByTract)

Variable Name	Description	Notes
tractID	Census Tract ID	
gp	General purpose lane observations during the baseline period	Total number of general purpose lane observations for the Census tract in spatialJoinCensusTract
hov	Carpool lane observations during the baseline period	Total number of carpool lane observations for the Census tract in spatialJoinCensusTract
Total	Sum of general purpose and carpool lane observations	Assumes that the sum of traffic is conserved before and after conversion
gp_modeled	Number of trips predicted in the general purpose lanes after conversion	Total baseline volume minus the predicted HOT volume below
hot_modeled	Number of trips predicted in the HOT lane after conversion	Total baseline volume multiplied by the predicted probability of using the HOT lane

Variable Name	Description	Notes
XB	Predicted value for the exponential function of the regression coefficients	Multiplied by the total baseline volume to predict HOT lane volume
hotLaneProbability	Probability of using the HOT lane	Based upon the XB value for each census tract
avgIncome	Average income in the tract	Found in atlantaTractDataACS2009to13 and used with dummy variables
avgHouseholdSize	Average household size in the tract	Found in atlantaTractDataACS2009to13 and used with dummy variables
avgCommute TravelTime	Average commute time for the tract	Found in atlantaTractDataACS2009to13 and used with dummy variables
Und18Yrs_per	Percentage of total Census tract population under 18 years of age	Found in atlantaTractDataACS2009to13 and used with dummy variables
A18to34Yrs_per	Percentage of total Census tract population between 18 and 34 years of age	Found in atlantaTractDataACS2009to13 and used with dummy variables
A35to64Yrs_per	Percentage of total Census tract population between 35 and 64 years of age	Found in atlantaTractDataACS2009to13 and used with dummy variables
Over65Yrs_per	Percentage of total Census tract population 65 years of age or older	Found in atlantaTractDataACS2009to13 and used with dummy variables
White_per	Percentage of households with white head of household	Found in atlantaTractDataACS2009to13 and used with dummy variables
Black_per	Percentage of households with black head of household	Found in atlantaTractDataACS2009to13 and used with dummy variables
Asian_per	Percentage of households with Asian head of household	Found in atlantaTractDataACS2009to13 and used with dummy variables
Otherrace_per	Percentage of households with other race head of household	Found in atlantaTractDataACS2009to13 and used with dummy variables
Hispanic_per	Percentage of households with Hispanic head of household	Found in atlantaTractDataACS2009to13 and used with dummy variables

2.10 Tract-Level Model Derivation (tractLevelModel)

The generalized linear model (GLM) is employed in development of the macro-modeling tool (Khoeini, 2014). In the analyses that follow, the dependent variable (or response variable) is predicted managed lane usage rate, between zero and one (π) for each Census tract, which prohibits the use of ordinary least square regression. The GLM is a flexible generalization of ordinary linear regression that allows for response variables to have other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function (Probit, in this case) and by allowing the magnitude of the variance of each measurement to be a function of its predicted value (MacCullagh and Nelder, 1989).

The predictor variables in the model are socioeconomic attributes at the Census tract-level. The Census tract-level socioeconomic data are expressed in either percentages (for example percent of white population in each tract) or average values (for example average annual household income in each tract). After extensive experience of data exploration and model generation (Khoeini, 2014), the research team found that dummy variables work better in socioeconomic related modeling to predict travel choice. Table 9 shows the values used to convert the original socioeconomic attributes for each Census tract into dummy variables for use in model development.

Table 9: Final Census-Tract Demographic Dummy Variables

near_dist_mile_bin	2	26mi<=
	1	8mi<= <26mi
	0	<8mi
age_under18_bin	1	25% <=
	0	<25%
ethnicity_bin	3	25% Asian<
	2	25% Hispanic<
	1	50% Black<
	0	50% White<
married_bin	1	50%<= married
	0	<50% married
education_bin	1	25%<= with BS or higher degree
	0	<25% with BS or higher degree
income_bin	2	21%<= have $125,000+ annual HH income
	1	The rest
	0	37%<= have <$30,000 annual income
travelmode_bin	1	5%<= use public transportation
	0	The rest
TT_bin	1	30min<= commute travel time
	0	<30min

The modeled response variable is managed lane usage rate between zero and one (π), for each Census tract. As mentioned above, the GLM approach generalizes linear regression by allowing the linear model to be related to the response variable via a link function, and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. For binary data (each license plate observation is either from the managed lane or general purpose lanes), the link function maps from $0<\pi_i<1$ to $\eta_i \in R$, and two link functions are commonly used: 1) Logit (Equation 1); or 2) Probit (Equation 2); where $\psi\ (\cdot)$ is the normal cumulative distribution function (MacCullagh and Nelder, 1989). In this study, Logit and Probit link functions were both employed, and compared using goodness of fit parameters; the Probit function was selected (Khoeini, 2014).

$$\boldsymbol{\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)}$$ Equation 1 (Logit link function)

$$\boldsymbol{\eta_i = \psi^{-1}\ (\pi_i)}$$ Equation 2 (Probit link function)

The model specification, response variables, standard error coefficients, and model performance parameters are presented in the tractLevelModel worksheet. Table 10 contains the final model Census-tract-level model variables and GLM regression coefficients with standard errors.

Table 10: Tract-Level Model Variables and Regression Coefficients

Model Parameter	B	Std. Error
(Intercept)	-1.455	.0137
[near_dist_mile_bin=2.00]	0.031	.0125
[near_dist_mile_bin=1.00]	0.045	.0061
[near_dist_mile_bin=.00]	0a	
[age_under18_bin=1.00]	0.061	.0065
[age_under18_bin=.00]	0a	
[ethnicity_bin=3.00]	-0.109	.0168
[ethnicity_bin=2.00]	-0.085	.0086
[ethnicity_bin=1.00]	-0.356	.0160
[ethnicity_bin=.00]	0a	
[married_bin=1.00]	0.042	.0083
[married_bin=.00]	0a	
[Education_bin=1.00]	0.176	.0078
[Education_bin=.00]	0a	
[Income_bin=2.00]	0.065	.0138
[Income_bin=1.00]	0.063	.0124
[Income_bin=.00]	0a	
[Travelmode_bin=1.00]	-0.086	.0134
[Travelmode_bin=.00]	0a	
[TT_bin=1.00]	0.091	.0055
[TT_bin=.00]	0a	
(Scale)	1b	

Because the model employs a link function, GLM regression coefficients (B) are not easy to interpret. Coefficients must be translated using the exponent function. When a logistic regression is calculated, the regression coefficient (B) is the estimated increase in the odds of the outcome per unit increase in the value of the exposure. The odds-ratio is the exponential of the coefficient, Exp(B), and is a measure of association between an exposure and an outcome. The odds-ratio represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. In other words, the exponential function of the regression coefficient (Exp(B)) is the odds-ratio associated with a one-unit increase in the exposure (Szumilas, 2010).

In practice, when there is a positive relationship between a predictor and an outcome (regression coefficient B > 0), the odds-ratio is greater than 1, and as the predictor B increases, the odds-ratio increases. The interpretation is that when the scale predictor increases by one unit, the probability that the outcome happens (vs. the other alternative happens) increases by a factor of the odds-ratio. Similarly, if there is a negative relationship between a predictor and an outcome (B < 0), the odds-ratio is less than 1, and as B decreases, the odds-ratio approaches zero. For non-scale predictors, the odds-ratio will be interpreted as a comparison. For example, if the predictor has two categories (male vs. female) the beta coefficient for one category (for example: male) will be set to zero and the odds-ratio for that category will be assumed to equal one, and the odds-ratio for the other category (in this case: female) is calculated respectively. If the calculated odds-ratio for female is more than 1, say 1.2, it implies that it is the outcome is 1.2x more probable when the predictor is female than if the predictor is male. The coefficients can be used to predict the probability that an outcome will occur, given the input values.

In logistic regression, instead of the standard R^2 parameter used in linear regression, other indicators must be used to assess model goodness of fit. The ρ^2 parameter measures how much the log likelihood (LL) of the fitted model improved compared to the null model (Equation 3). In logistic regression, deviance is analogous to the sum of squares in linear regression and is a measure of lack of fit to the data (Cohen and Cohen, 1975). Deviance (Equation 4) is calculated by comparing a given model with the saturated model; a model with a theoretically perfect fit. The Pseudo R^2 (Equation 5) shows the percentages of improvement in model fit (smaller deviance), by comparing the deviance of the fitted model to the deviance of the null model. As the model fit improves, the deviance should decrease, and the Pseudo R^2 moves closer to 1. Normally, the ρ^2 and $Pseudo\ R^2$ goodness of fit measures are very close, but not equal. The AIC (Equation 6) is another alternative for assessing goodness of fit, where k stands for number of parameters in the model and smaller AIC values indicate a better goodness of fit. Lastly, the Omni test examines the hypothesis of whether the built model is significantly better than the constant only model in predicting the response variables.

$$\boldsymbol{\rho^2} = \frac{\textbf{LL (Null Model)}-\textbf{LL(Fitted Model)}}{\textbf{LL (Fitted Model)}} \qquad \text{Equation 3}$$

$$\textbf{Deviance (Fitted Model)} = -\mathbf{2}\,\mathbf{ln}\,\frac{\textbf{LL (Fitted Model)}}{\textbf{LL (Saturated Model)}} \qquad \text{Equation 4}$$

$$\textbf{Pseudo R}^{2} = \frac{\textbf{Deviance (Null Model)} - \textbf{Deviance (Fitted Model)}}{\textbf{Deviance (Null Model)}} \qquad \text{Equation 5}$$

$$\textbf{AIC} = \textbf{2k} - \textbf{2 ln(LL)} \qquad \text{Equation 6}$$

The final model, associated parameters, and goodness of fit are presented in the tractLevelModel worksheet. In this case, all models are significant, with p-value less than 0.001 at 95% confidence (Khoeini, 2014).

2.11 Census-Tract-Level Model Calculations (tractLevelModelCalcs)

The tractLevelModelCalcs worksheet contains the values of the final model input dummy variables for each Census tract from the tractLevelModel worksheet. The final column in this worksheet (XB), presents the predicted value of the linear predictor for each tract. This same XB value is used in the laneUseByTract worksheet to predict HOT lane use. Proper transformation of the linear predictor value (column XB in the laneUseByTract worksheet) yields the estimated probability (column hotLaneProbability in the laneUseByTract worksheet) that a vehicle from the Census tract will choose to use the HOT lane after conversion (Khoeini, 2014). The calculated columns presented in the laneUseByTract worksheet (hot_modeled and gp_modeled) represent the predicted number of times that users from each Census tract will use the HOT lane and general purpose lanes, for the same morning and afternoon peak period durations in which pre-conversion data were collected.

2.12 Census-Tract-Level Model Outputs (tractLevel_SE_outputs)

The tractLevel_SE_outputs worksheet illustrates the socioeconomic attributes of the corridor commuters by lane type, before-and-after HOT lane conversion, in both tabular and graphic formats. Because the analyses employ aggregated tract-level data, the attributes of the four groups are expected to be close. The attribute that is most noticeably different is income, which is not surprising given the priced nature of the HOT lane. Figure 3 through Figure 5 illustrate the before and after household income, travel time to work, and race splits for the before and after conditions by lane type.

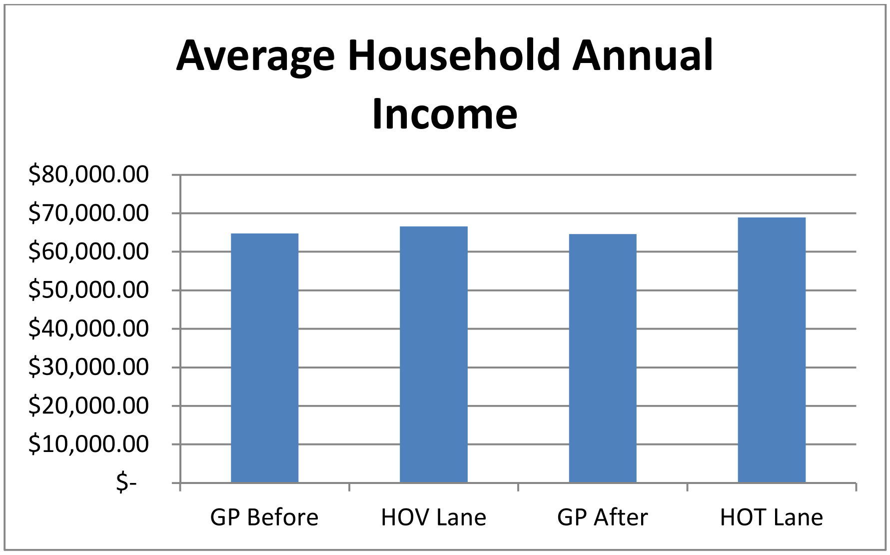

Figure 3: Average Before-and-After Income vs. Lane Use

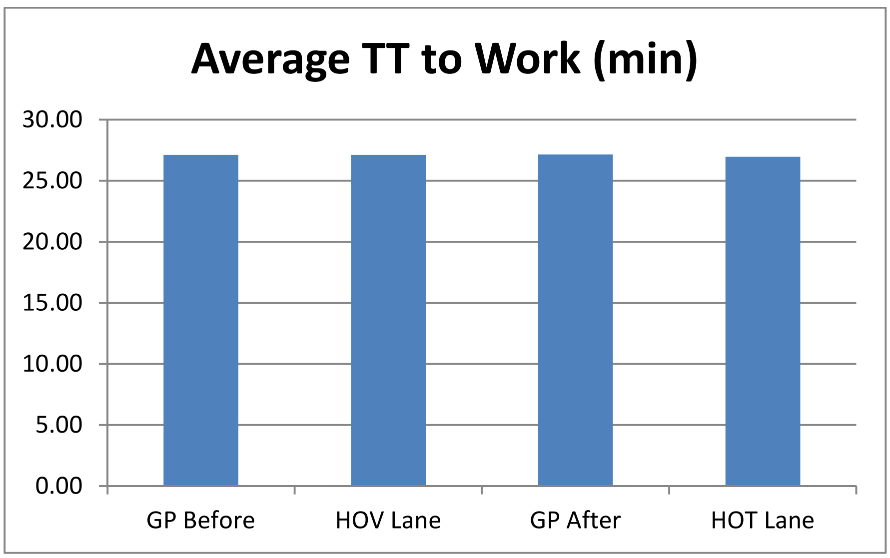

Figure 4: Average Before-and-After commute Travel Duration vs. Lane Use

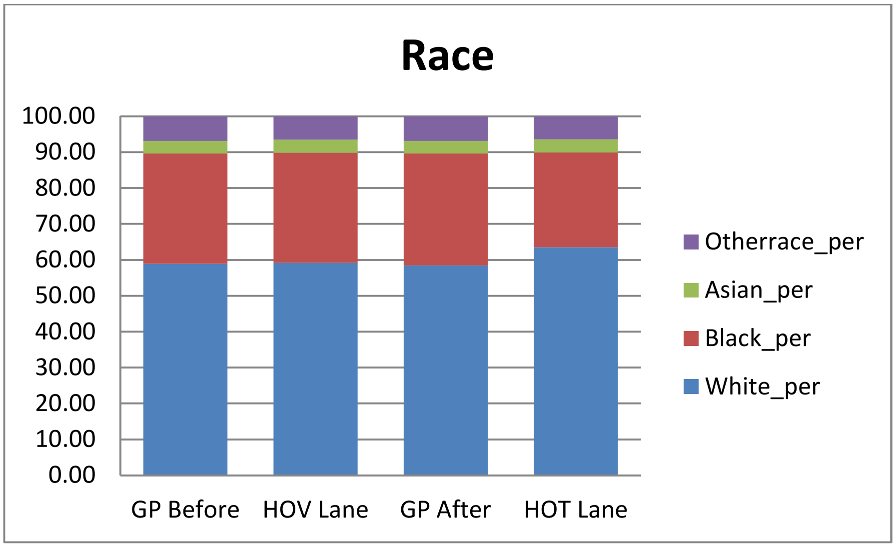

Figure 5: Average Before-and-After Race vs. Lane Use

2.13 Commutershed Analysis

The concept of the commutershed, or catchment area, is regularly employed by researchers to estimate facility travel demand (Horner and Groves, 2007). Commutersheds can be developed for any transportation facility such as highways, transit routes, and park and ride facilities to identify potential corridor users. As in other travel behavior studies, surveying has been the primary method of data collection and commutershed estimation. While the small sample size of typical surveys limits the use of direct GIS analysis, license plate data sets are so large that they are quite amenable to direct GIS analysis.

The modeling tool is based upon very large sampling of license plate data to define a detailed commutershed for the corridor under study. The comparison of the before- and after-conversion commutersheds helps researchers better understand whether the HOT lane significantly impacted the spatial distribution of corridor users. Because the HOT lane provides shorter and more reliable travel times, it is expected that more of the households that are farther away from the commutershed before conversion may begin to use HOT lanes on the corridor. It is important to note that detailed commutershed graphics can only be generated from household-level location data. Data aggregated to census tracts cannot generate such refined figures.

The ArcGIS Point Density Function was used to develop the corridor commutershed maps presented in the tractLevelComutershedOutput worksheet, and in Figure 6 and Figure 7. The Point Density Function calculates the density of point features around each output raster cell. Conceptually, a small neighborhood is defined around each raster cell center, and the number of household data points that fall within the neighborhood is totaled and divided by the area of the neighborhood. The population field is used to weight the observation data, where the weights differentiate between households with different numbers of observations. Accordingly, the weights for the households observed more frequently in the corridor are greater, relative to the households observed less frequently. Figure 6 shows the point density map (commutershed) for the general purpose lanes before and after conversion. Similarly, Figure 7 shows the commutershed for the managed lane before conversion (when it was a HOV carpool lane) and after conversion to the HOT lane.

2.13.1 Heat Maps

Heat maps enable researchers to visualize the changes in the corridor commutershed after the conversion. To create the heat maps, a linear transformation is performed on the raw density values so that output values can be mutually compared. The "Fuzzy Membership" function in ArcGIS transforms the input raster values to a 0 to 1 scale, indicating the strength of membership in a set (based on a fuzzification algorithm). In this case, a linear algorithm from 0 to 1 was used (Khoeini, 2014). A value of 1 indicates absolute membership and a value of 0 indicates absolute non-membership in the fuzzy set (ESRI, 2013).

To compare cell Fuzzy values before and after the conversion, the raster calculator in ArcGIS has been used (Khoeini, 2014). The raster calculator generates a new raster layer after applying the prescribed numerical function to the input layer cell values. In this case, the fuzzy values before conversion have been subtracted from the fuzzy values after conversion, and then multiplied by 100. The difference is multiplied by 100 to build a scale of impact between -100 and 100. Cells that experienced a value of change in fuzzy membership of 100 had the highest possible positive change. Cells with a value of change in fuzzy membership of -100 had the highest possible negative change. A zero value of change implies no change in corridor usage. The heat map for the observed data is contained in the tractLevelComutershedOutput worksheet, and presented below in Figure 8, where green indicates an increase (0 to 100 scale), and red indicates a decrease (0 to 100 scale) in set membership. In the case of the I-85 conversion, general purpose lane use appears to have increased directly along the corridor, while managed lane usage increased from areas upstream of the new HOT lane. A decrease is noted in both general purpose lane use and managed lane use southwest of the facility (both are red for this area), which may be related to route diversion to the Stone Mountain Freeway. Some traffic from north of the corridor may have diverted to SR13 or GA400 as well. However, monitoring/survey data are not available to confirm these hypotheses.

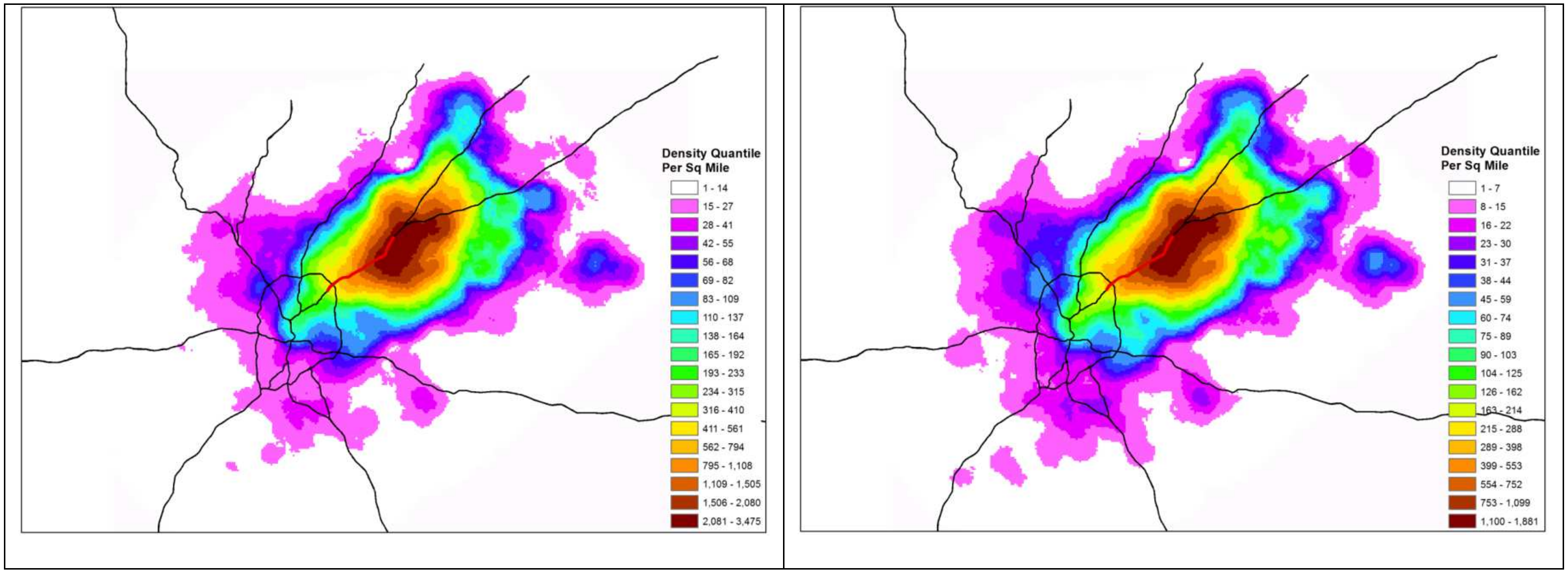

Figure 6: General Purpose Lane Commutershed Density Maps Before and After HOT Lane Conversion (Khoeini, 2014)

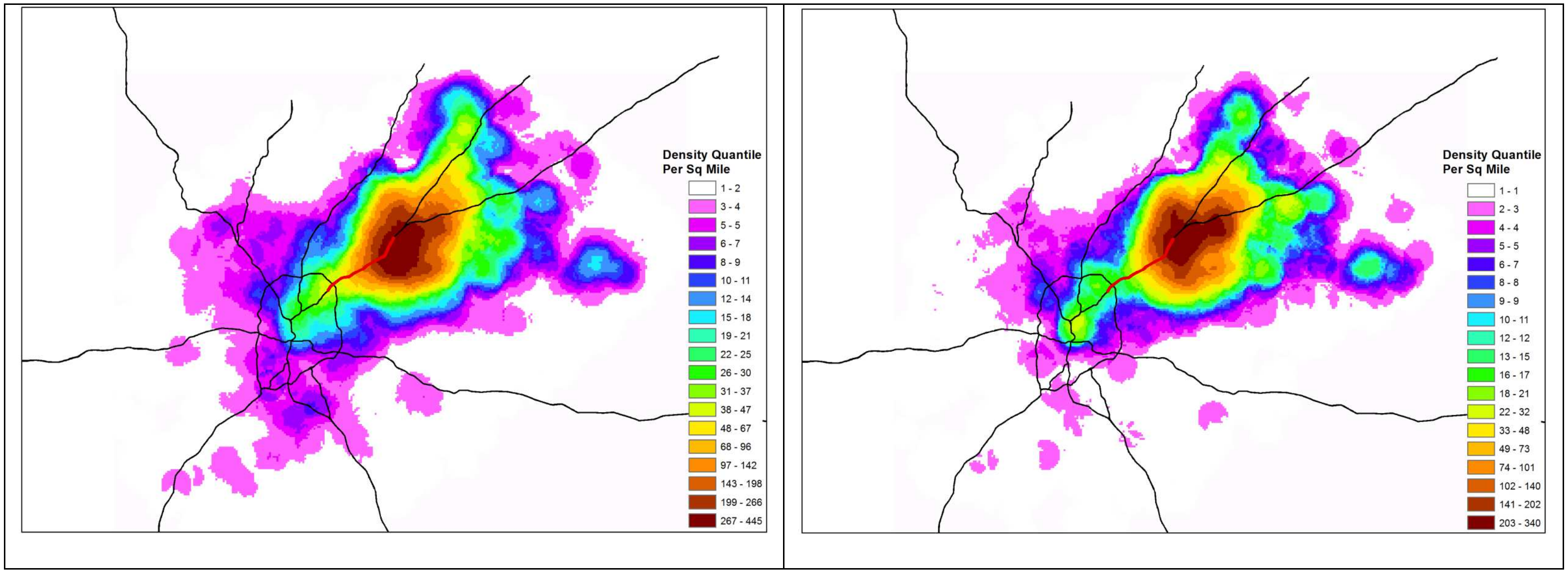

Figure 7: Managed Lane Commutershed Density Maps
Before (HOV) and After (HOT) HOT Lane Conversion (Khoeini, 2014)

2.13.2 Directional Distribution Ellipses

Another method of spatial distribution analysis is Directional Distribution Ellipse (see Khoeini, 2014). The ellipse is referred to as the standard deviation ellipse, because the method calculates the standard deviation of the x coordinates and y coordinates from the mean center to define the axes of the ellipse. For example, ellipses developed based on two standard distribution cover 95% of the observations in the map. A built-in tool in ArcGIS converts any set of point features to a directional distribution ellipse. Also, the ellipse allows researchers to see whether the distribution of features is elongated, and hence has a particular orientation. The ellipses for the general purpose lanes (before and after the conversion), HOV lane, and HOT lane are presented in the tractLevelComutershedOutput worksheet and in Figure 9 below. The ellipses employ two standard deviations and therefore include 95% of the observed commuters.

The directional distributional analysis is the best method to evaluate the overall directional displacement of the commutershed. However, it has the disadvantage of including the areas that are not actually part of the real commutershed in the analysis. Therefore, estimation error in area calculation is unavoidable. Fuzzy membership methods presented earlier help resolve this problem and produce a more accurate estimate of the change in the area of the commutershed.

2.14 Macro-Modeling Tool Caveats

It is important to note that not all the commuters observed actually live in the place that they have registered their vehicles (Nelson et al., 2008). Therefore, the registration address may differ from the actual residential address. For example, students and young professionals may register their vehicle at a parent's address to reduce insurance rates. Couples may live together in the corridor, while maintaining separate addresses. Governmental and commercial license plates also account for approximately 10% of the vehicles in our study. Users of these vehicles may use these vehicles for their daily commute trips. Leased vehicles are usually registered by the car owner household address instead of leasing company address, but not always.

Based on GIS spatial tools, 87% of the registered vehicle addresses did fall within the Atlanta metro area. Gwinnett County alone represented more than 66% of all the license plates. The next most highly-involved counties are Fulton and DeKalb County, the other two large counties adjacent to the corridor. After joining to the registration database and geocoding the addresses, 53% of the observed license plates in the field could be matched to a valid location in Atlanta metro area. However, a previous study by Granell (2002) indicated that perhaps 33% of the total vehicles are not registered in the same place that they begin their daily trip (especially for trips leaving apartment complexes). Further investigation of registered vs. garaged locations of vehicles are certainly warranted.

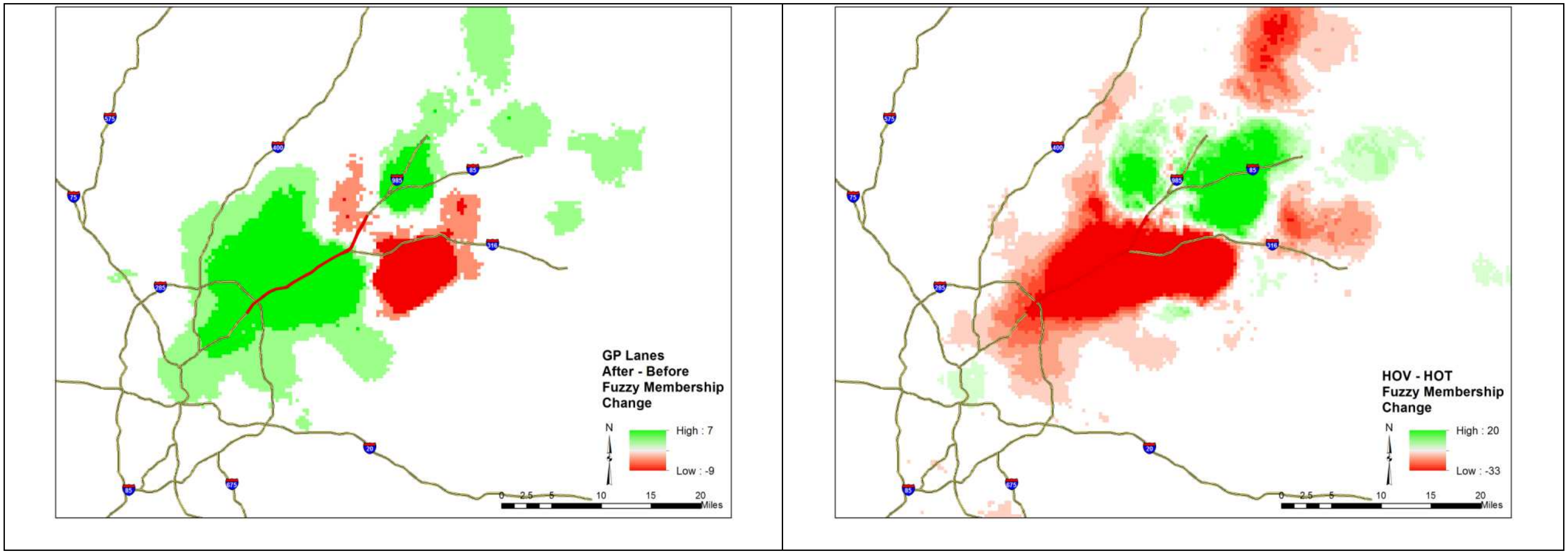

Figure 8: Heat Maps Before and After HOT Lane Conversion (Khoeini, 2014)

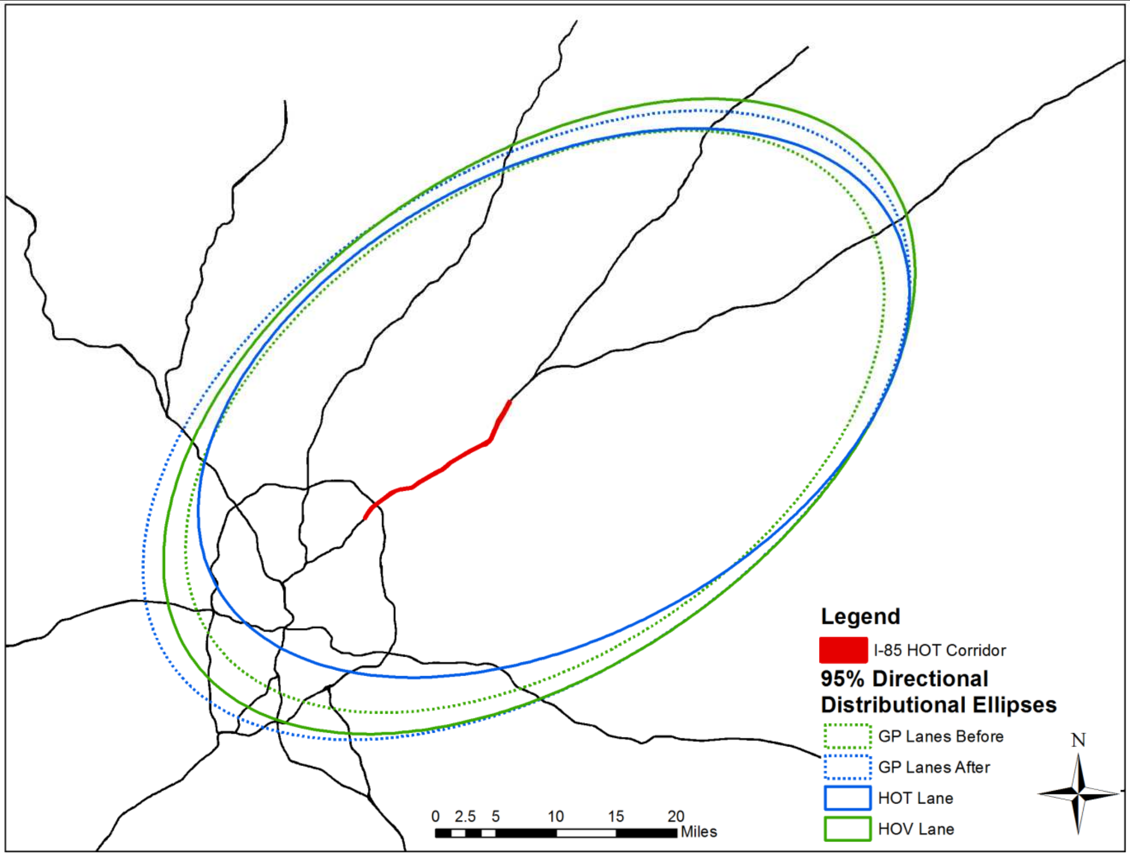

Figure 9: Directional Distributional Ellipses Before and After HOT Lane Conversion (Khoeini, 2014)

The Census-tract-based modeling tool operationalized in the spreadsheet is based upon the before-and-after data collected for the I-85 HOV-to-HOT conversion. Given the use of a large, revealed preference data set, the models derived and presented in the two dissertations (Khoeini, 2014; Sheikh, 2015) should, in theory, be transferrable to new sites and new projects in the metro area. However, there may be corridor-specific relationships affecting choice that have not been identified in these modeling efforts. As managed lane use data become available from new managed lane locations throughout the region, refined models should be developed from new observational data using the same general methods to reassess commuters response to managed lane addition at the macro-modeling level (overall response and managed lane usage) and to price and congestion at the micro-modeling level (day-to-day response of commuters to congestion and pricing).

3 Conclusions and Recommendations

This research presents a case study of the conversion of a high occupancy vehicle (HOV) carpool lane to a high-occupancy toll (HOT) lane, implemented in 15.5 miles of Atlanta I-85 on Oct, 1 2011. The focus of the research was to assess the impacts of socio-spatial characteristics of commuters on their travel behavior and choice to use or not use the HOT lane. The research team conducted the research using observational data for 1.5 million license plates, collected over two-year study period before and after HOV-to-HOT conversion and matched to household locations. The license plate basis of the study allowed the team to control research costs associated with conducting surveys, and to use revealed preference data in the analyses rather than stated preference data. The dissertation work by Khoeini (2014) was used to develop and implement a macro-level modeling tool from the before-and-after data that can be used to predict HOT lane usage as a function of publicly-available Census tract demographic data. The Census-tract-based model is operationalized in spreadsheet format for use in future corridor analysis. This report describes the methods used to develop the model, and the detailed content of the worksheets that comprise the spreadsheet modeling tool. Purchased marketing data, which include detailed household socioeconomic characteristics, can also be used to develop more refined models when tied to the license plate observation data (Khoeini, 2014; Sheikh, 2015).

At the general scale, this study enhances managed lanes' travel demand models with respect to users' characteristics and introduces a comprehensive modeling framework for socioeconomic analysis of managed lanes. The methods developed through this will inform future Traffic and Revenue Studies and help to better predict the socio-spatial characteristics of the target market. At the local level, the sponsored study also conducted a comprehensive socio-spatial analysis of Atlanta I-85 HOV to HOT conversion to investigate the impact on users' socio-economic attributes and on the commutershed (Sheikh, 2015). However, operationalizing the findings from the micro-level modeling in a spreadsheet format has proven too difficult to date.

The statistically-derived models presented in the two dissertations (Khoeini, 2014; Sheikh, 2015) should, in theory, be transferrable to new sites and new projects in the metro area. However, there may be corridor-specific relationships affecting choice that have not been identified in the modeling efforts. As managed lane use data become available from new managed lane locations in the region, refined models should be developed from new observational data to assess how commuters respond to changes in price and congestion level at the macro-modeling (overall response and managed lane usage) and micro-modeling levels (day-to-day response of commuters to congestion and pricing). These two types of models can fuel the assessment of aggregate- level response to implementation of managed lanes and then the pricing structures for efficient operation of these lanes. These types of models provide two significant benefits: 1) the models are based upon very large samples of observed license plates, rather than small samples surveys, and 2) the models will be based upon revealed preference data (actual observation) rather than stated preference (opinion) data. If the refined models prove accurate, demonstrated through applications in future corridors, license plate observational and model development methods should be implemented on a widespread basis. These models would likely be further improved by linking the observational data with stated preference survey data collected through regular household stated preference surveys of corridor users.

3.1 HOV-to-HOT Commutershed Macro-Modeling Tool Conclusions

Investigating the impact of users' socio-spatial characteristics and their HOT lane travel behavior can provide input to policy decisions concerning future managed lane investments and development (tolling and revenue studies), be used to improve travel demand models, and to assess and respond to socioeconomic concerns (Khoeini, 2014). In previous studies, traveler response toward managed lanes was often estimated using stated-preference or travel diary surveys, of small percent of the population, which are expensive, time-consuming, and labor-intensive. To minimize the cost and maximize the impact of this study, this research is based on one and a half million license plates, matched to household locations (using vehicle registration database), collected over two-year study period before and after HOV-to-HOT conversion. Identifying revealed relationships between socio-spatial characteristics and user response to the HOT lane conversion was the goal of this study.

Some additional conclusions derived from materials presented in this report and from the additional analyses in Khoeini's (2014) dissertation, include:

- Overall, the use of the HOT lane is lower than the baseline use of the HOV lane because the model is developed for the overall four-hour peak period. Users are much less likely to pay for use of a HOT lane during the shoulders of the peak, resulting in an overall lane use reduction. This is expected, and is not a negative consequence of implementation. Managed lanes are only needed under congested conditions.

- The sensitivity of the model to demographic data was lowest across days of week and higher across time of day and site of data collection within the corridor. Hence, there may be some local corridor-specific travel relationships that are omitted from the models.

- The Khoeini (2014) work also examined the application of vehicle value, which is less expensive and more convenient to collect, as a proxy for household income. The analysis demonstrated that the average vehicle value in the HOT lane was significantly higher, about $2,100 (23%), and the average vehicle model year was about one year newer, compared to the general purpose lanes. Furthermore, of 23% difference in vehicle value between HOT and GP lanes, 13% was associated with a difference (increase) in model year, and 10% was associated with changes in vehicles make/model rankings.

- Descriptive statistics were used to compare the socioeconomic differences between different groups of corridor users using both block group level and household-level data. To name a few major attributes, HOT lane user average household income is about 15% higher than users of adjacent GP lanes and HOV lane. In terms of vehicle ownership, HOV lane has the highest average vehicle ownership which accounts for 5% difference compared to adjacent GP lanes. Moreover, the original HOV lane represented 50% more Asian and 33% more Hispanic households, and 8% fewer White households compared to the adjacent general purpose lanes. On the other hand, HOT lane represents 8% more White households, and 28% fewer African-American, 33% fewer Hispanic, and 12% fewer Asian households. In terms of home ownership, the HOT lane has 44% fewer renters compared to the adjacent general purpose lanes.

- GIS raster analysis methods were used to visualize and quantify the impact of the HOV-to-HOT conversion on corridor commutershed. The HOT lane commutershed is smaller than the HOV lane commutershed and the general purpose lane commutershed expanded after the conversion (perhaps in part resulting from longer distance commuters switching from the HOV lane to general purpose lane, as well as the addition of new long-distance commuters). However, the amount of commutershed expansion by general purpose lanes dominates the amount of retraction produced by HOT lane, causing an overall expansion in the corridor commutershed.

- In the detailed dissertation, Khoeini (2014) also developed six models at two analytical levels: primary aggregated (block-group-level) and advanced disaggregated (household-level). The advantages of the block group level models are lower cost, and publically available socioeconomic data, and the disadvantage is lower predictive power. The advantage of household-level models is significantly higher predictive power, but at a cost of acquiring household-level marketing data. Household-level models accuracy increases as the sample size and resulting cost of data increase.

- Generally, the impacts of income, home ownership, and ethnicity (Hispanic/Asian/African-American) are the highest in these models. The fact that income and home ownership are significant is intuitive, considering the pricing scheme of the conversion. However, the impact of ethnicity after controlling for income is interesting and has not been identified in any previous studies. One potential reason might be the fact that some ethnic groups may be more hesitant to acquire transponders.

- The HOT usage model has substantially better goodness of fit compared to a similar HOV usage model in Khoeini (2014). Significant additional research appears warranted to assess the relationships between demographic characteristics and HOV formation and retention.

- The socioeconomic variables associated with household usage of the HOT lane corridor were derived from license plate observations linked to demographic data sources. For the models presented in this report, the socioeconomic data were retrieved from Census-tract-level level American Community Survey data. However, the Khoeini (2014) and Sheikh (2015) dissertations also explore the use of household-level marketing data. Marketing data provide very detailed household and individual level attributes with significant low amount of cost (10¢ per household), compared to travel surveys which cost about $200/household. Marketing data, used in conjunction with associated trip data, have been introduced as an alternative for conducting travel behavior studies. The research team believes that the model enhancements provided by the use of household level data are worth the investment.

3.2 HOV-to-HOT Commutershed Micro-Modeling Tool Conclusions

The overall research effort undertaken in this project employed revealed-preference data of I-85 Express Lane users to investigate the monetary value users ascribed to their time on the corridor, by examining the toll amounts they paid and the resulting time that they saved (Sheikh, 2015). The dissertation analyses examined the resulting value of travel time savings distributions across income segments and among trips of different lengths. As reported in the dissertation, the differences in these distributions among lower, medium, and higher income households were marginal at best. Differences among the mean, median, and other quartile values were on the order of cents, rather than dollars. The results did not indicate that higher income households had the highest value of travel time savings results, as may have been expected. The ranking of value of travel time savings by income segment was not consistent across time frames or direction of travel (morning vs afternoon commute travel). The trip length investigation revealed more distinct differences between users who traverse the entire length of the corridor and those that take partial trips; in that case, the southbound and northbound differences were also more pronounced. An important consideration in interpreting the results from Sheikh's (2015) dissertation is that they represent the Express Lane users only; that is, only users who chose to make paid trips in the HOT lanes. Non-users, and general purpose lane trips by HOT users, were excluded from this analysis.

The modeling work performed by Sheikh (2015) provided a number of insights into toll lane use and the determinants of lane choice decisions. The initial analysis involved binary logit mode choice models which were estimated across different income segments and household clusters to examine differences in decision making between low, medium, and higher income households and between demographically similar households. The results indicated that the income-segmented models yielded different results than the pooled model at the 95% confidence level, but the parameters were largely consistent across the three segments. The clustered households exhibited more variation in their responses, particularly for the older and larger households. For the year studied, rates of HOT lane use were fairly consistent across the three income groups for which data were available, differing by a maximum of 3.9%. Disaggregate elasticity values revealed low sensitivities to nearly all of the explanatory parameters with the exception of the problematic trip distance variable, and income among the higher income users. These elasticity values illustrated varying responses to household income and education, for example, across the segmented and clustered households.

The extensions of the preliminary analysis revealed the benefits of further segmenting households by income to illustrate the variety of behavior within the higher income households. This segmentation indicated that the three-segment strategy disguised substantial behavioral differences among the highest income households on the I-85 corridor. The determinants of lane choice decision-making in the morning peak had notable differences from the determinants of the afternoon peak, particularly with regards to toll rate sensitivity and the impact of the total corridor segments traversed. Afternoon peak models had better goodness of fit metrics overall, though the pseudo-R^2 measures for both time frames were under 0.40 in all but one of the cases. This indicates that the there are many other factors in play in lane choice decision making. Collection of travel survey and stated-preference data from these corridor users may play an important role in improving the models. The operational characteristics included in the lane choice models, including average lane speeds and transponder counts, yielded similar responses across the income segments under examination. It should be noted that the users examined in this study all had registered for Peach Pass transponders, and as such represent a self-selecting sample of corridor users. The similarities in decision-making factors across the different models and income groups may result from this effect. This issue could begin be addressed by providing transponders automatically and without cost to those users without Peach Pass accounts, though the sample would still be restricted to those users who choose to use them in their vehicles.

A mixed logit framework improved the modeling results by addressing the issue of serial correlation and by estimating the toll amount and household income coefficients as random rather than fixed parameters (Sheikh, 2015). The toll amount coefficients, for example, were more appropriately modeled as normal distributions that encapsulate both positive and negative values to reflect both the ‘signaling’ and demand-reducing effects of toll rates. Further segmenting the households showed that lane choice determinants varied more within the ‘higher’ income segment than across the original three-segment structure. In particular, the five-segment models illustrated lower elasticities with regard to corridor segment counts and toll levels for the highest-income households in the sample, as well.

3.3 A Managed Lane Socio-Spatial Modeling Framework

The objective of this section is to summarize the methods developed and applied in the previous sections and to suggest a preliminary analytical framework that could be applied to future assessments of similar managed lane and tolling projects. The data collection methodologies, analyses, and results were illustrated in detail in the previous sections. Based upon these results, this section proposes a step-wise framework for future socioeconomic analysis of managed lane facilities. Similarly, traffic and revenue studies could use the resulting analytical framework to forecast the characteristics and probable travel behavior of target market in response to pricing.

The first step is to collect travel data of the current conditions to establish baseline conditions and provide data for use in forecasting future activity levels. Travel data specifically refers to elements that identify users of the corridor and their current travel behavior with respect to operations on the corridor. For example, in this study, license plate data were used to identify the households that were currently using the corridor before HOT conversion, as well as the associated frequency of use along HOV and general purpose lanes. Considering the available budget and desired accuracy of any future study, different methods of data collection and different amount of data could be collected.

The collection and analysis of license plate data, using similar to the methods employed in this study, are recommended by the research team. Based upon the field experience of the research team, the net cost for collecting and processing a completed license plate record (i.e., a plate that yields matched records in the registration database with fewer than eight registered vehicles per address, and the registration address is in reasonable proximity to the corridor) is approximately 10¢ per plate. On average, one two-hour session of data collection on a six lane corridor (12 lane-hours) produces 7,719 complete license plate records at peak hour and costs less than $800 (including the cost of manual license plate extraction). Accordingly, one lane hour collected video produces 643 correct license plates costs less than $100. More advanced methodologies such as Automatic License Plate Readers (ALPR), RFID tag readers, and cell phone data, can increase the amount of data collected and decrease the cost of data collection once equipment is capitalized. Automated methods should reduce labor costs and improve the future efficiency of future data collection efforts.

Using the collected travel data, the next step is to establish the corridor commutershed. Accordingly, the commutershed could be developed at household-level or at the block-group-level. The household-level analysis needs enough data to provide reliable frequencies of corridor use per household and to identify frequent users (top 5% frequent license plates in this study). In this study, each household was observed an average of five times during the 1860 lane-hours data collection. This large amount of data (1,196,433 complete license plates matched to 241,466 households) enabled this research to establish target market groups and support household-level models. Although large datasets enhance the accuracy of the results of future studies, collecting such a large amount of data is not necessarily required for all future studies.

Supplemental studies can assess the amount of data that are required to develop reliable models, so that field data efforts can be minimized to control field data collection costs. The numbers of collected households and (correct) license plates as a function of amount of data collection for this study are illustrated in Figure 10. These functions could be used to plan future data collection efforts. Because there are so many regular users on a commuter corridor, plate data collection yields diminishing returns with respect to identification of new households. For example, if we assume that 1000 lane-hours yield 198,000 households, analysts can capture about 60% of these households in about 500 hours.

Figure 11 illustrates the cost of field data collection at 10¢ per completed license plate. For the example above, the cost of collecting about 198,000 households (or 653,000 license plates) is approximately $66,000. In the context of a $100 million project, this is an insignificant expenditure. However, this is the cost for data collection along only one corridor and studies would need to be conducted throughout the region. Again, however, these costs would be insignificant relative to the projected $16.1 billion cost for the complete managed lane system.

Figure 12 illustrates the relationship between average observation frequency per household as a function of amount of data collected (lane hours) with the blue line. The estimated power functions can be used by future researchers to estimate average observation frequency per household. Whereas the average observation frequency is estimated across all the license plates, the minimum frequency of the frequent corridor commuters (top 5% frequent license plates) has also been illustrated with the red line. This latter variable is important for identifying the frequent users for the application of developed models. The slope of variation for average observation frequency is relatively flatter than the minimum frequency of top 5% users. For example, by collecting 252 lane hours' worth of data (or 21 two-hours session for a six lane highway, which corresponds to one quarter of data collection conducted for this study), the average frequency is 2.3

and the minimum frequency of top corridor users is seven, which is large enough for applying all the developed models in this study.

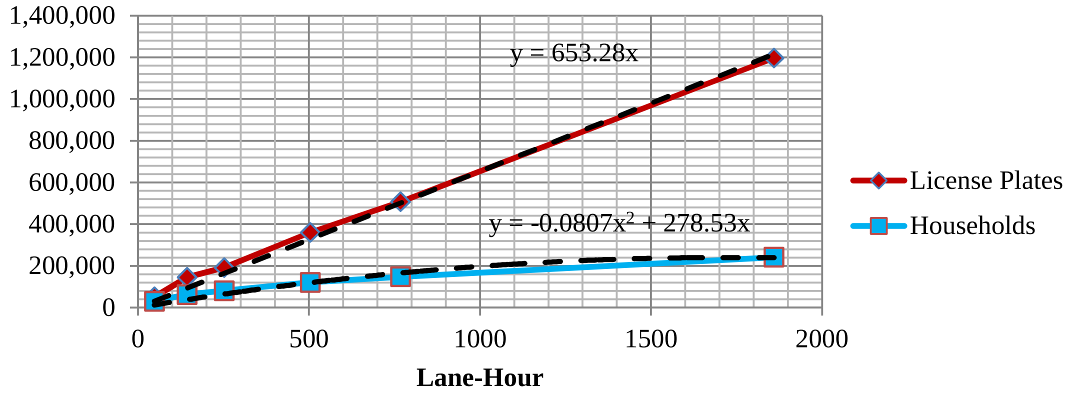

Figure 10: Count of License Plates and Households as a Function of Lane-Hours of License Plate Data Collection

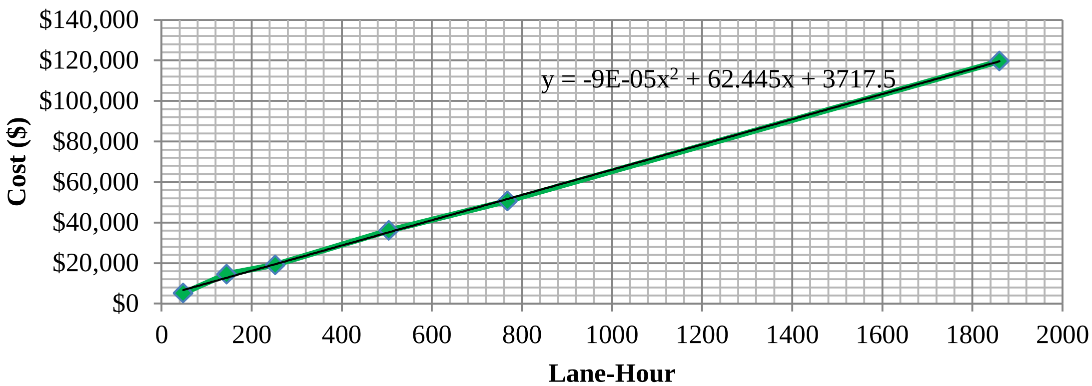

Figure 11: Data Collection and Processing Cost Estimate as a Function of Lane-Hours of License Plate Data Collection

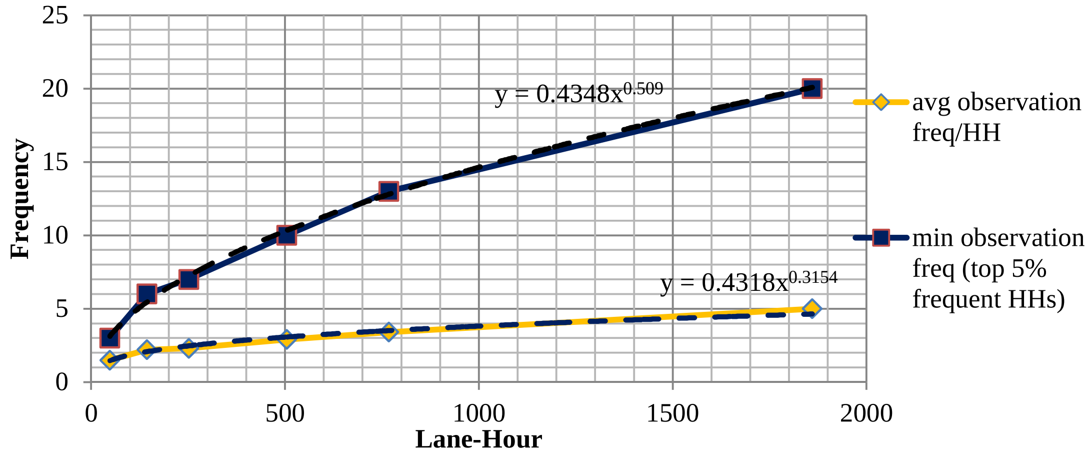

Figure 12: License Plate Data Frequency of Observation per Household as a Function of Lane-Hours of License Plate Data Collection

The sensitivity analysis of Khoeini's (2014) work, show that license plate data along a corridor are more sensitive to data collection location than to time (a.m. vs. p.m.), and are more sensitive to data collection time than to weekday. Thus, it is better to spread the data collection locations across the corridor and make sure to collect data in both the morning and afternoon peak periods.

The third step is to acquire socioeconomic and demographic data. For block-group-level analysis, the most recent publicly available American Community Survey data should be used. The block groups that need to be incorporated in the analysis are those that intersect with the developed commutershed. However, the modeling results based upon disaggregate household-level data are preferable (Khoeini, 2014). For household-level models, the licensing cost for the full set of marketing data used in this study was approximately 10¢ per household. Considering the multi-million cost of the entire project, the entire cost of license-plate data collection and socioeconomic data acquisition is perfectly reasonable.

Socioeconomic data can also be collected using other methods such as surveys and cell phone apps. However, the sample size of traditional surveys is very small compared to the large number of collected households and not strongly recommended for project-level analysis. However, the application of more innovative forms of surveys such as cellphone apps as part of a before-after panel survey may be justifiable. More specifically, if the travel data have been collected using cell phone data, the collection of socioeconomic data with cell phone apps would make the data collection process less expensive and more efficient.

In general, any big-data collection methodology, which can produce both travel and/or socioeconomic data, could be applied to the models developed in this study. Specifically, the proposed methodology in this study is the concurrent application of license plate data, registration data, and marketing data.

3.4 Additional Research

The research presented in this report indicates that macro-level and micro-level models can be used to predict corridor response. Although these models are developed from very large data sets of revealed preference observational data, these models cold still be significantly improved. The research team proposes that the following research efforts be conducted to assess the potential impacts of the pending $16.1 billion in managed lane investments slated for implementation in the Atlanta Metropolitan region:

- Targeted sub-regional household stated-preference surveys should be conducted along corridors of interest, in parallel with license plate data collection and analysis. Resulting data would enhance models designed to predict how users are likely to respond to the implementation of new managed lanes (or alternative managed lane strategies).

- Given the relatively low cast of marketing data (10¢ per household), which provide very detailed household and household-level socioeconomic attributes, additional efforts should focus on demonstrating the accuracy of these data and integrating these disaggregate data into managed lane corridor assessments. A larger research effort that combined purchased marketing data with stated preference survey data collection, and with follow-up focus groups for a subset of participants managed lane, would verify the accuracy and reliability of the household-level data and would provide new stated preference and revealed preference data that could be used to enhance model development.

- The acquisition of transponders, which are required for use of the HOT facility, may differ significantly across income and ethnic groups, which necessarily affects the model outcomes predicted in the dissertation work by Khoeini (2014) and Sheikh (2015). The research team believes that a separate choice model should be developed to predict the establishment of Express Lane accounts and acquisition of transponders for managed lane participation as a function of spatial and demographic variables. These new models could be used to inform strategies designed to increase transponder adoption rates and facility participation across demographic groups.

- Significant additional research appears warranted to assess the relationships between demographic characteristics and HOV formation and retention. If carpools are being considered as a viable strategy for managing future transportation demand, much more information on the causal variables affecting carpool formation and retention is needed.

- Future studies should be implemented to assess the sensitivity of model development to the amount of license plate data collected and processed. The goal of this supplemental research would be to predict the amount of data needed to obtain reliable results, so that field teams do not collect more data than are necessary to develop reliable models (controlling field data collection costs). Further assessment of the sensitivity of license plate data collection with respect to time, day, and location of data collection would also help in model development and to better understand potential uncertainty impacts of derived models.

4 References

AirSage (2013). http://www.ourenergypolicy.org/wp-content/uploads/2014/11/The_Future_of_Transportation_Studies-1.pdf

Atlanta Regional Commission (2011). Atlanta Regional Travel Survey Final Report. Atlanta, GA. 2011

Börjesson, M. (2008). Joint RP–SP Data in a Mixed Logit Analysis of Trip Timing Decisions. Transportation Research Part E: Logistics and Transportation Review, 44(6), 1025-1038.

Burris, M., S. Nelson, P. Kelly, P. Gupta, and Y. Cho (2012). Willingness to Pay for High-Occupancy Toll Lanes. *Transportation Research Record: Journal of the Transportation Research Board* 2297, 47–55.

Burris, M., K. Sadabadi, S. Mattingly, M. Mahlawat, J. Li, I. Rasmidatta, and A. Saroosh (2007). Reaction to the Managed Lane Concept by Various Groups of Travelers. *Transportation Research Record: Journal of the Transportation Research Board* 1996, 74–82.

Cohen, J., and P. Cohen (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.

Colberg, K. (2013). "Investigating the Ability of Automated License Plate Recognition Camera systems to Measure Travel Times in Work Zones." Master's Thesis, Georgia Institute of Technology.

D'Ambrosio, K. (2011). Methodology for Collecting Vehicle Occupancy Data on Multi-Lane Interstate Highways: A GA 400 Case Study. Master's Thesis. Georgia Institute of Technology, School of Civil and Environmental Engineering. August 2011.

Devarasetty, P.C., Burris, Mark, Douglass Shaw, W. (2012). The value of Travel Time and Reliability-evidence from a Stated-Preference Survey and Actual Usage. *Transportation Research Part A: Policy and Practice* 46, 1227–1240.

Dill, J., and A. Weinstein (2007). How to Pfor Transportation? A Survey of Public Preferences in California. *Transport Policy* 14, 346–356.

Doherty, S. (2009). Emerging Methods and Technologies for Tracking Physical Activity in the Built Environment. *Transport Survey Methods*: *Keeping up with a Changing World*, 153-190.

Douma, F., J. Zmud, and T. Patterson (2005). Pricing Comes to Minnesota: Attitudinal Evaluation of I-394 HOT Lane Project. *Transportation Research Board Annual Meeting*.

ELSAG North America (2013). "ALPR Products and Solutions". Available from: http://www.elsag.com/alpr_applications.htm.

Golob, T., D. Bunch, and D. Brownstone (1997). A Vehicle Use Forecasting Model Based on Revealed and Stated Vehicle Type Choice and Utilisation Data. *Journal of Transport Economics and Policy* , 69-92.

Goodall, N., and B.L. Smith (2010). What Drives Decisions of Single-Occupant Travelers in High-Occupancy Vehicle Lanes?. Transportation Research Record: Journal of the Transportation Research Board, 2178(1), 156-161.

Guensler, R., V. Elango, A. Guin, M. Hunter, J. Laval, S. Araque, S. Box, K. Colberg, F. Castrillon, K. D'Ambrosio, D. Duarte, K. Kamiya, S. Khoeini, E. Palinginis, L. Peesapati, C. Rome, A. Sheikh, K. Smith, C. Toth, T. Vo, and S. Zinner (2013). "Atlanta I-85 HOV-to-HOT Conversion: Analysis of Vehicle and Person Throughput." Prepared for the Georgia Department of Transportation, Atlanta, GA. Georgia Institute of Technology. Atlanta, GA. October 2013.

Hess, S., M. Bierlaire, and J. Polak (2005). Estimation of Value of Travel-time Savings using Mixed Logit Models. Transportation Research Part A: Policy and Practice, 39(2), 221-236.

Hultgren, L., and K. Kawada (1999). San Diego's Interstate 15 High-Occupancy/Toll Lane Facility Using Value Pricing. *ITE Journal* 69.6.

Khoeini, S. (2014). Modeling Framework for Socio-economic Analysis of Managed Lanes. Dissertation. Georgia Institute of Technology. Atlanta, GA. 2014.

Khoeini, S., M.O. Rodgers, V. Elango, and R. Guensler (2012). Sensitivity of Commuters' Demographic Characteristics to License Plate Data Collection Specifications: Case Study of I-85 High-Occupancy Vehicle to High-Occupancy Toll Lanes Conversion in Atlanta, Georgia. *Transportation Research Record: Journal of the Transportation Research Board* 2308, 37–46.

King, D., M. Manville, and D. Shoup (2007). The Political Calculus of Congestion Pricing. *Transport Policy* 14, 111–123.

Kressner, J.D., and L.A. Garrow (2012). Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia, Airport. *Transportation Research Record: Journal of the Transportation Research Board*, 2266, 20–30.

Kriger, D S., S. Shiu, and S. Naylor (2006). *Estimating Toll Road Demand and Revenue* (Vol. 364). Transportation Research Board.

Lam, T.C., and K.A. Small (2001). The value of Time and Reliability: Measurement from a Value Pricing Experiment. Transportation Research Part E: Logistics and Transportation Review, 37(2), 231-251.

Li, J. (2007). Potential Users' Perspectives on Managed Lanes Insights From a Focus Group Study. *Public Works Management & Policy* 12, 416–430.

Liu, H.X., X. He, and W. Recker (2007). Estimation of the Time-dependency of Values of Travel Time and its Reliability from Loop Detector Data. Transportation Research Part B: Methodological, 41(4), 448-461.

Lorenzo, G.D., J. Reades, C. Francesco, and C. Ratti (2012). Predicting Personal Mobility with Individual and Group Travel Histories. *Environment and Planning-Part B* 39.5, 838.

MacCullagh, P., and J. Nelder (1989). Generalized Linear Models.

Munnich, L., and J. Loveland (2005). Value Pricing and Public Outreach: Minnesota's lessons learned. *Transportation Research Record: Journal of the Transportation Research Board* 1932.1 (2005): 164-168.

Ross, C.L., R. Guensler, J. Barringer, A. Danner, M. Allen, E. Barrella, J. Doyle, J.I. Nelson, and L. Zuyeva (2008). Congestion Pricing Response: Study for Potential Implementation in the Metropolitan Atlanta Area. Downloaded from: "http://www.cqgrd.gatech.edu/sites/ files/cqgrd/files/gdot_congestion_pricing_study_full.pdf".

Schönfelder, S., H. Li, R. Guensler, J. Ogle, K.W. Axhausen (2006). "Analysis of Commute Atlanta Instrumented Vehicle GPS Data: Destination Choice Behavior and Activity Spaces." Published in the CD-ROM Proceedings of the 83rd Annual Meeting of the Transportation Research Board. Washington, DC. January, 2006.Sheikh, A. (2015). Consumer Response to Road Pricing: Operational and Demographic Effects. Dissertation. Georgia Institute of Technology. Atlanta, GA. 2015.

Sheikh, A. (2015). Consumer Response to Road Pricing: Operational and Demographic Effects. Dissertation. Georgia Institute of Technology. Atlanta, GA. 2014.

Small, K.A., C. Winston, and J. Yan (2005). Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability. Econometrica, 73(4), 1367-1382.

Stopher, P.R., and H.M. Metcalf (1996). *Methods for Household Travel Surveys* (Vol. 236). Transportation Research Board.

Suh, W., R. Guensler, M. Hunter, A. Guin, J. Anderson, K. Colberg, and S. Zinner (2013). "Work Zone Technology Testbed." Final Report. Prepared for the Georgia

Department of Transportation, Atlanta, GA. GDOT Research Project RP 11-15. Georgia Institute of Technology. Atlanta, GA. October 2013.

Sullivan, E. (1998). Evaluating The Impacts of the SR91 Variable-Toll Express Lane Facility. Sacramento, CA.

Supernak, J., and J. Golob (2002). San Diego's Interstate 15 Congestion Pricing Project: Attitudinal, Behavioral, and Institutional Issues. *Transportation Research Record: Journal of the Transportation Research Board* 1812, 78–86.

Szumilas, M. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 227–9.

Wang, P., T. Hunter, A.M. Bayen, K. Schechtner, and M.C. González. (2012). Understanding Road Usage Patterns in Urban Areas. *Scientific Reports* 2, 1001.

Xi, Yanzhi (2010). Effective GPS-Based Sample Size for Urban Travel Behavior Studies. Dissertation. Georgia Institute of Technology. Atlanta, GA. 2010.