Title

Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis

Authors

Lu, Hongyu

Rodgers, Michael O.

Guensler, Randall

Publication Date

2024-09-01

DOI

10.7922/G2G73C16

Data Availability

The data associated with this publication are available at:
https://doi.org/10.5281/zenodo.13864619

Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis

September 2024

A Research Report from the National Center for Sustainable Transportation

Hongyu Lu, Georgia Institute of Technology

Dr. Michael O. Rodgers, Georgia Institute of Technology

Dr. Randall Guensler, Georgia Institute of Technology

National Center for Sustainable Transportation
A USDOT University Transportation Center

Georgia Tech School of Civil and Environmental Engineering
College of Engineering

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. NCST-GT-RR-24-26	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis		**5. Report Date** September 2024 **6. Performing Organization Code** N/A
7. Author(s) Hongyu Lu, https://orcid.org/0000-0002-0170-7169 Michael O. Rodgers, Ph.D., https://orcid.org/0000-0001-6608-9333 Randall Guensler, Ph.D., https://orcid.org/0000-0003-2204-7427		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address Georgia Institute of Technology School of Civil and Environmental Engineering 790 Atlantic Dr., Atlanta, GA 30332		**10. Work Unit No.** N/A **11. Contract or Grant No.** USDOT Grant 69A3551747114
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Research Report (October 2020 – April 2023) **14. Sponsoring Agency Code** USDOT OST-R

15. Supplementary Notes
DOI: https://doi.org/10.7922/G2G73C16
Dataset: https://doi.org/10.5281/zenodo.13864619

16. Abstract
In this study, a modeling framework was developed to generate high-resolution synthetic fleets, for use with synthetic household modeling in activity-based travel models, by integrating various data sources. The synthetic households were generated by pairing household locations and demographic attributes, and synthetic fleets were assigned to the households so that travel demand model outputs would have vehicles associated with each model-predicted tour for energy and emissions analysis. The CO emissions were modeled for each vehicle and each link traversed by vehicles as predicted by the travel demand model, and the results of the synthetic fleet (by employing Monte Carlo simulations and Bootstrap techniques) were compared with those from standard regional and sub-regional fleet configurations. The results demonstrated that using a traditional sub-regional fleet scenario produced 30% higher predicted emissions than when the synthetic fleet was employed with predicted vehicle trips, and that using a regional average fleet (applied throughout the region) produced emissions that were more than 50% higher than synthetic fleet emissions. Lowest household emissions were associated with low-income and non-working households, and highest emissions were associated with moderate-income households and one-person high-income household groups. The results presented in the research are not necessarily conclusive, because the licensed vehicle data procured for Atlanta appear to be biased toward older vehicles. Model year penetration rates are accounted for in these analyses, but the authors believe that the variability in the registration mix for newer vehicles is likely underestimated in the data procured for these analyses. The authors conclude that access to statewide registration data will be required to remove potential biases that exist in licensed private data sets. Nevertheless, the study does demonstrate that properly pairing vehicle model years with the most active households (and their daily trips) significantly impacts energy and emissions analysis.

17. Key Words Synthetic Household, Synthetic Fleet, Emission Modeling, Travel Demand Model		**18. Distribution Statement** No restrictions.	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 107	**22. Price** N/A

About the National Center for Sustainable Transportation

The National Center for Sustainable Transportation is a consortium of leading universities committed to advancing an environmentally sustainable transportation system through cutting-edge research, direct policy engagement, and education of our future leaders. Consortium members include: the University of California, Davis; California State University, Long Beach; Georgia Institute of Technology; Texas Southern University; the University of California, Riverside; the University of Southern California; and the University of Vermont. More information can be found at: ncst.ucdavis.edu.

Disclaimer

Acknowledgments

This study was funded, partially or entirely, by a grant from the National Center for Sustainable Transportation (NCST), supported by the U.S. Department of Transportation (USDOT) through the University Transportation Centers program. The authors would like to thank the NCST and the USDOT for their support of university-based research in transportation, and especially for the funding provided in support of this project. The authors would also like to thank staff from the City of Atlanta and Atlanta Regional commission for providing network data and assistance.

NCST

Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis

A National Center for Sustainable Transportation Research Report

September 2024

Hongyu Lu, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta

Michael O. Rodgers, Ph.D., Schools of Civil and Environmental Engineering and Public Policy, Georgia Institute of Technology, Atlanta

Randall Guensler, Ph.D., School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta

NCST

[page intentionally left blank]

TABLE OF CONTENTS

List of Tables

List of Figures

Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis

EXECUTIVE SUMMARY

Current regional-scale and sub-regional-scale energy use and emissions models (and downstream modeling applications, such as microscale dispersion modeling) often use travel demand model outputs as impact assessment model inputs, because large-scale on-road activity and household trip data are difficult and expensive to collect. Although existing travel demand models generate synthetic households that are designed to match sub-regional distributions of household demographics for predicting vehicle activity, these models do not currently integrate detailed vehicle fleet specifications. Hence, vehicle class and model year information is not currently assigned to individual trips by the regional model. This means that travel demand models do not account for the differences in fleet composition by time of day and when transportation network traffic volumes and operating conditions are predicted. The integration of detailed vehicle information into the travel demand modeling process will facilitate the modeling of energy use and emissions at much higher spatial and temporal resolution. This research developed a modeling framework to integrate travel demand model outputs, licensed household marketing demographics data, and vehicle registration data to generate sub-region synthetic fleets and assign these vehicles to synthetic households in travel demand models and to each model-predicted household tour (connected series of trips).

The study focused on developing a generator for synthetic households and fleets by combining demographic data from ARC's ABM2020 model and Epsilon 2022 household data with regional vehicle data licensed from the R.L. Polk® Vehicle Registration Database. The synthetic fleet was generated by pairing the household locations (TAZ vs. longitude and latitude, and Epsilon household addresses vs. Polk® vehicle registration addresses), and a variety of demographic attributes (household size, number of children, number of vehicles, and household income). A case study was then carried out in metro Atlanta, Georgia, where emission results predicted using the synthetic fleet were compared to those from the traditional county-based fleet approach, where average emission rates from the regional fleet are applied to each trip by average speed and facility type. The case study compared these results across the 21+ million ABM-modeled daily trips.

Using Monte Carlo analysis across 1,000 primary iterations, the team derived average metrics of fleet ownership, VMT distributions, emissions, and differences across the two scenarios. The stability of these metrics was further verified using Bootstrap techniques. The synthetic fleet generation generated an average of 3,8 million vehicles for metro Atlanta. The tight 95% CI range spans from 3,836,204 to 3,836,965 vehicles, underscoring the stability of the synthetic fleet generation process. However, a deeper dive into vehicle ownership by type and age revealed an amplified representation of older vehicles in the synthetic fleet, even when juxtaposed with the original Polk® data. Hence, bias in licensed data associated with vehicle

ownership toward older vehicles will generate a multiplicative bias effect in energy use and emissions predictions, given the non-linear relationship between fleet age and on-road emission rates. The observations in this report underscore the need for using unbiased vehicle ownership and use data, which requires access to the state-managed vehicle registration database and smog check mileage accrual rates, to avoid using biased licensed data in synthetic household and fleet generators.

The temporal emission trends, as expected, aligned with traffic volume and speed variations, especially during peak hours. White-collar work trips emerged as the predominant contributors to CO emissions, followed by blue-collar work trips. Interestingly, emissions from discretionary and maintenance activities were very similar, both slightly less than the work trips. However, given the bias noted above in licensed vehicle registration data, it is possible that these results may not hold once unbiased fleet registration data are obtained and employed in future analyses (older vehicles tend to be correlated with use in lower income households).

As expected, the households that did not own vehicles emitted the least CO emissions, while high-income single households and moderate-income working households exhibited higher emissions. The most pronounced emissions were observed in groups with significant commute demands using older vehicles due to what appears to be less frequent vehicle turnover.

A spatial analysis of CO emissions differences between the two scenarios revealed downtown links to have smaller differences, while exurban links exhibited larger variability. Specific areas like the vicinity of Six Flags Georgia consistently underestimated emissions across all hours, and major northern interstates, including I-75, I-85, GA-400, and parts of I-285, displayed a consistent trend of overestimation. Such spatial biases underscore the need for caution when applying county-based emission rates to individual roadway links.

1. Overview

Travel demand model outputs are frequently utilized as inputs to energy and environmental sustainability analyses, and fleet composition is often addressed as a post-processing activity (assigning source type and vehicle age distributions to roadway links after travel demand model runs are finished). The development of a comprehensive modeling framework that integrates synthetic fleet generation directly into travel demand models offers an opportunity to enhance regional and sub-regional energy and emissions modeling by ensuring that appropriate fleets are assigned to network corridors based upon household travel choices (i.e., ensuring that synthetic household travel activities and their detailed trip characteristics are assigned to appropriate vehicles).

It is important for project-level emission assessment to represent the traffic operations (including fleet composition) at the time and location of interest, as a small change in model inputs can lead to large changes in energy and emissions predictions, given the variability and sensitivities of emission rates across source types and model years. Synthetic household and fleet data play a crucial role in large-scale emission modeling by better capturing the spatial and temporal variability of on-road fleet compositions, and helping to mitigate uncertainty associated with using aggregated input (Lu, et al., 2021; Granell, et al., 2002; Bachman, et al., 1998).

The synthetic fleet used in this study was developed using a “bottom-up” methodology, where vehicles in the fleet were assigned to households based upon household demographic data. Then, each vehicle-tour (and therefore trips within the tour) was assigned a household vehicle and predicted emissions were aggregated by trip, individual, and household level. The use of synthetic household and fleet data can also facilitate high-resolution equity analyses by aggregating energy use and emissions across demographic subgroups. These energy use and emission results can be further integrated with other equity analyses that employ travel demand model outputs, such as pollutant concentration exposure modeling.

This research aimed to develop a synthetic fleet generator and assignment system by analyzing and combining data from various sources: 1) licensed household-level demographic data from Epsilon®, Inc. that includes detailed information on household and individual demographics along with geographic coordinates (longitudes and latitudes); 2) vehicle registration data licensed from Polk® that provides vehicle make, model, and model year information (used to develop source type and vehicle age distributions) along with registration addresses (that could be paired with Epsilon® household addresses); 3) demographic information at the household level that is embedded in the Atlanta Regional Commission’s (ARC’s) activity-based travel demand model (ABM); and 4) ABM model outputs for trip purpose and retained trip path data. This comprehensive system generates a high-resolution synthetic fleet by pairing households using demographic characteristics and by allocating synthetic vehicles to each tour (trip chain).

A previously-developed regional emission modeling framework was used to assess trip-level and household-level emissions of Carbon Monoxide (CO) (Liu, et al., 2019; Xu, et al., 2018a; Xu, et al., 2018b). The mean and 95% confidence intervals were predicted by employing Monte

Carlo simulations and Bootstrap techniques on the synthetic fleet, and the results were compared against the emissions derived from sub-regional (county-based) results that are based on average emission rates by source type and by model year, which are extracted from the scenario of synthetic fleet. A separate stand-alone research paper was published by the research team that proposed a methodological model to assess the potential of electric vehicle (EV) penetration and corresponding reduction in energy use and emissions (Dai, et. al, 2022).

2. Data and Methodology

The synthetic household and fleet generator pairs demographic and vehicle information from various sources, including the Activity-Based Travel Demand Model (ABM) from the Atlanta Regional Commission (ARC), the commercial licensed household demographic dataset of Epsilon®, Inc., and licensed vehicle registration database from R.L. Polk® Company. Epsilon® vehicle ownership data was used in a cross comparison for 9+ year-old vehicles.

The ABM was configured to retain the modeled travel paths of every household trip (more than 21 million trips) along the 149,967-roadway links modeled in the metro Atlanta transportation network. The resulting emissions from these modeled trips were estimated using the Georgia Tech MOVES-Matrix implementation of the EPA MOtor Vehicle Emission Simulator (MOVES) 2014b emission rate model. The resulting link-level CO emissions were compared between two scenarios: 1) a link-specific fleet, using the synthetic fleet given the driver/household demographics of the trips expected to traverse that link, given the ABM trip outputs; and 2) link-by-link emission inventory developed by county, using fleet average emission rates (derived from the synthetic fleet scenario) weighted by household VMT. The Epsilon® households were first paired with Polk® vehicle information based on household address vs. registration address information, and a two-stage randomized pairing and allocation process was developed to associate: 1) the ABM households with the Epsilon®-Polk® households, and 2) vehicle information (source type and model year) with ABM trips. A Monte Carlo simulation and Bootstrap techniques were employed for 1,000 random draws, to derive the mean and variances of the fleet and emissions results. The emissions per person and per households derived from the synthetic fleet generation process were compared across the demographic groups, and the link-by-link emission inventory by hour was compared against the results that were based on county-level average emission rates.

2.1 Individual/Household Travel Behavior and the Roadway Network

The activity-based travel demand model (ABM) by Atlanta Regional Commission (ARC) was used in this study to provide the roadway network and the travel behavior at the households and individual levels using ABM path retention (Zhao, 2021; Zhao, et al., 2019; Zhao and Guensler, 2019). The specific version employed was the ABM2020-TIPA1-2020, which is same version used by ARC staff for the Atlanta Metropolitan Transportation Improvement Program (TIP), for calendar year 2020, modeled with the ARC planning assumptions for the transportation network, land use, and household demographics for calendar year 2020 (Zhao, 2021). The ABM2020-TIPA1-2020 model data is available in geodatabase format from the ARC and includes a links layer and a nodes layer. In total, the model network for this ARC scenario and modeling year includes 149,967 links and 66,418 unique nodes.

The team applied the path retention method that allows modelers to retain the paths between origin and destination predicted by the ABM's internal Franke-Wolfe algorithms. With path retention, model-predicted link-by-link vehicle traverses through the road network are available for analysis, and 21.3 million model-predicted household trips were retained.

2.2 Demographic and Vehicle Registration Information

Three datasets that were used in developing the synthetic fleet generator in this study as follows. The ABM2020 and Epsilon® 2022 data provided demographic information at both the individual and household level, and the R.L. Polk® Vehicle Registration Data provided vehicle information that can be paired with demographic data, by allocating vehicles to registration addresses.

- ABM2020: The ABM provides the demographic information at the individual and household level of 2.39 million households which couples with the household trip information.
- Epsilon® 2022 Demographic Data: The commercial licensed dataset of Epsilon® (version 2022) provides the household and individual demographic information of 3.21 million households in metro Atlanta.
- R.L. Polk® Vehicle Registration Data 2022: These data provide the vehicle makes, model and model year information of 1.35 million vehicles in metro Atlanta. These data were found to have significant limitations as discussed in Section 2.2.2 below.

2.2.1. ABM Demographic Information

The ABM retains household demographic data and person assignment for each trip, which includes household size, number of workers, annual household income, number of vehicles, household traffic analysis zones (TAZ), etc. for the 2.39 million households modeled in the travel demand model. ABM also provides demographic information for every person in each household, which includes age, gender, worker type (full-time worker, university student, non-worker, etc.), etc., and the team paired the person information with the household data to derive number of children and whether the household was a single-parent family with children.

The 16 demographic groups used in this study was consistent with the previous study as definitions that are both mutually exclusive and collectively exhaustive (Zhao, 2021; Zhao, et al., 2019; Zhao and Guensler, 2019). A breakdown of the individual group characteristics is presented in Table 1. These groupings were previously established through a collaboration between U.S. Department of Energy staff and Georgia Tech researchers for the 2018 ARPA-E TRANSNET-Atlanta project for equity impact assessment (Zhao, 2021; Zhao, et al., 2019; Zhao and Guensler, 2019). These mutually exclusive groups include lifecycle stages, but were also specifically designed to reveal transportation mobility and accessibility impacts on households that own no vehicles, households with very low-income households, and single parent households with children (i.e., severe constraints to mobility and accessibility).

The households were classified first by vehicle ownership (see Zhao, 2021 for more information on grouping methods). Households with no vehicles were placed into Group #1 (households with no vehicles). The households with at least one vehicle were then examined and divided based on annual income; those below the threshold of $25,000 were classified as Group #2 (low-income households). The remaining households were then categorized based on household size (number of adults and children), and further classified based on number of workers, presence of children, and whether the household includes a single parent with children (for households with three or more persons only). The one-person households were placed into Group #3 (non-working households with annual income lower than $60,000), Group #4 (working households with annual income lower than $60,000), and Group #5 (annual income more than $60,000). Households with two or more persons were examined by annual income and categorized into intervals of $25,000 to $60,000 (Groups #6, #7, #8, and #9), $60,000 to $120,000 (Groups #10, #11, #12, #13, and #14), and $120,000 or more (Groups #15 and #16). For households with annual income lower than $60,000 (and with two or more persons), the non-working households were classified into Group #6, and two-person working households were classified into Group #7. Then, the working households with three or more persons were classified based on whether they are single parent with children (Group #8) or without children (Group #9). For the households with annual income from $60,000 to $120,000, the households with either no or one worker were classified based on whether they have at least one kid (Group #10) or not (Group #11), and the two-person households with two workers were classified into Group #12. The three-person households with at least two workers were classified into Group #13, and those households with four or more persons and with at least two workers were classified into Group #14. The households with annual income of $120,000 or more were classified into Group #15 (with no or one worker) and Group #16 (with two or more workers). The classifications result in each household being placed into one and only one group (exclusive and exhaustive classification).

It is worth noting that although 6.3% of the households in the model do not own any vehicles, the ABM still predicts some household vehicle trips (e.g., by ride-sharing services, taxis, rideshare passenger, etc.). These households were treated as special cases in this study and were not allocated in the emissions comparison (see the following sections).

Table 1. Definition of the 16 Demographic Groups based on ABM household information.

Group #	Own Vehicles	Low Income	HH Size	Annual Income	Workers	Vehicles	With Kid(s)	Single Parent w/kid(s)	Household Counts	Percentage
1	No	Any	Any	Any	Any	0	Any	Any	150,300	6.3%
2	Yes	Yes	Any	$0 - $25k	Any	1+	Any	Any	289,947	12.1%
3	Yes	No	1	$25k - $60k	0	1+	Any	Any	59,005	2.5%
4	Yes	No	1	$25k - $60k	1	1+	Any	Any	205,224	8.6%
5	Yes	No	1	$60k+	0 or 1	1+	Any	Any	165,176	6.9%
6	Yes	No	2+	$25k - $60k	0	1+	Any	Any	101,743	4.2%
7	Yes	No	2	$25k - $60k	1+	1+	Any	Any	155,804	6.5%
8	Yes	No	3+	$25k - $60k	1+	1+	Any	Yes	21,853	0.9%
9	Yes	No	3+	$25k - $60k	1+	1+	Any	No	236,925	9.9%
10	Yes	No	2+	$60k - $120k	0 or 1	1+	Yes	Any	98,867	4.1%
11	Yes	No	2+	$60k - $120k	0 or 1	1+	No	Any	152,092	6.4%
12	Yes	No	2	$60k - $120k	2	1+	Any	Any	127,920	5.3%
13	Yes	No	3	$60k - $120k	2+	1+	Any	Any	112,760	4.7%
14	Yes	No	4+	$60k - $120k	2+	1+	Any	Any	172,743	7.2%
15	Yes	No	2+	$120k+	0 or 1	1+	Any	Any	104,872	4.4%
16	Yes	No	2+	$120k+	2+	1+	Any	Any	238,157	10.0%
Total	All	All	All	All	All	All	All	All	2,393,388	100.0%

2.2.2 Licensed Demographic and Vehicle Registration Datasets

In addition to the ABM demographic information, the licensed demographic dataset from Epsilon® and the vehicle registration data from R.L. Polk® were also used in the synthetic fleet generator.

The R.L. Polk® vehicle registration dataset provides information on 1.35 million vehicles in metro Atlanta, which includes the vehicle identification number (VIN), registration address, 9-digit ZIP code, vehicle make, model, model year, body type (sedan, hatchback, coupe, truck, etc.), Polk® vehicle type (cars, mini-vans, sport utility vehicles (SUVs), and pick-up trucks), odometer (only for approximately 20% of the vehicles), and other parameters. Because these proprietary data also include personally identifiable information (PII), such as individual first name and last name, the team processed these data with utmost care and confidentiality. All PII data used in this study were stored on a secure server located on the Georgia Tech campus and data were not permitted to be removed from the secure location to ensure data privacy and protection. All PII information was removed from the data before processing (no PII is needed by the synthetic fleet generator), and the data were processed using local computers located in a secure office on Georgia Tech campus. Access to this office was limited to authorized team members who passed human subject reviews, and only authorized members of this research project were granted access to the computer that stored the project (processed) data. These measures were taken to prevent any unauthorized access or disclosure of PII and to protect the privacy and confidentiality of the individuals whose data were used in this study.

The Epsilon® 2022 dataset provided information of 3.21 million households in metro Atlanta, and a total of 95 attributes that provide household and individual demographics, including household size (number of adults and number of children), household income (relative index compared to national average), ethnicity of every household member, household address, household ZIP code, and the longitude and latitude of the household.

The synthetic fleet generator was based on pairing the geographic locations of households across various datasets, and thus it was important to perform quality assurance and quality control (QA/QC) checks on the data by comparing the Epsilon® and R.L. Polk® address information. Out of the 3,215,967 records in Epsilon dataset, the research team identified 101,688 addresses that were P.O. Boxes (about 3.2%), and 45,335 records out of 1,356,745 vehicles in Polk® vehicle registration dataset were P.O. boxes (about 3.3%). The team decided to retain these records, given that they appear in both datasets, for use in verification comparisons.

Duplicate addresses were also identified in Epsilon® 2020 data and in the R.L. Polk® dataset, and the team reviewed the distribution of the addresses by number of repeats. The maximum number of repeats of Polk® dataset was eight, which is not regarded as a data quality issue (one household can have multiple vehicles). The team performed a manual review of 66,763 duplicate Epsilon® data records (representing 180 unique addresses) by sampling addresses within each repeat interval (defined by the number of times each address is repeated). The results indicated approximately half of the records that come with 1,000+ repeats are commercial addresses, such as post offices, parking garages, and auto services (not residential addresses). These locations with large number of duplicates represent fleets, but the largest values appear to represent leasing companies. The vehicles at these commercial addresses were removed, because they could not be assigned to individual households. However, excluding leased vehicles, which are generally less than five-years-old, will necessarily bias synthetic fleet generation and assignment. Future implementations of synthetic fleet models need to account for the assignment of leased vehicles to specific households, which might be obtained directly from the state vehicle registration database, or derived by combining registration records with other sources, such as smog check data or vehicle insurance data.

The team was unable did to perform a more thorough review of all addresses manually due to the following constraints: 1) manual reviewing of all Epsilon® addresses takes a significant amount of human labor, and the team is working on developing a machine learning-based algorithm to facilitate an automatic process of QA/QC; 2) commercial addresses only account for 5% to 10% records (estimated), or 1% of the unique addresses of the total households; and 3) the R.L. Polk® vehicle registration data were paired with the Epsilon® households on a many-to-one relationship (multiple vehicles were allocated to households base on having the same registration address), and the duplicates only impacted the similarity pairing between ABM-Epsilon® (which was based on random sampling).

The team also received additional vehicle ownership information from Epsilon (purchased March 2023), which included 19 new attributes and provided the number of vehicles (can be more than five vehicles), make, model, and model year information (for up to five vehicles per household), for 770,563 households. The team performed a preliminary QA/QC process of the Epsilon vehicle ownership data, and it was indicated that 393,041 households out of 770,563 records (approximately 51.0%) were either an exact match with the synthetic fleet results (including make, model and model year information), or a subset of the synthetic fleet (the synthetic fleet allocated more vehicles to the households, but the rest of the vehicles are exactly the same). This indicates that the generated synthetic fleet had a good pair rate with commercial dataset, and that the Epsilon® and R.L. Polk® data may have been derived from (or partially derived from) the same or similar sources. The team did not consider this agreement to indicate that both sources were accurate. The Epsilon® vehicle ownership dataset was not used in this study based on issues described below.

A preliminary QA/QC process was performed targeting the Epsilon® vehicle ownership dataset (documented in Appendix A). The vehicle ownership information in the Epsilon® licensed data were found to contain little, if any, information on vehicles less than nine years old, and thus these data could be used only in allocating only older (9+ years old) to households. Similarly, the vehicle age distributions derived from the R.L. Polk® Vehicle Registration Dataset were also found to have an unrealistically low proportion of younger (less than nine years old, post 2014 model years) based on historic vehicle sales data. These R.L. Polk® are shown in Figure 1. These commercial data sets were compared the vehicle age distributions from Georgia Department of Natural Resources (DNR) based on state vehicle registration data and the on-road observed fleet age distribution for a recent project for the Georgia State Road and Toll Authority (Randall Guensler, et al., 2022). This comparison is shown in Figure 2 and Figure 3. From the figures, it is clear that the R.L. Polk® Vehicle Registration Data show many fewer younger vehicles (less than nine years old) than the other data sources. This difference was reported to R.L. Polk® who is evaluating the source of the bias. In the interim, for the purposes of this study the R.L. Polk® data were used only for allocation of older (nine years or older) vehicles within the synthetic fleet generator and the allocation of younger vehicles (less than nine years of age) was randomized, using proportions derived from the DNR data. The proposed methodology (pairing demographics, travel demand model and vehicle registration data, and comparing the synthetic fleet vs. regional fleet vs. sub-regional fleet in terms of emission modeling results) was used for a case study of vehicles that are nine years or older. The results of all model years vs. those of vehicles of 9+ years are discussed in the Results and Discussion section below. The full methodology can be applied once more reliable registration data become available for the younger vehicles in the fleet (as noted earlier, the team has concluded that access to state registration data will be required).

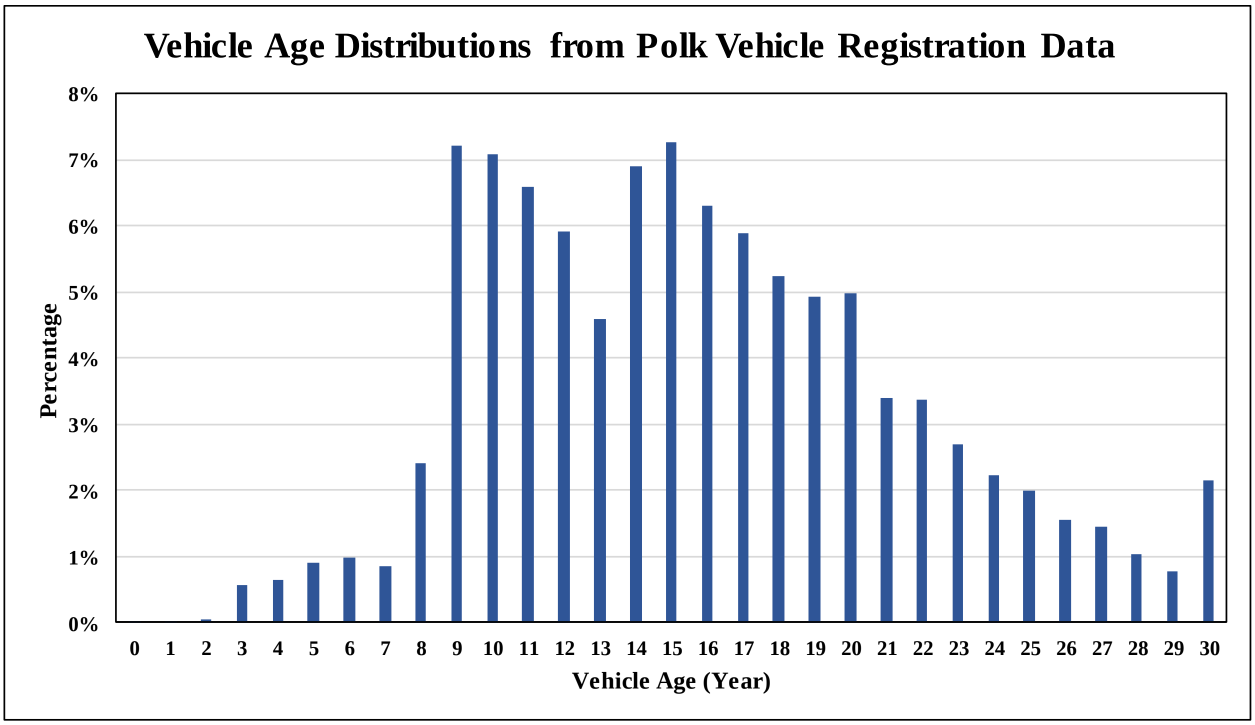

Figure 1. Vehicle age distributions of Polk® Vehicle Registration Data.

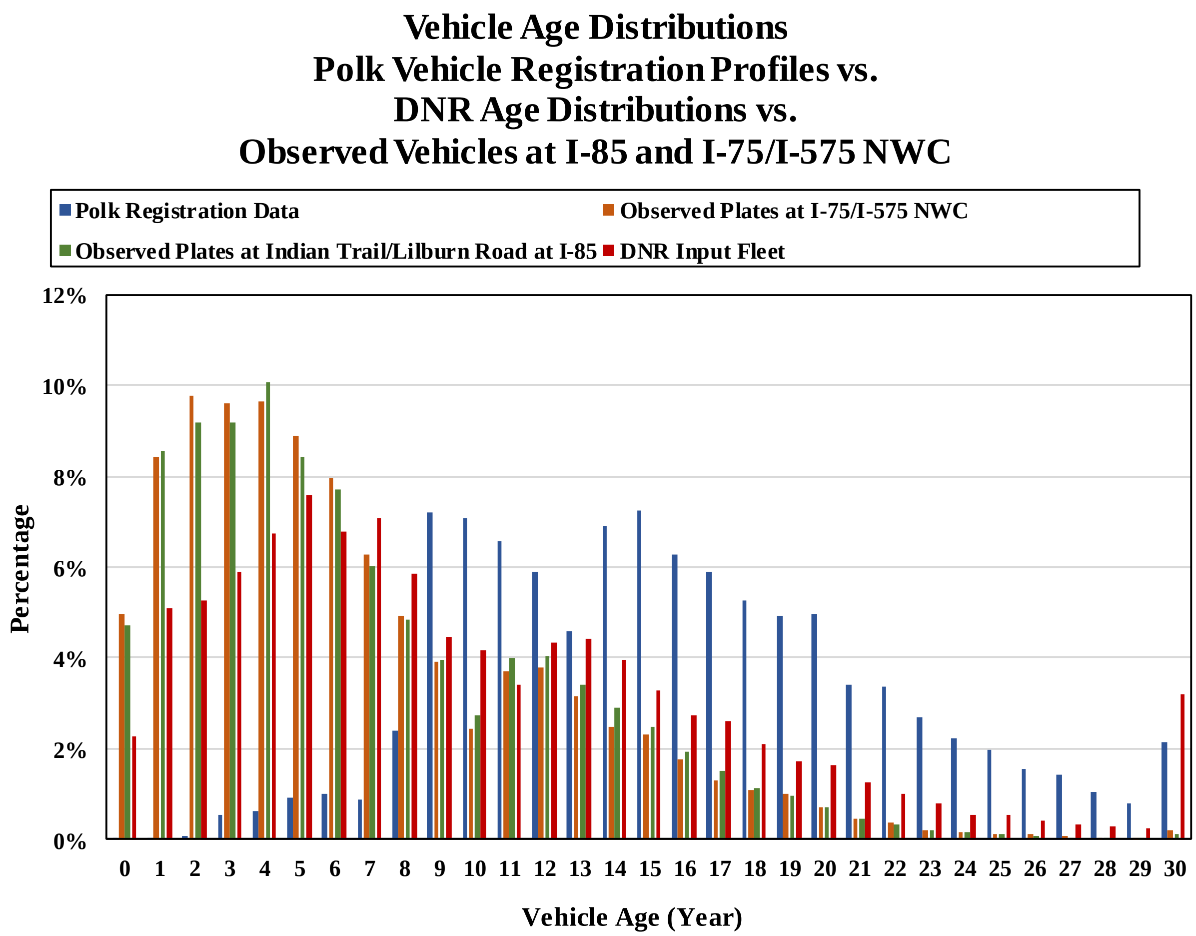

Figure 2. Vehicle age distributions of Polk® Vehicle Registration, observed license plates and DNR fleet.

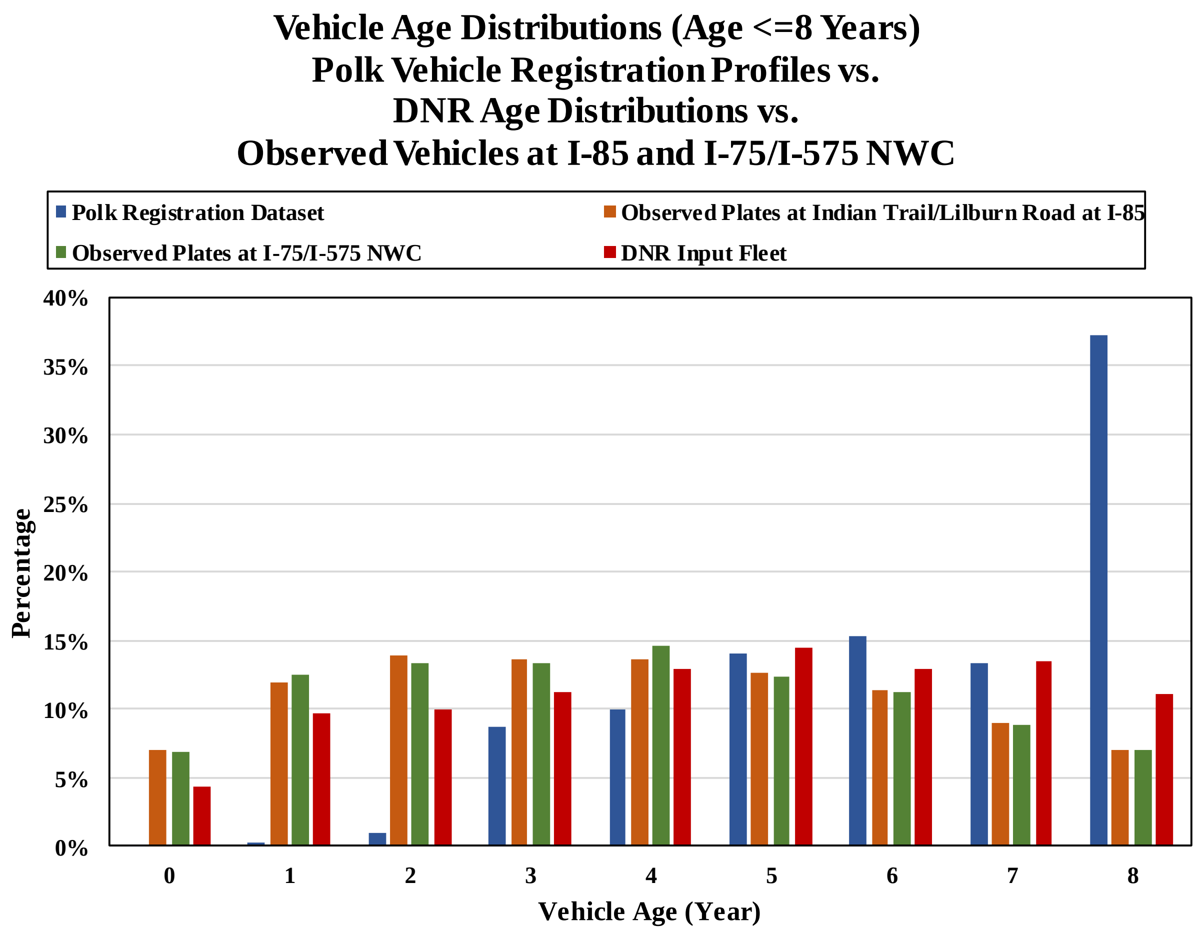

Figure 3. Vehicle age distributions (vehicle age <= 8 Years) of Polk® Vehicle Registration, observed license plates and DNR fleet.

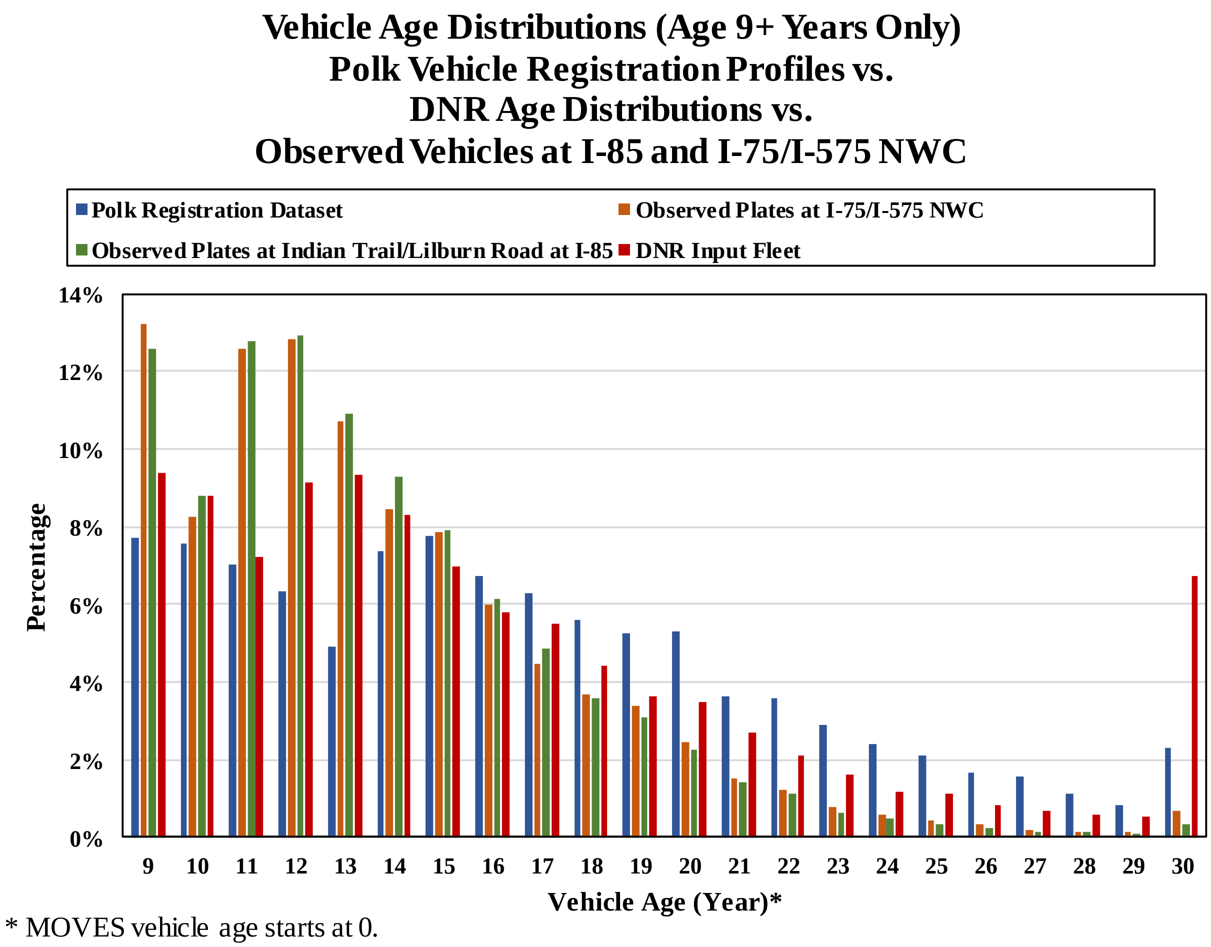

Figure 4. Vehicle age distributions (vehicle age > 9 Years) of Polk® Vehicle Registration, observed license plates, and DNR fleet.

2.3 Synthetic Household Generator

The synthetic fleet generator works in a two-step process initiated by allocating the R.L. Polk® vehicle information for the older vehicles to Epsilon® demographic dataset (address pairing), and by pairing the demographics of ABM vs. Epsilon® 2022 (similarity based on household size, household income, number of vehicles, and number of children). The R.L. Polk® vehicle registration information was first allocated to the Epsilon® demographic data by pairing the household addresses. The ABM and Epsilon® households were then integrated by a comprehensive similarity assessment (using weighted Euclidean distance) that was consistent with the previous study (Dai, et al., 2022).

The synthetic household generation was fleet-oriented (source type and model year needed to model vehicle emissions), and it is important to include the variables that impact household vehicle ownership (number of vehicles, vehicle make, model, and model year). The Epsilon® dataset does not come with the vehicle ownership data (until the late purchase in March 2023), and the Polk® vehicle registration data was allocated to the Epsilon® households by pairing the registration addresses to the residential addresses. Before conducting the pairing process, address information from both datasets were first converted to a standard address format by replacing the address suffixes. The table of address suffixes is provided in Appendix B. Given the duplicate addresses in Epsilon® 2022 dataset, when a single Polk® address corresponded to multiple Epsilon® households sharing the same address, the vehicle information was duplicated and paired with each of these households (one to many relationship). This duplication was intentional to create a pool of potential vehicle ownership options (rather than defining a fixed fleet composition), so that the subsequent random draws are based on a diverse set of potential vehicle ownership scenarios.

A total of 1,047,965 vehicles were paired to the Epsilon® households (approximately 82.57% of the R.L. Polk® Vehicle Registration Data after removing younger vehicles), and the vehicles (no younger than nine years) that were not paired to any Epsilon® household were reviewed to make sure no vehicles were missing. A subset of 300 vehicles (randomly sampled out of the 221,187 vehicles that were not paired to any Epsilon® household) were manually reviewed to make sure these addresses were not in the Epsilon® datasets (no pairing needed).

Most of the unpaired records are those in external counties where Epsilon® records are sparse, as shown in Figure 5, and the three ZIP codes with high percentages in midtown/downtown Atlanta are 30332 (Campus of Georgia Tech), 30322 (Campus Emory University), and 30334 (State Capitol and Legislative Office), in which Epsilon® has very limited households.

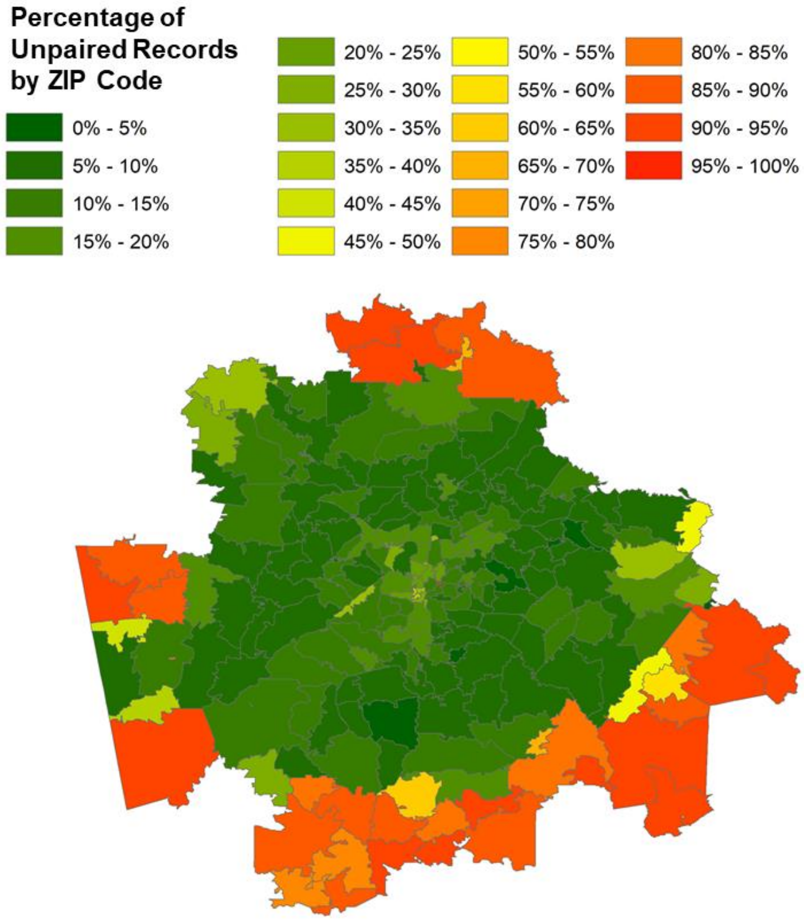

Figure 5. Percentages of Polk® Vehicle Records that were not paired by ZIP code.

The Epsilon® households that were not allocated any R.L. Polk® vehicle were temporarily removed before pairing with Epsilon households to ABM, and the ABM households with no vehicle (demographic group #1) were also excluded from the subsequent pairing, to make sure that all integrated results have at least one vehicle as defined.

The inherent one-to-many relationships that exist between the ABM, Polk®, and Epsilon® datasets introduce variability in the possible pairings, and a deterministic approach to pairing would not capture the full spectrum of potential outcomes. The team employed a combination of Monte Carlo simulation and Bootstrap technique based on random resampling of the Epsilon® households associated with Polk® vehicle information, and performed a total of 1,000 iterations of the synthetic household and fleet generator.

The Epsilon® 2022 household locations are by longitude and latitude, and these households were first allocated to TAZs in metro Atlanta (ABM household locations are at TAZ level). The ABM households by TAZ were then iteratively examined by assessing their similarities to the Epsilon® households within the same TAZs, and weighted Euclidean distance was used to represent a comprehensive similarity between a pair of ABM and Epsilon® households based on household size, household income, number of vehicles, and number of children, as shown in Equation (1).

$$D_{m,n} = \left(\sum_{i=1}^{4} w_i \left(x_{m,i} - b_{n,i}\right)^2\right)^{\frac{1}{2}} \quad (1)$$

where $D_{m,n}$ is the weighted Euclidean distance between the mth Epsilon® household vs. the nth ABM household, w_i represents the weight of attribute i (from 1 to 4, and represents household size, household income, number of vehicles, and number of children), $x_{m,i}$ and $b_{n,i}$ represent the value of the attribute i for mth Epsilon® household and the nth ABM household.

The weights of the input variables were derived based on a consistent methodology with the previous study (Dai, et al., 2022) by manually reviewing the permutation importance of a decision tree regression and the feature importance of a random forest allocation. Permutation feature importance is a model-agnostic method used to estimate the importance of input features by evaluating the decrease in model performance when the values of a specific feature are randomly shuffled (Brière, et al., 2021), and it works by breaking the relationship between the feature and the target variable, and by subsequently measuring the impact on prediction accuracy (if permuting a feature leads to a significant drop in model performance, that feature is considered important for model's predictions). The team also assessed the feature importance using a Random Forest Regressor, an ensemble learning method that offers improved accuracy and robustness by averaging impurity reduction across multiple decision trees (Breiman, 2001; Breiman, et al., 1984). The weights of the four input variables were shown in Table 2, and a detailed description of the weights and the feature importance can be found in Appendix C.

Table 2. Assigned weights for ABM Epsilon household similarity.

Variable	Weights Assigned
Household Size	0.30
Household Income	0.35
Number of Vehicles	0.35
Number of Children	0.10

Each ABM household was then paired with the Epsilon® household that was allocated with the greatest similarities (minimum weighted Euclidean distance) to integrate the demographic information of ABM vs. Epsilon®. One ABM household can correspond to multiple Epsilon® households (within the same TAZ) that all exhibited the same similarities (i.e., distances that are all minimum), and these households form a pool of potential households that can be paired with the ABM household. One of these Epsilon households was selected to be paired with the ABM household in the subsequent random process of the synthetic fleet generator.

The team assured all Epsilon®-Polk® households were paired with ABM household with at least one vehicle, by excluding Epsilon® households that was not allocated with any record from R.L. Polk®. For ABM households with no vehicle ownership (Demographic Group #1) that were temporally excluded, no Epsilon®-Polk® record was allocated to these households at this stage (a random vehicle from the fleet population was assigned to each of their trips in the subsequent processes).

2.4 Synthetic Fleet Generator

The inherent one-to-many relationships between the ABM, Polk®, and Epsilon® datasets introduce variability in the possible pairings, and a deterministic approach to pairing would not capture the full spectrum of potential outcomes. The team employed a combination of Monte Carlo simulation and Bootstrap techniques to assess the fleet composition and the resultant CO emissions. A two-stage random allocation algorithm was designed to: 1) pair the ABM households with the Epsilon®-Polk® households, and 2) allocate a random vehicle to each of the ABM-predicted tours and trips. A total of 1,000 primary iterations of random allocation were performed to estimate the average and 95% confidence interval of the results.

The first stage of the synthetic fleet generator paired each ABM household with an Epsilon®-Polk® household, based on the previously mentioned similarities. If multiple Epsilon®-Polk® households exhibited the same highest degree of similarity, a random Epsilon®-Polk® household was selected from this pool for pairing. The random pairing algorithm ensured that no Epsilon®-Polk® household was paired more than once, and when two ABM households were matched with the same Epsilon®-Polk® household, a re-draw was conducted for all households within that TAZ.

The second stage of the generator allocated vehicles of each ABM households to their trips. ABM trips are structured as tours, which are essentially chains of consecutive trips that start and end sequentially, and the team assumed that a single vehicle is used throughout a tour (i.e., no vehicle switch mid-tour). In each primary iteration of the synthetic generator, an Epsilon®-Polk® household was paired with each ABM household, and the paired household could have multiple vehicles assigned to it. For ABM households that was paired with only one vehicle, that vehicle was automatically allocated to all the trips that ABM predicted for that household. For households that were assigned multiple vehicles, a specific vehicle was chosen for all trips in each tour.

This allocation of vehicles was performed through an iterative random sampling process that is weighted based on the relative mileage accumulative rates (RMAR) that are used in MOVES2014 (U.S. EPA, 2016) and MOVES3 (U.S. EPA, 2021), as shown in Table 3 (using the same RMAR for both MOVES versions). The derivation of RMAR for passenger cars was based on the NHTSA report on survivability and mileage schedules (NHTSA, 2006), and a comprehensive explanation of how RMARs were derived is provided in Appendix D.

Table 3. Relative Mileage Accumulative Rates (RMAR) for passenger cars and passenger trucks from MOVES2014 and MOVES3.

MOVES Vehicle Age ID	Passenger Car (Miles)	Passenger Truck (Miles)
0	12,594	16,085
1	13,961	15,782
2	13,669	15,442
3	13,357	15,069
4	13,028	14,667
5	12,683	14,239
6	12,325	13,790
7	11,956	13,323
8	11,578	12,844
9	11,193	12,356
10	10,804	11,863
11	10,413	11,369
12	10,022	10,879
13	9,633	10,396
14	9,249	9,924
15	8,871	9,468
16	8,502	9,032
17	8,144	8,619
18	7,799	8,234
19	7,469	7,881
20	7,157	7,565
21	6,866	7,288
22	6,596	7,055
23	6,350	6,871
24	6,131	6,739
25	5,940	6,663
26	5,780	6,648
27	5,654	6,648
28	5,562	6,648
29	5,508	6,648
30	5,494	6,648

In our study, every primary iteration that paired ABM households with Epsilon®-Polk® households generated a set of fleet ownership profiles (vehicle counts by source type and vehicle age), and the weights for random vehicle allocation were initially determined by dividing the RMAR by the respective vehicle counts (i.e., average mileage per vehicle). Vehicles with higher weights were more likely to be selected (higher mileage), and these weights were applied in a relative manner within the context of each household's set of vehicles. For example, a household with three vehicles would have their weights computed separately and relatively, even if two vehicles belong to the same source type and age.

The weights were further adjusted through sub-iterations to make sure that the overall VMT distributions by source type and by vehicle age aligns with the RMAR in MOVES. The adjustment factors of each sub-iteration were based on the deviation between the VMT distributions generated in the prior sub-iteration vs. the desired VMT distributions (i.e., the ones derived from RMAR), as shown in Equation (2). The average deviation across the age groups was used as a metric to assess the overall difference between the generated vs. desired VMT distributions, as shown in Equation (3), and a value closer to 0 indicates that the generated distribution closely aligns with the desired RMAR, while values further from 0 denotes greater deviations.

$$Adj_{s,i} = \frac{Desired\ VMT\ Distributions_{s,i}}{Generated\ VMT\ Distributions_{s,i}} \quad (2)$$

$$Deviation_{avg} = \frac{1}{N}\sum_{s,i}\left|Adj_{s,i} - 100\%\right| \quad (3)$$

where s is the source type (passenger cars and passenger trucks), i is the MOVES vehicle age ID (from 9 to 30), and N is the total number of age groups across both source types (44 in this case).

Through empirical testing across 30 runs with varying parameters, it was observed that an average deviation threshold of 0.05 serves as an optimal convergence criterion. When the average deviation falls below this threshold, the generated distributions tend to oscillate rather than further converge, which indicated a stabilization in the alignment with the desired RMAR. The empirical tests indicated that 20 sub-iterations per primary iteration took 10+ hours to finish (not surprising given the substantial size of the input data), while the deviations ceased to decrease at around sub-iteration #10 to sub-iteration #15, and the team capped the maximum number of sub-iterations at 25 to prevent potential infinite loops.

The weighted random allocation approach systematically refined the VMT distributions to draw them closer to the desired RMAR with each sub-iteration. However, it is important to note that the generated VMT distributions may not perfectly align with the RMAR even after achieving convergence. The vehicle allocation process is constrained by the vehicles that were already assigned to households. While vehicle ownership does not dictate the VMT distributions, it inherently determines the upper limit of the variability of the generate VMT across source type and vehicle age, regardless of our vehicle allocation strategy. More discussion can be found in Chapter 3.

The preferences of vehicles are related to a variety of other factors that are dependent on trip purposes and trip lengths, such as fuel efficiency, passenger and cargo capacity, vehicle age and reliability, comfort and features, weather, driving conditions (e.g., terrain), and even personal attachment or emotional factors, and these preferences may vary across the households. However, a comprehensive preference modeling process requires extensive data and will likely require separate model development efforts (Choo, S., and Mokhtarian, P. L., 2004), that are beyond the scope of this study. The synthetic fleet generator can incorporate vehicle choice models into the modeling framework, and the team is currently assessing the uncertainty associated with the random vehicle selection (the assumption that vehicles have equal chances to be allocated to trips regardless of trip purpose and trip length).

2.5 Emissions Modeling

Two scenarios were developed based on the generated synthetic households and fleet, and the emission of CO were modeled for both scenarios. The first scenario of synthetic fleet was based on allocating the vehicles to the predicted routes that they were predicted to travel. For each primary iteration of the synthetic generator, the emissions were modeled for each traversed roadway link for all ABM-predicted trips, and these emission results were further aggregated by roadway link, by tour purpose, by households, and by demographic groups. The second scenario adopted a county-based approach. The emission rates per vehicle per mile by county was extracted from the results of the first scenario, and the emission rates by source type and by vehicle age were applied to each link to derive the link-by-link emission inventory, as shown in Equation (4). For each primary iteration, the link-by-link differences between the results from two scenarios were compared.

$$ER_{fleet} = \sum_{ST} \sum_{MY} ST\% \times MY\%_{ST} \times ER_{ST,MY} \quad (4)$$

The emission rates were extracted from MOVES-Matrix, that produces precisely the same results as running the USEPA MOVES 2014b model for the analysis (Kim, et al., 2020; Liu, et al., 2019; Xu, et al, 2018a; Xu, et al, 2018b). By running MOVES about thirty thousand times for a region (i.e., areas that employ the same fuel specification and inspection and maintenance programs), across all combinations of input variables that affect emission rates, a multi-dimensional emission rate matrix of 90 billion energy and emission rates is generated. Users can query the emission rates directly from the matrix and thus improve run time efficiency (Liu, et al., 2019); performing a matrix query is about 200x faster than a MOVES run.

MOVES-Matrix can also support rapid analyses of engine starts, truck hoteling, evaporative sources, brake/tire wear (Xu, et al., 2018b) with MOVES 2014b. MOVES-Matrix can be easily coupled with vehicle activity analysis (Liu, et al., 2019; Xu, et al., 2018a; Li, et al., 2018; Guensler, et al., 2017; Xu, et al, 2016) by importing second-by-second vehicle operations. The tool can also be used to compare emissions across individual vehicles (Guensler, et al., 2017) when monitored vehicle activity data are avalable. MOVES-Matrix can be applied to a variety of transportation models, such as travel demand models (Xu, et al., 2018), and microscopic traffic simulation models (Xu, et al., 2016), or applications of emissions modeling that require high-efficient model performance, such as sensitivity assessment (Lu, et al., 2021; Lu, et al., 2020).

MOVES-Matrix is highly-desirable for regional-scale dispersion analysis (Kim, et al., 2020; Lu, et al., 2021), with high-performance to deal with links from large-scale networks, variations in meteorology, and traffic operation input, and with its user-friendly nature to minimize potential human error in running MOVES (especially when it comes to increased number of input links).

MOVES-Matrix provides a look-up table for each modeling region, and for the Atlanta metro area, MOVES-Matrix contains sub-matrices based on combinations of calendar year, season (Spring/Fall, Summer, Winter fuel season), temperature (0º-110º F with 1º F-bin intervals, 111 bins for Atlanta), and relative humidity (0%-100% with 5%-bin intervals, 21 bins for Atlanta).

The hourly AERMET meteorological profiles were provided by the GA EPD (24 hours × 365 days of the year) and the temperature and humidity profiles of the calendar year of 2021 were employed, as 2022 were not yet available when the runs were processed (Georgia EPD, 2023). Meteorological data from AERMET were rounded to the appropriate temperature and humidity to link with appropriate sub-matrices for each MOVES-Matrix run.

The default ABM output by link also includes non-household trips (e.g., delivery, long term truck trips, etc.), but only the modeled household trips were investigated. ABM household trips were reallocated to the roadway links by “following” their travel paths (all links they traversed) to generate the household traffic volumes by link, while keeping the ABM speed output. The household traffic volumes were used for emission modeling for the second scenario. Model-predicted average link speeds were used in MOVES emission rate selection and remained constant across scenarios.

The ABM output predicts trips of typical workdays, and the emission modeling was performed for every working hour in 2022. For each hour, the emissions of all 149,967 links were assessed, and a total of 249 days × 24 hours per day = 5,976 working hours were modeled (249 workdays), after excluding weekends and Georgia holidays.

The synthetic fleet scenario generates the travel paths for individual vehicles, and the team modeled each vehicle separately by “following” their travel paths (in this scenario, each MOVES-Matrix link essentially assesses the emissions of one vehicle on one roadway link). A total of 523,979,692 MOVES links per day × 249 days = 130,470,943,308 MOVES links were modeled per primary iteration for the first scenario.

The second scenario models each link of the roadway network, for each hour in all workdays. A total of 149,967 links per hour × 249 days × 24 hours per day = 896,202,792 MOVES links were modeled per primary iteration for the second scenario.

Upon completing the 1,000 primary iterations, the average metrics (emissions by link, differences by link, emissions by demographic group, etc.) were calculated by taking the mean values of the results across all iterations, and the 95% confidence intervals (CIs) were determined using the 2.5 and 97.5 percentiles across the iterations (Bootstrapping allows the mean to follow normal distributions due to the Central Limit Theorem).

A total of 130 trillion MOVES links were modeled in this study, as shown in Table 4, and the MOVES-Matrix runs were launched and finished on the Georgia Tech PACE supercomputing cluster.

Table 4. Number of modeled MOVES matrix links.

Scenario	Number of MOVES-Matrix Links	Note
Synthetic Fleet	130,470,943,308,000	523,979,692 Links/Day × 249 days ×1,000 Iterations
County-based	896,202,792,000	149,967 Links/Hour × 5,976 Hours×1,000 Iterations
Total	130,471,839,510,792	

3. Results and Discussion

This section presents the generated vehicle ownership and VMT distributions, the link-by-link emission results, and the emission results by tour purpose and demographics. All metrics presented in this section are derived from 1,000 iterations, and we provide both the average values and their corresponding 95% confidence intervals (CI) to ensure a robust understanding of the results. In the column and bar charts of this chapter, error bars indicate the upper and lower bonds of the 95% CIs (they might be difficult to distinguish due to their tight ranges).

3.1 Fleet Composition and VMT Distributions

The synthetic fleet generation resulted in an average of 3,836,602 vehicles owned by metro Atlanta, with the 95% CI range from 3,836,204 vehicles to 3,836,965 vehicles. The small half-width of the 95% denotes a standard deviation of smaller than 0.01%, and it was indicated that the synthetic fleet generations tend to be stable. The results of other metrics, such as emission results, indicated similar stability.

The vehicle ownership by source type and vehicle age is shown in Figure 18 (where the error bars indicate the upper and lower bounds of the 95% CIs), which indicates that more passenger cars were observed than passenger trucks, except for model years 2000 to model years 2003 (vehicle ages from 18 years to 20 years). Vehicle ownership (vehicle counts) are listed in Appendix E. The comparison between the R.L. Polk® data vs. the synthetic fleet is shown in Figure 7, indicating that the generated fleet tends to exhibit an even larger share of old vehicles than the original Polk® data.

It's important to note that vehicles younger than nine years old were excluded from our dataset prior to the synthetic fleet generation. Hence, the age distributions derived from the Polk® data indicated a potential bias, with a higher prevalence of older vehicles than expected. This initial bias also appears to be amplified in the post-synthetic pairing.

Several hypotheses might explain this observed amplification. The removal of newer vehicles might have had a disproportionate impact on certain areas. For instance, wealthier households, which might predominantly own newer or classic older vehicles, could have been affected more by the removal. This could lead to an overrepresentation of older vehicles in these areas, skewing the overall distribution. Also, the inability to incorporate more than 3% of the fleet representing leased vehicles registered to a leasing company also will have biased fleet age (to be older) and will have affected assignment to those households that may have been more likely to lease. Even after excluding the younger vehicles, there also might be other underlying biases or issues in the Polk® data that were not immediately evident from the overall distributions. It will be critical to repeat this research using data from the state vehicle registration database (if access can be obtained) to avoid these biases.

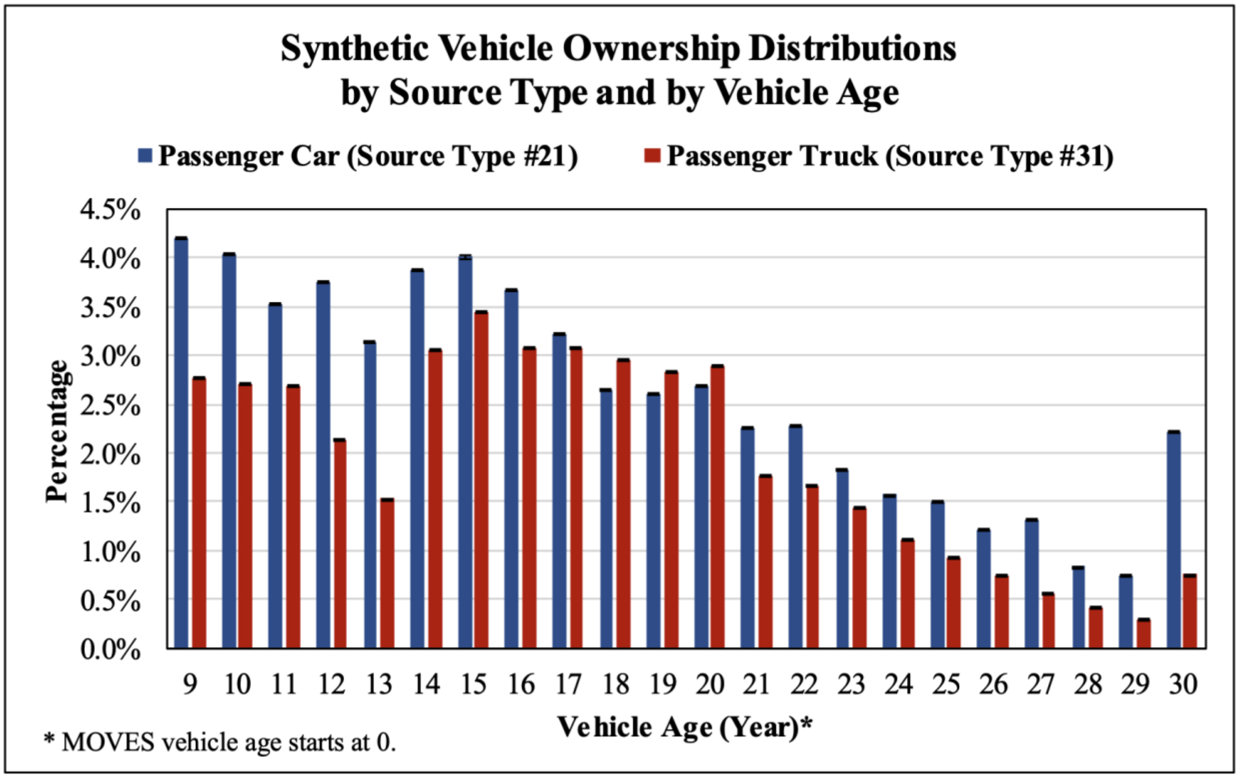

Figure 6. Regional vehicle distributions by source type.

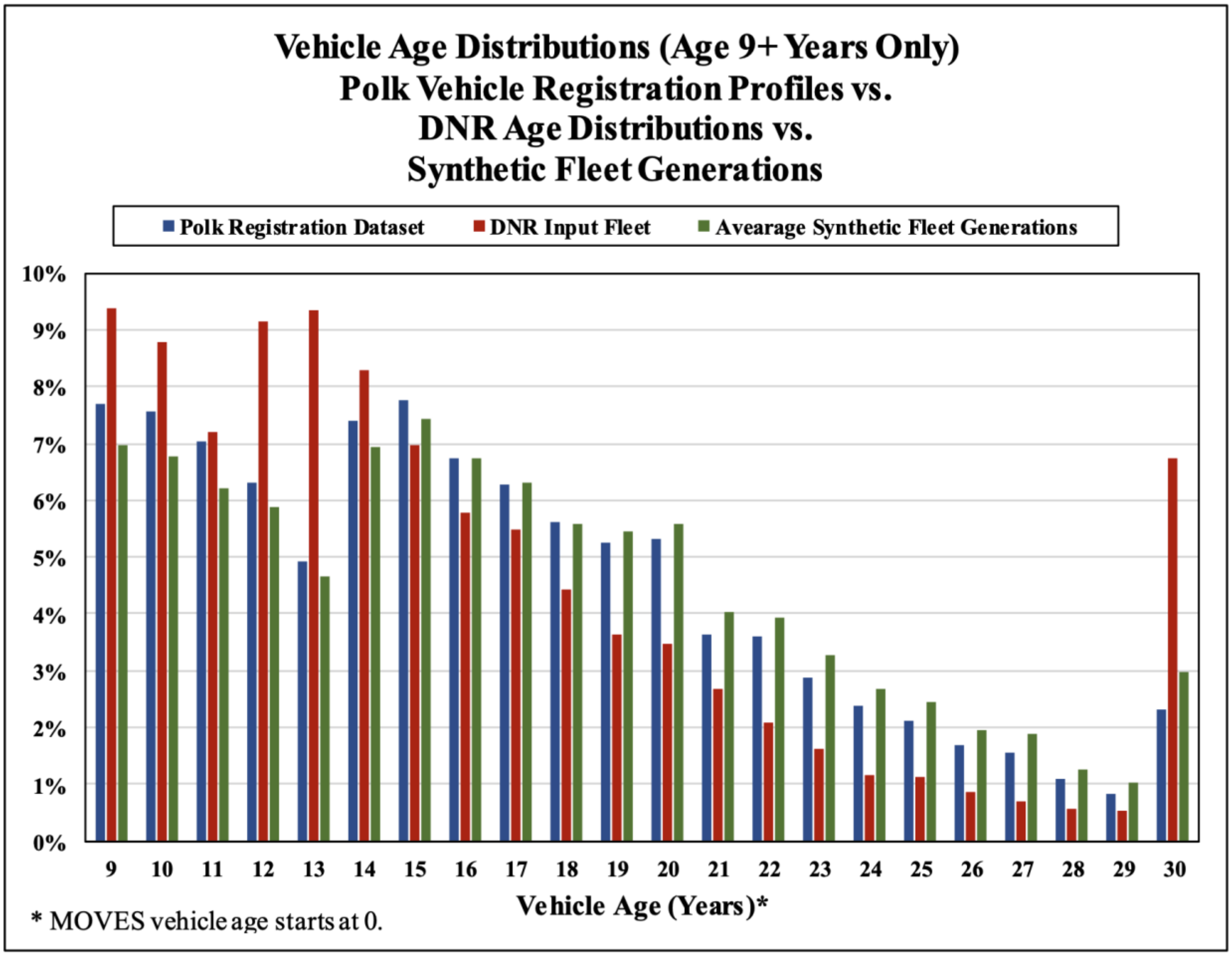

Figure 7. Vehicle age distributions of synthetic fleet vs. DNR inputs vs. Polk® data.

The total VMT per day for the metro Atlanta region was approximately 73,910,118.6 miles, and a cross-reference of the vehicle ownership results indicated an average mileage of approximately 19.26 miles per vehicle per day, which does not deviate from what the team would consider reasonable. The generated VMT distributions by source type and by vehicle age is shown in Figure 8. The small ranges of the CIs indicate that generated VMT distributions were stable, and the comparison against the RMAR indicated a general good alignment with the desired VMT distributions, despite the deviations in the generated fleet ownership profiles.

However, notably lower distributions were still observed for vehicles older than 13 years (model year of 2008) and those older than 25 years. This was due to the much lower ownership of these model years in the fleet ownership data and their distributions in the paired households. The random allocation of vehicles could only adjust VMT distributions based on households that were assigned with multiple vehicles. For those households that were allocated only one vehicle, that vehicle was allocated to all trips, irrespective to its weight. This again indicates that access to the official state vehicle registration database appear to be essential for developing representative synthetic household and fleet generators.

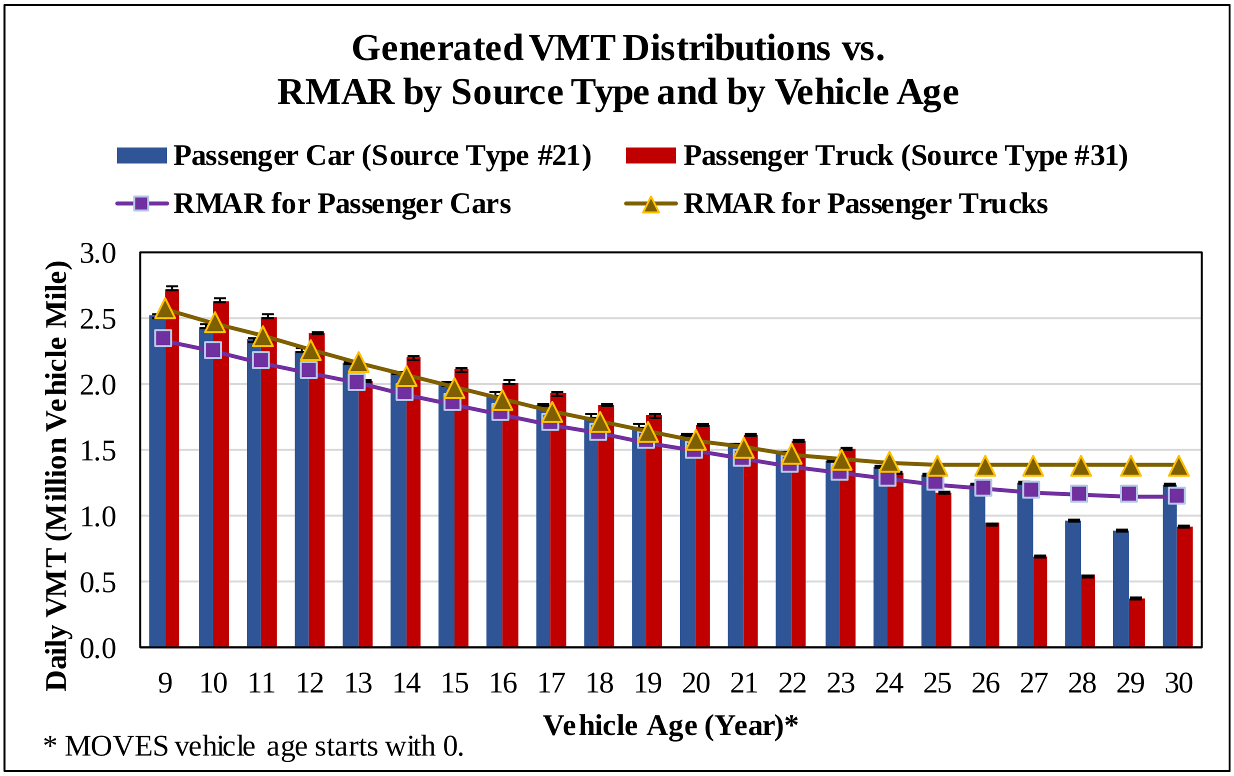

Figure 8. Generated VMT distributions vs. RMAR by source type and by vehicle age.

3.2 Network Emission Inventories

The differences in average CO emissions between the two scenarios were visualized link-by-link, with the 95% confidence intervals (CIs) also plotted for clarity. Links with no emissions, such as managed lanes that remain closed during specific hours, were excluded from the visualization to maintain focus on the active and relevant links. As a representative sample, the differences for the time slot from 9 AM to 10 AM were plotted, as shown in Figure 9 (average differences) and Figure 10 (confidence intervals of differences), and it was observed that similar patterns and findings were consistent across other hours. Comprehensive plots for all hours have been provided in Appendix F for those interested in a deeper dive.

To interpret the plotted CIs: if both upper and lower bounds of the CI (for a particular link) indicate a negative difference, it suggests with 95% confidence that the county-based approach resulted in an underestimation of emissions for that link (negative bias). Conversely, if both bounds are positive, it indicates an overestimation with 95% confidence (positive bias). Figure 11 illustrates those links with both CI bounds of the differences either positive or negative at 9 AM to 10 AM. In terms of spatial distribution of these differences, a few patterns emerged. Negative biases, or underestimations, were predominantly observed on busy arterial links and a few isolated links near the periphery of the metro area. On the other hand, positive biases, or overestimations, were primarily seen on the outer links, with a minor presence on the interstate highways. Links situated around the downtown area or town centers in the metro area (e.g., Sandy Springs) generally exhibited smaller differences, and links on the periphery or the edge of the metro area displayed larger differences. When focusing solely on the "biased" links, or those with CIs bounds of differences that are both positive or negative, the observed trends remained consistent. Positive biases were predominantly seen on the outer links and highways, while negative biases were more common on local arterial roads.

The county-based approach essentially aggregates the impact of average speed on emission rates, and this aggregation leads to loss of resolution in the input. This spatial distribution of biases is likely to be associated with the sensitivity of emission rates against speed, and the variabilities in the sensitivity across source types, model years, and facility types (restricted highway vs. unrestricted highway). The distinct patterns in the ABM-predicted fleet composition and travel speed across different links (e.g., certain links might consistently operate under free-flow conditions especially those near the edge of the modeled network, while other links might predominantly be used by specific source types, such as passenger cars) intersects with the varied sensitivities which could lead to the biases.

Very few links exhibit consistent biases across all hours, as shown in Figure 12. A notable observation was the consistent underestimation of emissions on several links near the Six Flags area across all hours. This persistent bias suggests specific local factors or conditions that might not be adequately captured by the county-based approach. Furthermore, the northern interstate highways of Atlanta, specifically I-75, I-85, GA-400, and parts of I-285, consistently exhibited positive biases. This indicates a tendency for the county-based approach to overestimate emissions on these major transportation routes. However, a wide span of the links was demonstrated to have at least one hour of biases, as shown in Figure 13. Please note positive and negative biases do not necessarily cancel each other off, especially for hot-spot analysis where peak hour traffic is often used to model the emissions.

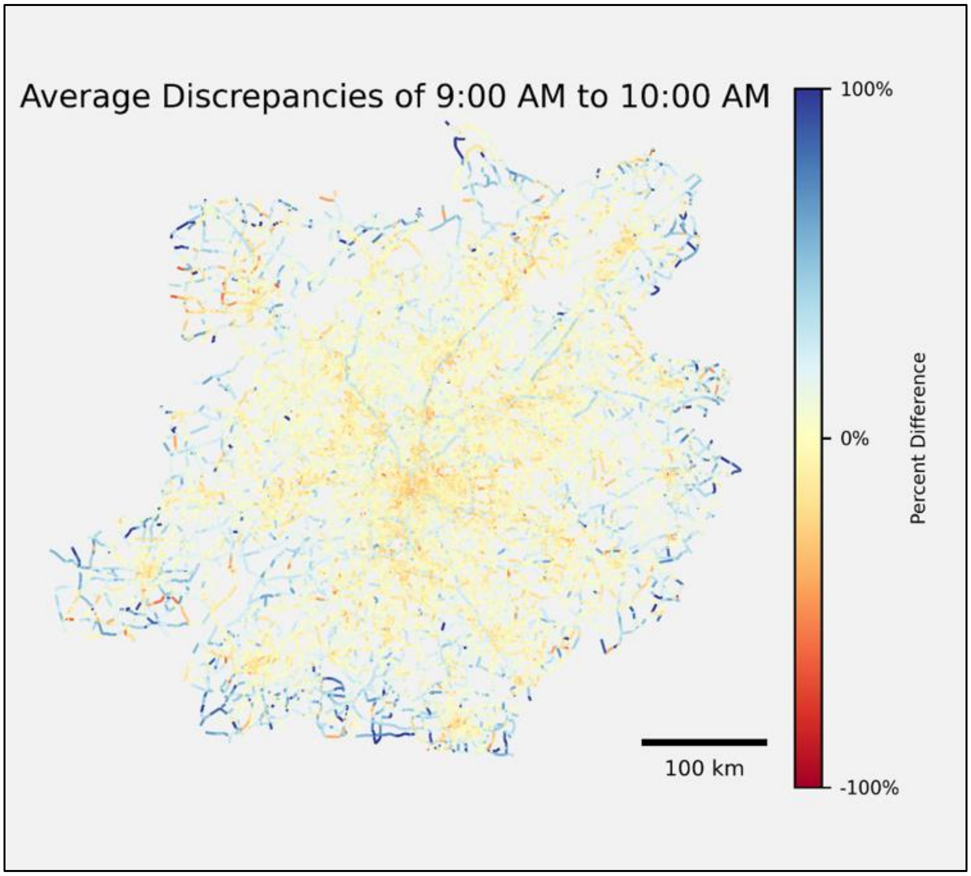

Figure 9. Link-by-link average differences between the scenarios, 9 AM to 10 AM.

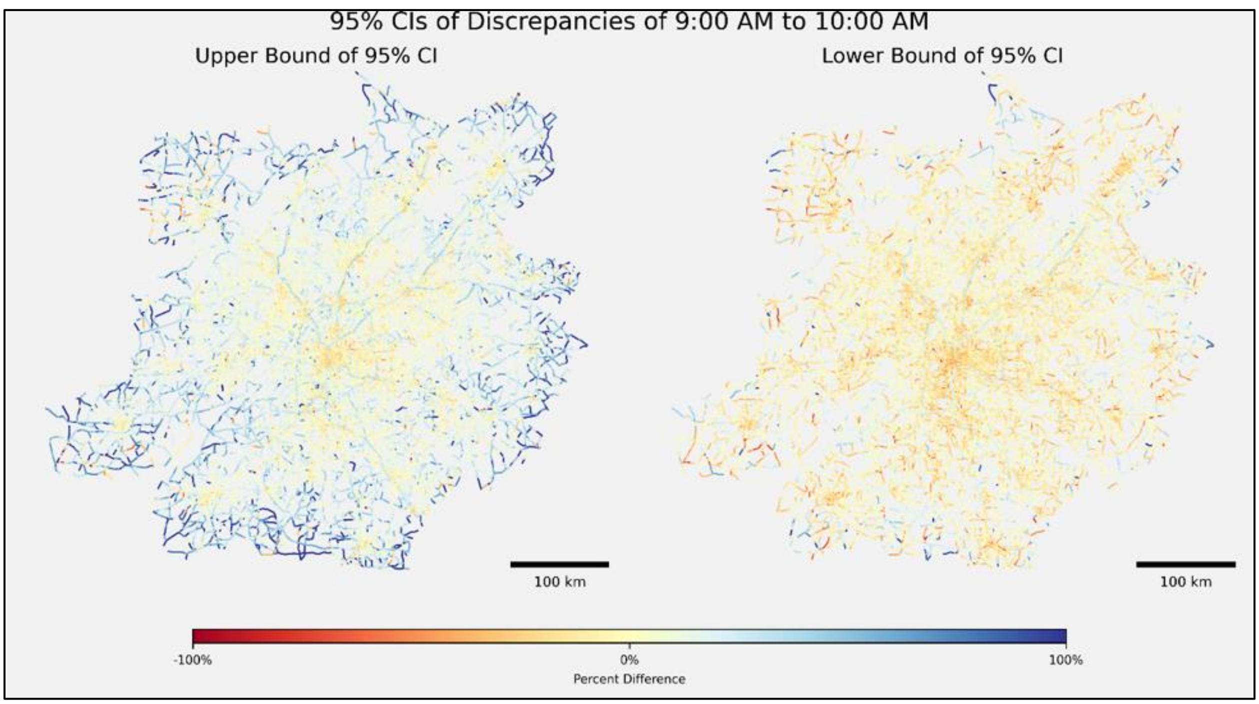

Figure 10. Link-by-link CIs of difference between the scenarios, 9 AM to 10 AM.

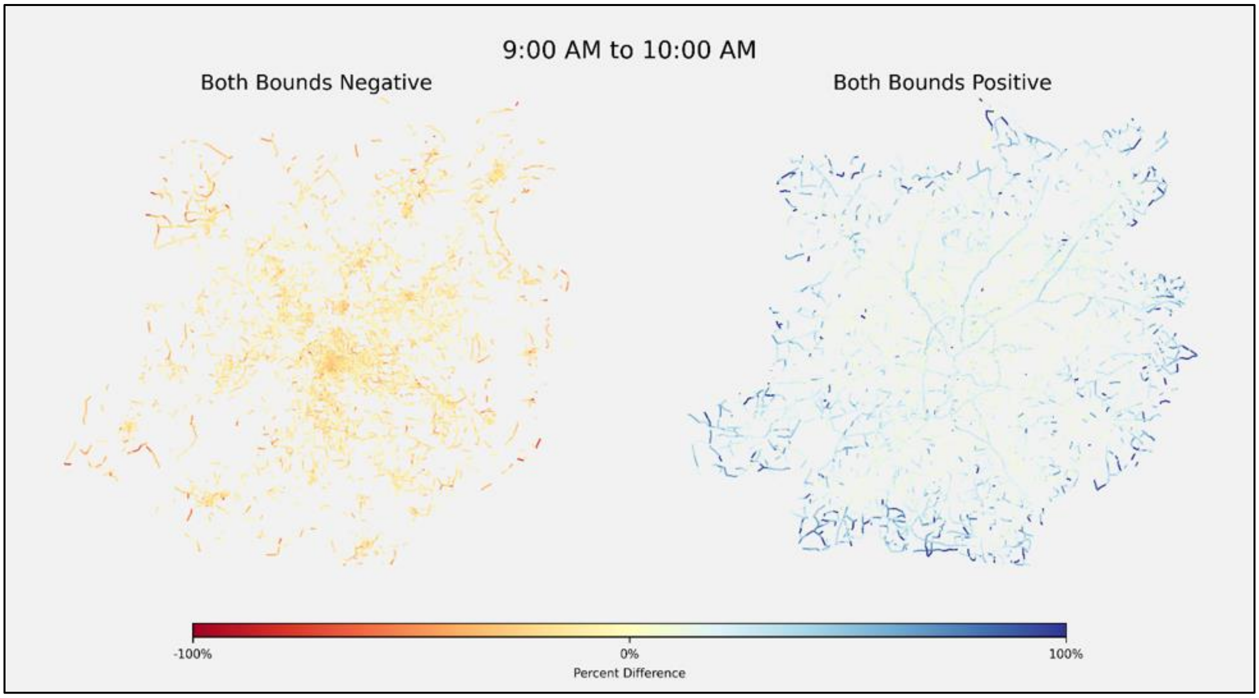

Figure 11. Links with both CI bounds negative or positive, 9 AM to 10 AM.

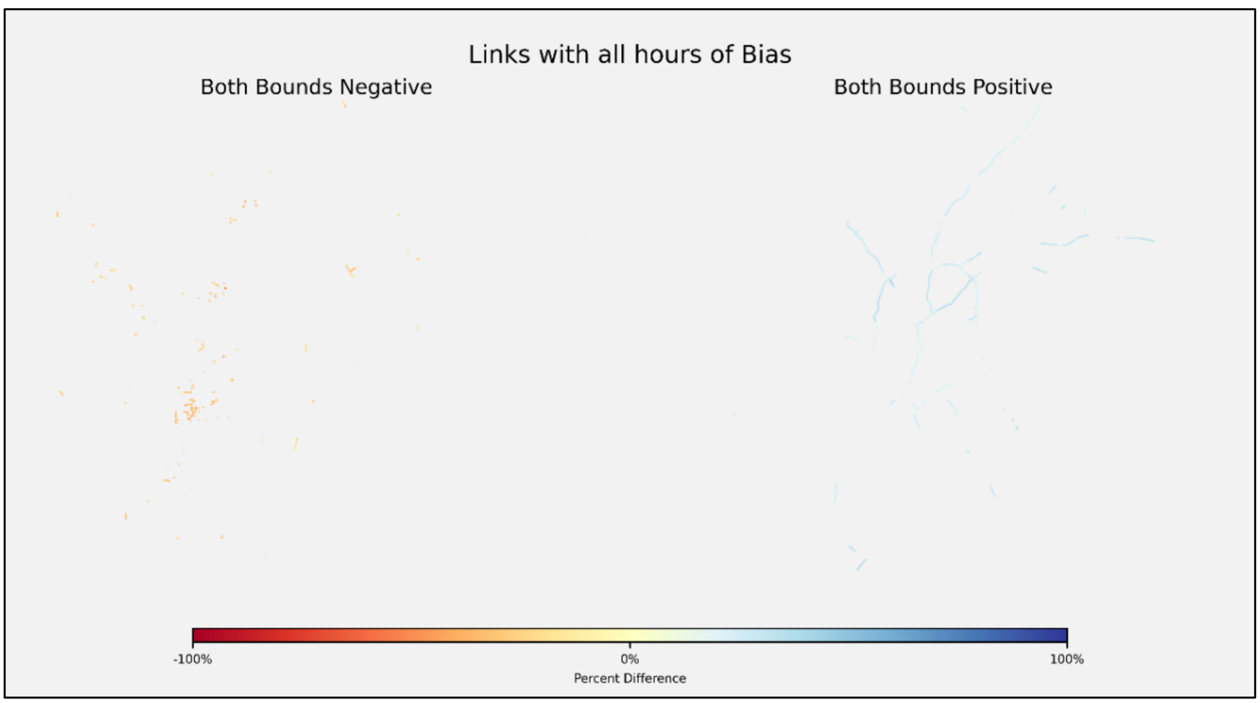

Figure 12. Roadway links with all CI bounds negative or positive across all hours.

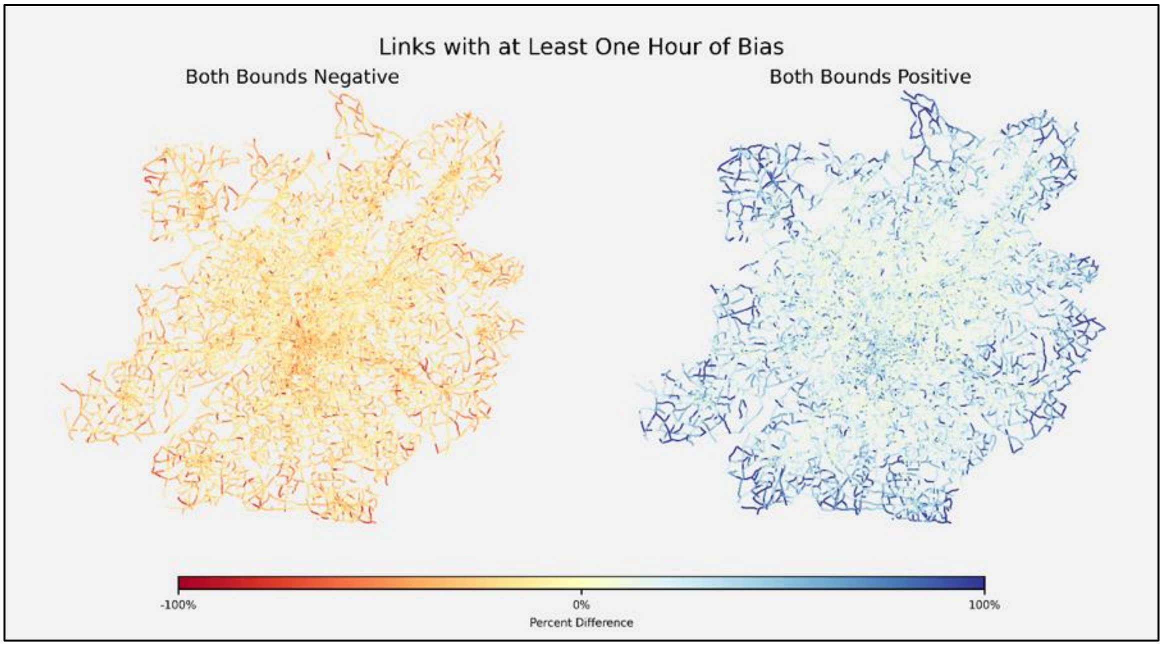

Figure 13. Roadway links with both CI bounds negative or positive for at least one hour.

3.3 Emission Results by Tour Purpose and by Demographic Groups

The emission results aggregated by half-hour bin are shown in Figure 14, and the error bars indicates the bonds of the 95% CIs. Similar to the other results, the tight ranges of the CIs indicate stable predictions of the synthetic fleet. The emission trends by time of day aligns with the anticipation of the variances in average speed and traffic volumes (e.g., peak hours vs. off peak hours).

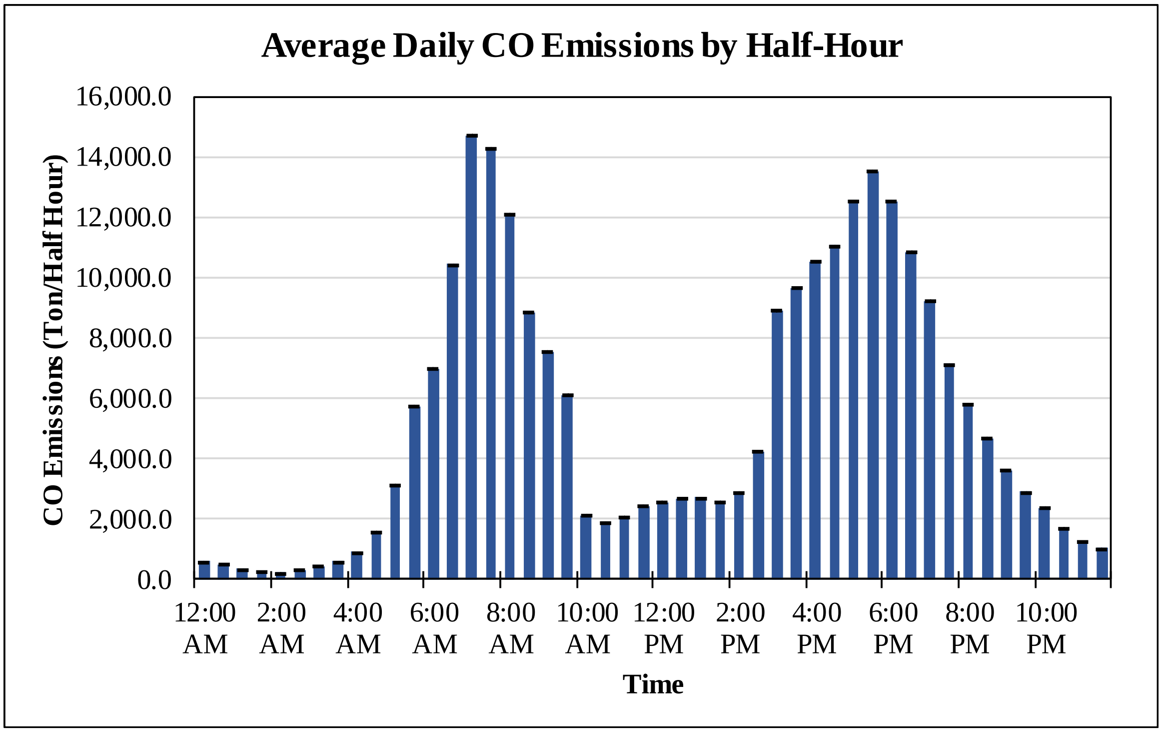

Figure 14. Average daily CO emissions by half-hour.

The emission predicted by the synthetic fleet generations were aggregated by tour purpose, as shown in Figure 15. The white-collar work trips stand out with the highest CO emissions, (significantly higher than other categories), followed by blue-collar work trips. These large emissions are not surprising given the prevalent demands of commute trips and that these trips predominantly occur during peak hours (lower speed and larger emission rates). This indicates the importance of sustainable commuting solutions, such as carpooling or transit incentives.

Interestingly, the emissions from other discretionary activities and other maintenance activities rank closely to each other, only next to white-collar and blue-collar work tours. This equivalency indicates that both essential (maintenance) and non-essential (discretionary) activities contribute similarly to CO emissions.

Emissions associated with pre-school education tours are almost double those linked to university education tours. This difference could be attributed to factors like the frequency of trips, the distance traveled, or the type of vehicles predominantly used by parents or guardians for preschool drop-offs and pick-ups (even for metro Atlanta with various college campuses).

The emissions from dining out from home and dining out from work are relatively close in value, and both are considerably lower than most work-related categories. This suggests that while eating out does contribute to CO emissions, its impact is less pronounced than regular work commutes.

The findings are further corroborated when examining the average CO emission by time of day and by tour purpose, as illustrated in Figure 16. Emissions from work-related categories predominantly cluster around morning and afternoon peak hours, and a similar temporal pattern is observed for discretionary activities and educational trips. The maintenance activities display a more consistent emission profile throughout the day time, which indicates less variation across the half-hour bins. The eat-out-at-work trips tend to have higher emissions during off-peak hours, and eat-out trips have higher emissions during evening trips. These observations align intuitively with the typical temporal characteristics associated with each tour purpose.

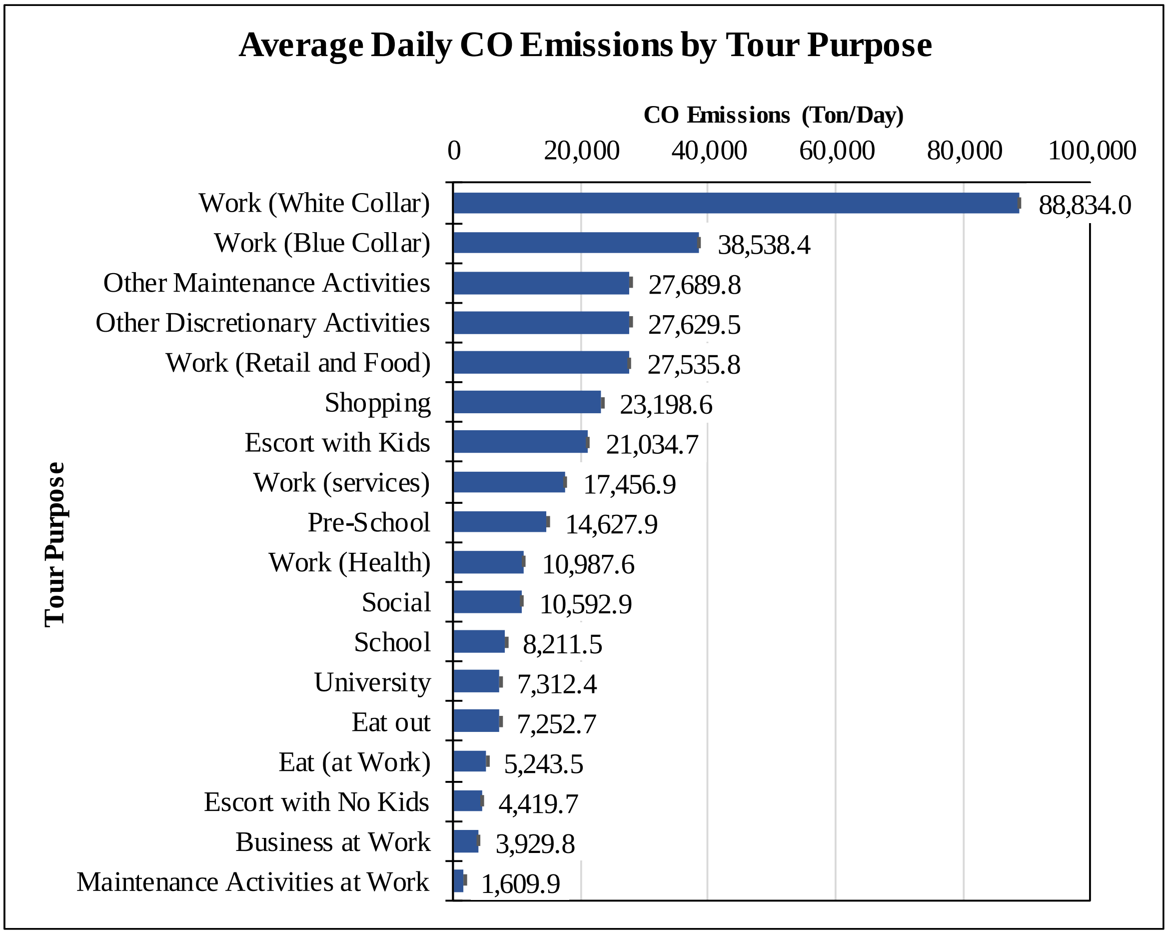

Figure 15. Average daily CO emissions by tour purpose.

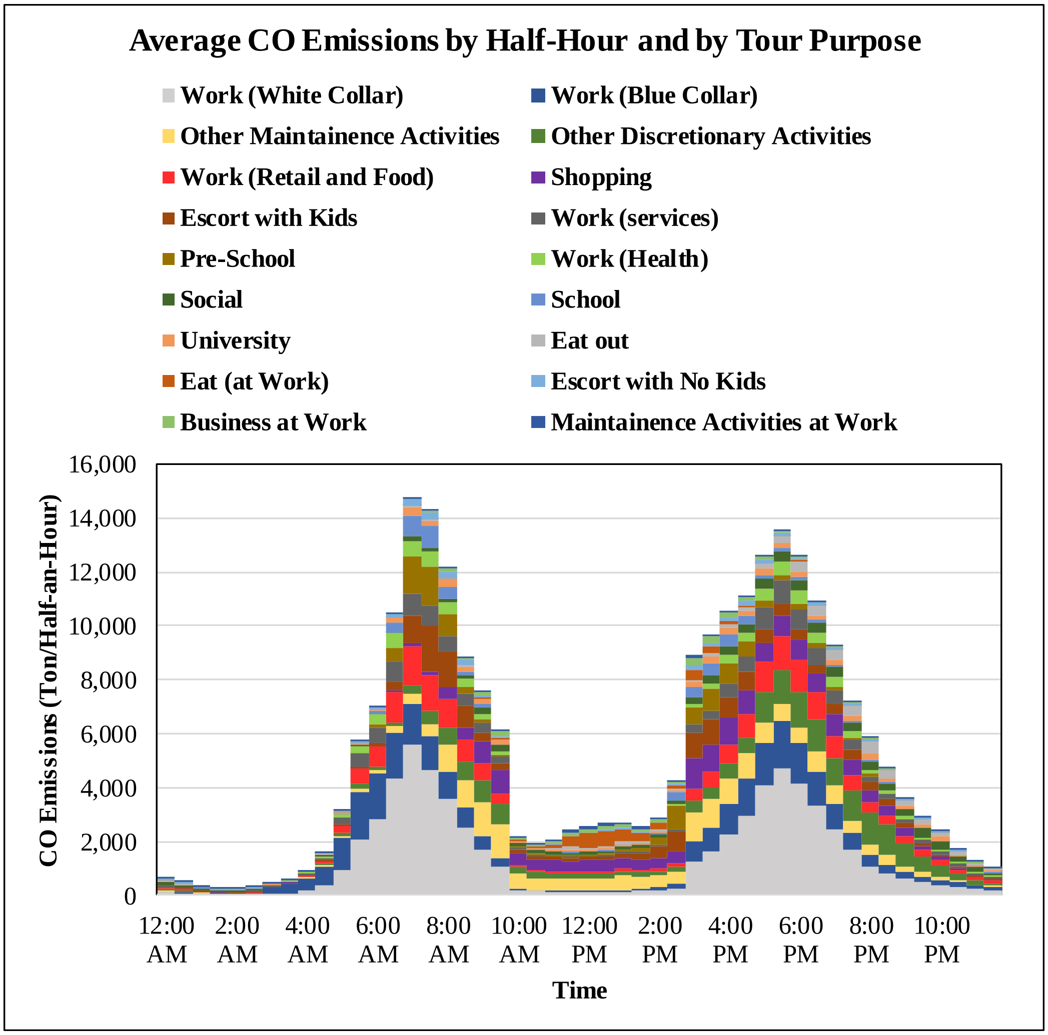

Figure 16. Average CO emissions by half-hour and by tour purpose.

The emission results predicted from on the synthetic fleet were based on individual vehicle and link pairs aggregated to the household level. The household emission results were further aggregated to the 16 demographic groups to provide average emission results per hour (workday) as shown in Table 5. The rationale for creating the demographic groups can be found in section 2.2.1.

Demographic Group #1 (household with no vehicles) were found to emit the least among all groups, which is not surprising given that they have limited access to automobiles (e.g., ride-sharing services). This was accommodated by the significantly fewer trips in the ABM for Group #1. Group #2, Group #3 and Group #6 followed closely, and generated slightly more than Group #1 but significantly lower than the other groups. These groups are either low-income or non-working households, and are likely to have fewer trips (less demand) than other households.

The demographic groups with higher emissions (i.e., Groups #4, #5, #7, #12, and #13) were either one-person households with high income (Group #5), or working households with moderate incomes. The highest emissions were observed at Group #12 and #13, which exhausted more than double of the average emission than either Groups #1, #2, #3, or #6. These groups were estimated to have a significant commute demand yet less able to regularly update of their vehicles relative to the higher-income groups.

The patterns across the demographic groups could be related to the variances in travel demands across households (e.g., commute vs. entertainment trips), and could also be related to the variability of spatial distributions across demographics (e.g., shorter vs. longer trips, or low-speed trips vs. high-speed trips). The team is working on analyzing the emissions by trip purpose, and by demographic groups, and coupling the analysis results with the spatial characteristics of the demographics.

A further enhanced equity assessment can be performed by integrating the emission results with population exposure modeling, and the team is working on expanding the modeling efforts another on-going research project.

Table 5. Average hourly emission predictions by demographic group.

Group #	Own Vehicles	Low Income	HH Size	Annual Income	Workers	Vehicles	With Kid(s)	Single Parent w/kid(s)	Emission Per Household (mg/hour)	Emission Per Person (mg/hour)
1	No	Any	Any	Any	Any	0	Any	Any	1,010.6	491.4
2	Yes	Yes	Any	$0 - $25k	Any	1+	Any	Any	1,821.5	890.0
3	Yes	No	1	$25k - $60k	0	1+	Any	Any	817.4	817.4
4	Yes	No	1	$25k - $60k	1	1+	Any	Any	2,297.5	2,297.5
5	Yes	No	1	$60k+	0 or 1	1+	Any	Any	2,120.1	2,120.1
6	Yes	No	2+	$25k - $60k	0	1+	Any	Any	2,437.3	992.4
7	Yes	No	2	$25k - $60k	1+	1+	Any	Any	4,008.2	2,004.1
8	Yes	No	3+	$25k - $60k	1+	1+	Any	Yes	5,381.9	1,566.5
9	Yes	No	3+	$25k - $60k	1+	1+	Any	No	6,548.6	1,602.4
10	Yes	No	2+	$60k - $120k	0 or 1	1+	Yes	Any	7,079.5	1,838.7
11	Yes	No	2+	$60k - $120k	0 or 1	1+	No	Any	3,889.2	1,733.8
12	Yes	No	2	$60k - $120k	2	1+	Any	Any	5,118.9	2,559.5
13	Yes	No	3	$60k - $120k	2+	1+	Any	Any	7,669.8	2,556.6
14	Yes	No	4+	$60k - $120k	2+	1+	Any	Any	8,637.8	1,835.5
15	Yes	No	2+	$120k+	0 or 1	1+	Any	Any	4,832.3	1,663.0
16	Yes	No	2+	$120k+	2+	1+	Any	Any	6,586.0	1,958.8
Total	All	All	All	All	All	All	All	All	4,374.6	1,696.6

4. Conclusions and Future Work

In this research, the team proposed a synthetic household and fleet generator by integrating three datasets: Epsilon® demographic data, R.L. Polk® vehicle registration data, and ARC's ABM2020. The integration process paired households and vehicles based on addresses and Traffic Analysis Zones (TAZs) through a series of random draws, and the synthetic fleet was generated by allocating vehicles to the trips based on MOVES's RMAR to capture the predicted vehicle ownership and usage patterns. Emissions of CO were modeled and compared in two distinct scenarios. The first scenario used a refined trip-based approach that employed the synthetic fleet applied to each trip, modeling emissions for each individual trip by following the assigned synthetic vehicle along each ABM-predicted route. The second scenario adopted a broader county-based approach which applies average emission rates by source type and by model year extracted from the first scenario to derive link-by-link emission inventory by county. After 1,000 primary iterations, average metrics of fleet ownership, VMT distributions, emissions, and differences between the two scenarios were derived, and the 95% confidence intervals were determined using Bootstrap techniques.

The synthetic fleet generation produced an average of 3,836,602 vehicles for metro Atlanta, with a 95% CI ranging narrowly from 3,836,204 to 3,836,965 vehicles, which suggested a stable generation process with minimal variance. The vehicle ownership by type and age was observed to be stable, which demonstrated larger shares of passenger cars over trucks, except for vehicles aged 18 to 20 years (which tend to remain in the fleet longer). However, a noteworthy observation is the amplified representation of older vehicles in the synthetic fleet, even compared to the original Polk® data. The team recommends using the state vehicle registration data in future analyses to avoid potential biases that appear to exist in the licensed data used in this research.

The analysis of CO emissions differences between the two scenarios revealed distinct spatial patterns across the metro Atlanta region, given the licensed registration data. Downtown links consistently showed generally smaller differences (and more negative biases), while peripheral links exhibited larger variances, with a relatively widespread overestimation on major northern interstates like I-75, I-85, GA-400, and parts of I-285. Additionally, specific areas, such as the vicinity of Six Flags, consistently underestimated emissions across all hours. Significant differences across one hour modeled predictions were observed for a large area of the metro area. Despite any potential issues with representativeness of the licensed registration data, the results indicate a clear need to very cautious in applying county-based average emission rates to individual roadway links. Roadway fleet composition, which is based upon local fleet ownership and local travel patterns, especially between local arterials and freeways used for commute and business trips, is very important for accurately reflecting sub-regional and regional energy use and emissions. Hence, this accuracy will also significantly impact health impact assessments using predicted pollutant concentration immediately downwind from transport facilities.

The temporal emission trends align with expected traffic volume and speed variations, particularly during peak hours. When examining emissions by tour purpose, white-collar work trips significantly lead in CO emissions, followed by blue-collar work trips. Notably, emissions from discretionary and maintenance activities are comparable, suggesting both essential and non-essential activities have similar CO emission contributions. Pre-school tours exhibit almost double the emissions of university tours, likely due to trip frequency, distance, or vehicle source types used for preschool commutes. Emissions from eating out and eating at work are relatively modest, especially when compared to work-related emissions. A deeper dive into the temporal emission patterns reveals that work-related emissions peak during rush hours, while maintenance activities have a more uniform distribution throughout the daytime.

When aggregating emissions to demographic groups, Group #1 (households without vehicles) unsurprisingly emits the least, given their limited vehicular access. Groups #2, #3, and #6, (primarily low-income or non-working households) also exhibit lower emissions. In contrast, Groups #4, #5, #7, #12, and #13 (high-income single-person households and moderate-income working households), exhibit higher emissions. The highest emissions are seen in Groups #12 and #13 (working households with moderate incomes), emphasizing the impact of commute demands and vehicle updating frequency. These emission patterns across demographic groups can be influenced by variations in travel demands and spatial distributions of activity. The team is further exploring these emissions in relation to spatial demographics and is considering integrating results with population exposure modeling in future research.

The modeling approach presented in this report can be applied to other metropolitan areas or regions, provided that similar detailed data for fleet ownership and travel activity are available (and demographics that can be paired with vehicle ownership data). However, the team also recommends that other regions consider exploring the use of vehicle ownership data embedded in local travel demand models when detailed vehicle ownership data are not available from external sources. Allocating vehicle ownership at the TAZ level, and then using detailed travel paths for each household, will still represent on-road fleet composition better (in terms of the temporal and spatial variances) than randomly assigning vehicles across an entire county or using county-wide averages in energy and emissions modeling. The proposed modeling framework can also be applied to fleets that include electric and hybrid vehicles, given proper geographic and demographic data. Incorporating electric vehicles will lead to even greater differences in emissions modeling results given the vehicle procurement and use differences across household demographics (Dai, 2023; Dai, et al., 2022). That is, electric vehicle ownership and use characteristics (i.e., assignment of vehicle activity to the vehicles within a household) are closely tied to demographics (such as income levels, charging infrastructure availability, and ownership costs). Because hybrids and battery-electric vehicles use significantly less energy and produce much lower emissions than conventional vehicles, corridor-level energy use and emission rates are very sensitive to the percentage of EVs in the fleet.

This study developed a synthetic household and fleet generator by integrating demographic information from ARC's ABM2020 and Epsilon® 2022 marketing database with vehicle ownership information from the R.L. Polk® Vehicle Registration Database. The team identified what appears to be a significant bias in the licensed vehicle data sets toward the presence of older vehicles in the fleet, when compared to the registration mix used by the Department of Natural Resources based upon their access to the state registration database. The licensed data bias appears significant against the presence of newer vehicles (newer than 8 years old) in the fleet. The research team faced some challenges to converge on VMT distributions based on RMAR from MOVES, but the methods developed are sound. The basic findings that local registration mix significantly affects sub-regional emissions predictions also remain clear, despite any potential presence in vehicle age in the licensed data. The research team also concludes that an updated version of the synthetic household and fleet generator created for this research should be prepared as soon as possible, using the methods outlined in this research, once full access to the state vehicle registration database can be arranged and proper protocols are put in place to use these personally-identifiable data.

References

Bachman, W., J. Granell, R. Guensler, and J. Leonard (1998). “Research Needs in Determining Spatially Resolved Subfleet Characteristics.” Transportation Research Record. Number 1625, pp. 139-146. Transportation Research Board. Washington, DC. 1998.

Brière, C., Obozinski, G., and Vayatis, N. (2021). The Permutation Importance Criterion: Model-Independent Variable Importance Measure. Machine Learning, 110(6), 1271–1300. https://doi.org/10.1007/s10994-021-05981-3.

Choo, S., and Mokhtarian, P. L. (2004). What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. Transportation Research Part A: Policy and Practice, 38(3), 201-222.

Dai, Z., Liu, H., Rodgers, M.O., and Guensler, R. (2022). Electric vehicle market potential and associated energy and emissions reduction benefits. Applied Energy, 307, 1179-1189.

Dai, Z. (2023). An Assessment of EV Adoption and Potential Growth under Evolving Techno-Policy Scenarios. Dissertation. School of Civil and Environmental Engineering, Georgia Institute of Technology. August 2023.

Georgia EPD. (2023). Georgia AERMET Meteorological Data. Retrieved January 8, 2023, from https://epd.georgia.gov/air-protection-branch-technical-guidance-0/air-quality-modeling/georgia-aermet-meteorological-data

Granell, J., R. Guensler, and W. H. Bachman (2002). “Using Locality-Specific Fleet Distributions in Emissions Inventories: Current Practice, Problems and Alternatives.” Published in the CD-ROM Proceedings of the 79th Annual Meeting of the Transportation Research Board. Washington, DC. January 2002.

Guensler, R., Liu, H., Lu, H., Chang, C. “Chris,” Dai, Z., Xia, T., Fu, Z., Liu, D., Kim, D., Zhao, Y., and Guin, A. (2022). Georgia Express Lane Corridors Vehicle Occupancy and Throughput Study 2018-2022. Volume I: Vehicle and Person Throughput Analysis: Before and After the I-75 Northwest Corridor and I-85 Express Lanes Extension. Report Prepared for GA State Road and Tollway Authority (SRTA).

Guensler, R., H. Liu., Y. Xu, A. Akanser, D. Kim, M.P. Hunter, M.O. Rodgers (2017). "Energy Consumption and Emission Modeling of Individual Vehicles." 10.3141/2627-11. Transportation Research Record. Number 2627. pp. 93-102. National Academy of Sciences. Washington, DC. 2017.

Guensler, R., and Leonard II, J. D. (1995). A Monte Carlo Technique for Assessing Motor Vehicle Emission Model Uncertainty. Transportation Congress, Volume 2; American Society of Civil Engineers, 2098–2115. New York, NY, U.S.

Guensler, Randall, Liu, H., Xu, X., and Xu, Y. “Ann.” (2016). MOVES-Matrix: Setup, Implementation, and Application. 95th Annual Meeting of the Transportation Research Board. Washington, D.C., U.S.

Guensler, Randall, Liu, H., Xu, Y. (Ann), Akanser, A., Kim, D., Hunter, M. P., and Rodgers, M. O. (2017). Energy Consumption and Emissions Modeling of Individual Vehicles. Transportation Research Record: Journal of the Transportation Research Board, 2627, 93-102. https://doi.org/10.3141/2627-11

Kim, D., Liu, H., Rodgers, M. O., and Guensler, R. (2020). Development of Roadway Link Screening Model for Regional-level Near-road Air Quality Analysis: A Case Study for Particulate Matter. Atmospheric Environment, 237, 117677. https://doi.org/10.1016/j.atmosenv.2020.117677

Kim, D., Liu, H., Xu, X., Lu, H., Wayson, R., Rodgers, M. O., and Guensler, R. (2020). Distributed computing for region-wide line source dispersion modeling. Computer-Aided Civil and Infrastructure Engineering, 1–15. https://doi.org/10.1111/mice.12639

Li, H., Liu, H., Xu, X., Xu, Y. "Ann," Rodgers, M. O., and Guensler, R. (2016). Emissions Benefits from Reducing Local Transit Service Deadheading: An Atlanta Case Study. 95th Annual Meeting of the Transportation Research Board. Washington, D.C., U.S.

Li, H., Wang, Y., Xu, X., Liu, H., Guin, A., Rodgers, M. O., Hunter, M., Laval, J.A., Khaled, A., and Guensler, R. (2017). Assessing the Time, Monetary, and Energy Costs of Alternative Modes. 97th Annual Meeting of Transportation Research Board. Washington, D.C., U.S.

Liu, H., Guensler, R., Lu, H., Xu, Y., Xu, X., and Rodgers, M. O. (2019). MOVES-Matrix for High-Performance On-Road Energy and Running Emission Rate Modeling Applications. Journal of the Air & Waste Management Association, 10962247.2019.1640806. https://doi.org/10.1080/10962247.2019.1640806

Lu, H., Kim, D., Zhao, Y., Liu, H., Guin, A., Laval, J. A., Hunter, M., Rodgers, M. O., and Guensler, R. (2021). Evaluating of Population Exposure to Traffic-related Air Pollution across Demographic Characteristics: Activity-based Model with Path Retention and Streamlined Dispersion Modeling in Atlanta, GA. CARTEEH 2nd Transportation, Air Quality, and Health Symposium.

Lu, H., Liu, H., Xia, T., Guin, A., Rodgers, M. O., and Guensler, R. (2021). Investigations of MOVES and AERMOD Uncertainty: Impact of Temporal- and Spatial-Aggregation of On-Road Operating Conditions on Emissions and Dispersion Modeling. CARTEEH 2nd Transportation, Air Quality, and Health Symposium.

Lu, Hongyu, Liu, H., Guin, A., Michael O. Rodgers, and Randall Guensler. (2020). Sensitivity Analysis of Emissions Estimation when Cluster Analysis is used to Trim Machine Vision Traffic Data. 99th Annual Meeting of the Transportation Research Board. Washington, D.C., U.S.

Lu, Hongyu, Liu, H., Xia, T., Angshuman Guin, Rodgers, M. O., and Randall Guensler. (2021). Investigation of MOVES Project-Level Uncertainty: Impact of Temporal- and Spatial-Aggregation of Onroad Operating Conditions on Emission Rate Estimates. 100th Annual Meeting of the Transportation Research Board. Washington, D.C., U.S.

National Highway Traffic and Safety Administration (NHTSA) (2006), Vehicle Survivability and Travel Mileage Schedules. DOT HS 809 952. http://wwwnrd.nhtsa.dot.gov/Pubs/809952.pdf

U.S. EPA (2021). Population and Activity of Onroad Vehicles in MOVES3. EPA Report EPA-420-R-21-012. https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P1011TF8.pdf

U.S. EPA (2016). Population and Activity of On-road Vehicles in MOVES2014. EPA Report EPA-420-R-16-003a. https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100O7PS.pdf

Xu, X., Liu, H., Anderson, J. M., Xu, Y., Hunter, M. P., and Rodgers, M. O. (2016). Estimating Project-Level Vehicle Emissions with VISSIM and MOVES-Matrix. 95th Annual Meeting of the Transportation Research Board. Washington, D.C., U.S.

Xu, X., H. Liu, Y. Xu, M.O. Rodgers and R. Guensler (2018a). Regional Emission Analysis with Travel Demand Models and MOVES-Matrix (18-05363). 97th Annual Meeting of the Transportation Research Board (presentation only, full paper review, extended abstract in proceedings). Washington, DC. January 2018.

Xu, X., H. Liu, H. Li, M.O. Rodgers, R. Guensler (2018b). "Integrating Engine Start, Soak, Evaporative, and Truck Hoteling Emissions into MOVES-Matrix." DOI: 10.1177/0361198118797208. Transportation Research Record. Washington, DC. 2018.

Xu, Y., Li, H., Liu, H., Rodgers, M. O., and Guensler, R. L. (2017). Eco-driving for Transit: An Effective Strategy to Conserve Fuel and Emissions. Applied Energy, 194, 784–797. https://doi.org/10.1016/j.apenergy.2016.09.101

Zhao, Y. (2021). Distributive Justice Impact Assessment using Activity-based Modeling with Path Retention. Ph.D. Dissertation. Georgia Institute of Technology, School of Civil and Environmental Engineering.

Zhao, Y. and R. Guensler (2019). Gender Differences in Predicted Travel Activity in Atlanta using an Activity Based Model with Path Retention. 6th International Conference on Women's Issues in Transportation. Irvine, CA. September 10-13, 2019.

Zhao, Y.; X. Xu, H. Liu, H. Li, C. Wang, A. Guin, M.O. Rodgers, and R. Guensler (2019). Using ABM15, Path Retention, and MOVES-Matrix to Assess Differences in Energy Use across 16 Demographic Groups in Atlanta. TRB ADA10 Fresh Ideas for Advancing Equity. 98th Annual Meeting of the Transportation Research Board. Washington, DC. January 2019.

Data Summary

As described in this report, the team modeled the emission of CO by roadway link based on massive MOVES 2014b model runs, for both synthetic fleet results and county-level results.

Products of Research

The traffic volume and speed data used in this study were derived from the Activity-based Model (ABM) by Atlanta Regional Commission (ARC), and the fleet composition profiles were developed by paring the ABM households with Epsilon® and Polk® datasets. Under the data user agreement, interested parties need to obtain these data from the corresponding contractors. The energy and emission rate matrices applied are public domain and can be found at https://zenodo.org/records/13864619.

Data Format and Content

The format and content of the MOVES-Matrix (MOVES2014b) data sets are documented in the NCST MOVES-Matrix overview and training documents at https://github.com/gti-gatech/moves_training/.

Data Access and Sharing

The MOVES-Matrix data are open source and can be downloaded and freely shared from the link provided above.

Reuse and Redistribution

The MOVES-Matrix data are open source can be downloaded, used, and freely redistributed using the link provided above. The following citation should be used:

> Lu, H. (2024). Synthetic Fleet Generation and Vehicle Assignment to Synthetic Households for Regional and Sub-regional Sustainability Analysis (Version 093024) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13864619

Appendix A: QA/QC of the Epsilon Vehicle Ownership Database

The new Epsilon dataset provides vehicle ownership data for 770 thousand households record (empty vehicle records for the rest of the dataset) with a total of 1.2 million vehicles. Given the high vehicle ownership of metro Atlanta, the team assumes that the empty vehicle information does not necessarily indicate these households do not have vehicles.

The team first performed a comparison of the vehicle records vs. the Polk® vehicle registration in terms of vehicle age distribution and source type distribution, and the results indicated they were not identical, but were similar. The comparison between Epsilon vehicle ownership data vs. the generated synthetic fleet indicated that 393,041 Epsilon households (more than 50%) have same vehicle ownership with the regional synthetic fleet (which essentially is based on pairing the addresses between Polk® and Epsilon datasets) that were either 1) 100% same with the synthetic fleet results, including number of vehicles, and the make, model, and model year of every vehicle, or 2) the Epsilon vehicle ownership is a subset of the synthetic fleet results.

The team found the Epsilon vehicle ownership still worth looking at for future studies, given that 265,038 of the Epsilon households have more vehicles than what were generated in the synthetic fleet. The vehicle age and source type distribution comparison across the Epsilon vehicle ownership data, the Polk® vehicle registration data and the regional synthetic fleet results are shown in Figure 17, and Figure 18, respectively.

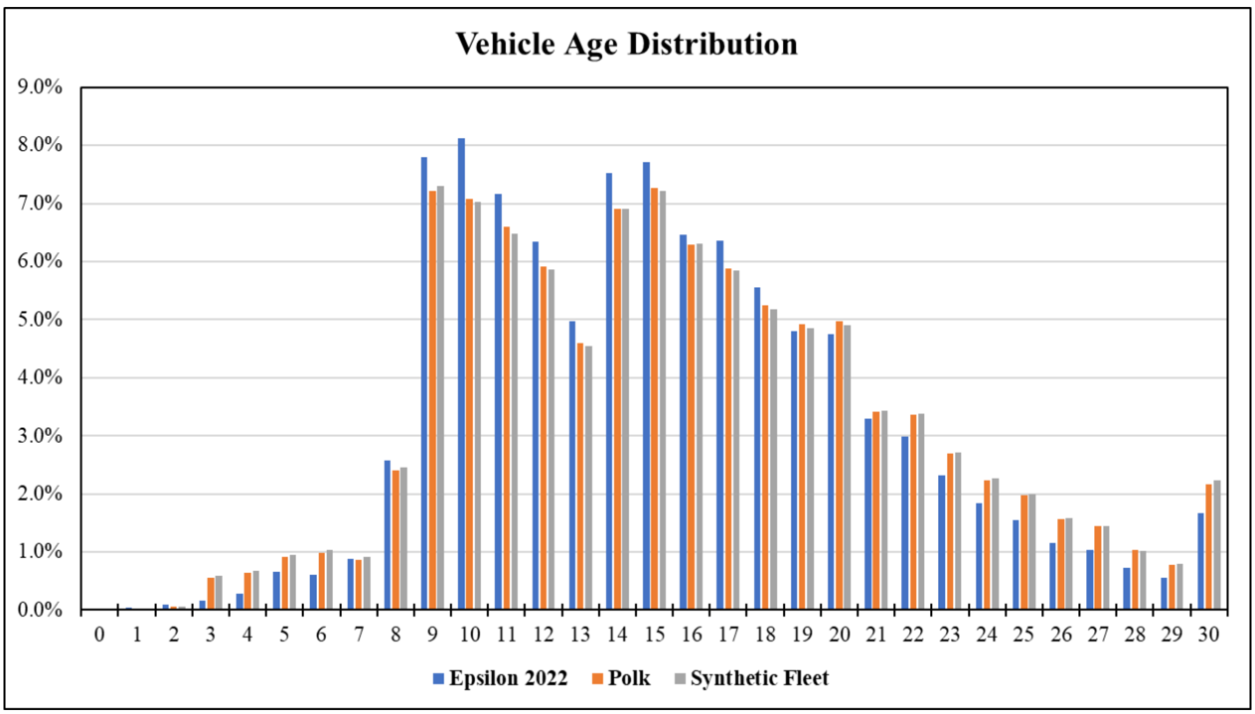

Figure 17. Vehicle age distributions of Epsilon vehicle ownership Data, Polk® Vehicle Registration Data, and the regional synthetic fleet results.

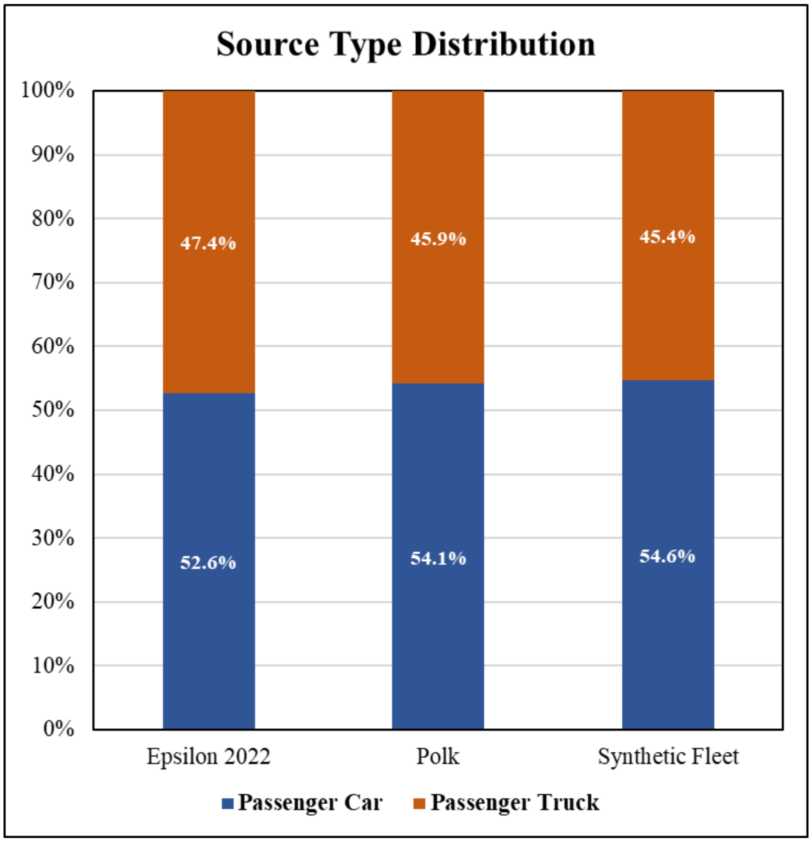

Figure 18. Source type distributions of Epsilon vehicle ownership data, Polk® Vehicle Registration Data, and the regional synthetic fleet results.

Appendix B: Address Suffix Table for Synthetic Household Generation

The suffix used to pair the synthetic households (addresses from Epsilon 2022 demographics vs. Polk® vehicle registration data) is shown in Table 6.

Table 6. Address suffixes for synthetic household generation.

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
APT	UNIT	APT	LAKE	LAKE	LK
APT	#	APT	LAKE	LK	LK
ALLEY	ALLEE	ALY	LAKES	LAKES	LKS
ALLEY	ALLEY	ALY	LAKES	LKS	LKS
ALLEY	ALLY	ALY	LAND	LAND	LAND
ALLEY	ALY	ALY	LANDING	LANDING	LNDG
ANNEX	ANEX	ANX	LANDING	LNDG	LNDG
ANNEX	ANNEX	ANX	LANDING	LNDNG	LNDG
ANNEX	ANNX	ANX	LANE	LA	LN
ANNEX	ANX	ANX	LANE	LANE	LN
ARCADE	ARC	ARC	LANE	LANES	LN
ARCADE	ARCADE	ARC	LANE	LN	LN
AVENUE	AV	AVE	LIGHT	LGT	LGT
AVENUE	AVE	AVE	LIGHT	LIGHT	LGT
AVENUE	AVEN	AVE	LIGHTS	LIGHTS	LGTS
AVENUE	AVENU	AVE	LOAF	LF	LF
AVENUE	AVENUE	AVE	LOAF	LOAF	LF
AVENUE	AVN	AVE	LOCK	LCK	LCK
AVENUE	AVNUE	AVE	LOCK	LOCK	LCK
BAYOO	BAYOO	BYU	LOCKS	LCKS	LCKS
BAYOO	BAYOU	BYU	LOCKS	LOCKS	LCKS
BEACH	BCH	BCH	LODGE	LDG	LDG
BEACH	BEACH	BCH	LODGE	LDGE	LDG
BEND	BEND	BND	LODGE	LODG	LDG
BEND	BND	BND	LODGE	LODGE	LDG
BLUFF	BLF	BLF	LOOP	LOOP	LOOP
BLUFF	BLUF	BLF	LOOP	LOOPS	LOOP
BLUFF	BLUFF	BLF	MALL	MALL	MALL
BLUFFS	BLUFFS	BLFS	MANOR	MANOR	MNR
BOTTOM	BOT	BTM	MANOR	MNR	MNR
BOTTOM	BOTTM	BTM	MANORS	MANORS	MNRS
BOTTOM	BOTTOM	BTM	MANORS	MNRS	MNRS
BOTTOM	BTM	BTM	MEADOW	MDW	MDW
BOULEVARD	BLVD	BLVD	MEADOW	MEADOW	MDW

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
BOULEVARD	BOUL	BLVD	MEADOWS	MDWS	MDWS
BOULEVARD	BOULEVARD	BLVD	MEADOWS	MEADOWS	MDWS
BOULEVARD	BOULV	BLVD	MEADOWS	MEDOWS	MDWS
BRANCH	BR	BR	MEWS	MEWS	MEWS
BRANCH	BRANCH	BR	MILL	MILL	ML
BRANCH	BRNCH	BR	MILL	ML	ML
BRIDGE	BRDGE	BRG	MILLS	MILLS	MLS
BRIDGE	BRG	BRG	MILLS	MLS	MLS
BRIDGE	BRIDGE	BRG	MISSION	MISSION	MSN
BROOK	BRK	BRK	MISSION	MISSN	MSN
BROOK	BROOK	BRK	MISSION	MSN	MSN
BROOKS	BROOKS	BRKS	MISSION	MSSN	MSN
BURG	BURG	BG	MOTORWAY	MOTORWAY	MTWY
BURGS	BURGS	BGS	MOUNT	MNT	MT
BYPASS	BYP	BYP	MOUNT	MOUNT	MT
BYPASS	BYPA	BYP	MOUNT	MT	MT
BYPASS	BYPAS	BYP	MOUNTAIN	MNTAIN	MTN
BYPASS	BYPASS	BYP	MOUNTAIN	MNTN	MTN
BYPASS	BYPS	BYP	MOUNTAIN	MOUNTAIN	MTN
CAMP	CAMP	CP	MOUNTAIN	MOUNTIN	MTN
CAMP	CMP	CP	MOUNTAIN	MTIN	MTN
CAMP	CP	CP	MOUNTAIN	MTN	MTN
CANYON	CANYN	CYN	MOUNTAINS	MNTNS	MTNS
CANYON	CANYON	CYN	MOUNTAINS	MOUNTAINS	MTNS
CANYON	CNYN	CYN	NECK	NCK	NCK
CANYON	CYN	CYN	NECK	NECK	NCK
CAPE	CAPE	CPE	ORCHARD	ORCH	ORCH
CAPE	CPE	CPE	ORCHARD	ORCHARD	ORCH
CAUSEWAY	CAUSEWAY	CSWY	ORCHARD	ORCHRD	ORCH
CAUSEWAY	CAUSWAY	CSWY	OVAL	OVAL	OVAL
CAUSEWAY	CSWY	CSWY	OVAL	OVL	OVAL
CENTER	CEN	CTR	OVERPASS	OVERPASS	OPAS
CENTER	CENT	CTR	PARK	PARK	PARK
CENTER	CENTER	CTR	PARK	PK	PARK
CENTER	CENTR	CTR	PARK	PRK	PARK
CENTER	CENTRE	CTR	PARKS	PARKS	PARK
CENTER	CNTER	CTR	PARKWAY	PARKWAY	PKWY
CENTER	CNTR	CTR	PARKWAY	PARKWY	PKWY
CENTER	CTR	CTR	PARKWAY	PKWAY	PKWY
CENTERS	CENTERS	CTRS	PARKWAY	PKWY	PKWY
CIRCLE	CIR	CIR	PARKWAY	PKY	PKWY

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
CIRCLE	CIRC	CIR	PARKWAYS	PARKWAYS	PKWY
CIRCLE	CIRCL	CIR	PARKWAYS	PKWYS	PKWY
CIRCLE	CIRCLE	CIR	PASS	PASS	PASS
CIRCLE	CRCL	CIR	PASSAGE	PASSAGE	PSGE
CIRCLE	CRCLE	CIR	PATH	PATH	PATH
CIRCLES	CIRCLES	CIRS	PATH	PATHS	PATH
CLIFF	CLF	CLF	PIKE	PIKE	PIKE
CLIFF	CLIFF	CLF	PIKE	PIKES	PIKE
CLIFFS	CLFS	CLFS	PINE	PINE	PNE
CLIFFS	CLIFFS	CLFS	PINES	PINES	PNES
CLUB	CLB	CLB	PINES	PNES	PNES
CLUB	CLUB	CLB	PLACE	PL	PL
COMMON	COMMON	CMN	PLACE	PLACE	PL
CORNER	COR	COR	PLAIN	PLAIN	PLN
CORNER	CORNER	COR	PLAIN	PLN	PLN
CORNERS	CORNERS	CORS	PLAINS	PLAINES	PLNS
CORNERS	CORS	CORS	PLAINS	PLAINS	PLNS
COURSE	COURSE	CRSE	PLAINS	PLNS	PLNS
COURSE	CRSE	CRSE	PLAZA	PLAZA	PLZ
COURT	COURT	CT	PLAZA	PLZ	PLZ
COURT	CRT	CT	PLAZA	PLZA	PLZ
COURT	CT	CT	POINT	POINT	PT
COURTS	COURTS	CTS	POINT	PT	PT
COURTS	CTS	CTS	POINTS	POINTS	PTS
COVE	COVE	CV	POINTS	PTS	PTS
COVE	CV	CV	PORT	PORT	PRT
COVES	COVES	CVS	PORT	PRT	PRT
CREEK	CK	CRK	PORTS	PORTS	PRTS
CREEK	CR	CRK	PORTS	PRTS	PRTS
CREEK	CREEK	CRK	PRAIRIE	PR	PR
CREEK	CRK	CRK	PRAIRIE	PRAIRIE	PR
CRESCENT	CRECENT	CRES	PRAIRIE	PRARIE	PR
CRESCENT	CRES	CRES	PRAIRIE	PRR	PR
CRESCENT	CRESCENT	CRES	RADIAL	RAD	RADL
CRESCENT	CRESENT	CRES	RADIAL	RADIAL	RADL
CRESCENT	CRSCNT	CRES	RADIAL	RADIEL	RADL
CRESCENT	CRSENT	CRES	RADIAL	RADL	RADL
CRESCENT	CRSNT	CRES	RAMP	RAMP	RAMP
CREST	CREST	CRST	RANCH	RANCH	RNCH
CROSSING	CROSSING	XING	RANCH	RANCHES	RNCH
CROSSING	CRSSING	XING	RANCH	RNCH	RNCH

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
CROSSING	CRSSNG	XING	RANCH	RNCHS	RNCH
CROSSING	XING	XING	RAPID	RAPID	RPD
CROSSROAD	CROSSROAD	XRD	RAPID	RPD	RPD
CURVE	CURVE	CURV	RAPIDS	RAPIDS	RPDS
DALE	DALE	DL	RAPIDS	RPDS	RPDS
DALE	DL	DL	REST	REST	RST
DAM	DAM	DM	REST	RST	RST
DAM	DM	DM	RIDGE	RDG	RDG
DIVIDE	DIV	DV	RIDGE	RDGE	RDG
DIVIDE	DIVIDE	DV	RIDGE	RIDGE	RDG
DIVIDE	DV	DV	RIDGES	RDGS	RDGS
DIVIDE	DVD	DV	RIDGES	RIDGES	RDGS
DRIVE	DR	DR	RIVER	RIV	RIV
DRIVE	DRIV	DR	RIVER	RIVER	RIV
DRIVE	DRIVE	DR	RIVER	RIVR	RIV
DRIVE	DRV	DR	RIVER	RVR	RIV
DRIVES	DRIVES	DRS	ROAD	RD	RD
ESTATE	EST	EST	ROAD	ROAD	RD
ESTATE	ESTATE	EST	ROADS	RDS	RDS
ESTATES	ESTATES	ESTS	ROADS	ROADS	RDS
ESTATES	ESTS	ESTS	ROUTE	ROUTE	RTE
EXPRESSWAY	EXP	EXPY	ROW	ROW	ROW
EXPRESSWAY	EXPR	EXPY	RUE	RUE	RUE
EXPRESSWAY	EXPRESS	EXPY	RUN	RUN	RUN
EXPRESSWAY	EXPRESSWAY	EXPY	SHOAL	SHL	SHL
EXPRESSWAY	EXPW	EXPY	SHOAL	SHOAL	SHL
EXPRESSWAY	EXPY	EXPY	SHOALS	SHLS	SHLS
EXTENSION	EXT	EXT	SHOALS	SHOALS	SHLS
EXTENSION	EXTENSION	EXT	SHORE	SHOAR	SHR
EXTENSION	EXTN	EXT	SHORE	SHORE	SHR
EXTENSION	EXTNSN	EXT	SHORE	SHR	SHR
EXTENSIONS	EXTENSIONS	EXTS	SHORES	SHOARS	SHRS
EXTENSIONS	EXTS	EXTS	SHORES	SHORES	SHRS
FALL	FALL	FALL	SHORES	SHRS	SHRS
FALLS	FALLS	FLS	SKYWAY	SKYWAY	SKWY
FALLS	FLS	FLS	SPRING	SPG	SPG
FERRY	FERRY	FRY	SPRING	SPNG	SPG
FERRY	FRRY	FRY	SPRING	SPRING	SPG
FERRY	FRY	FRY	SPRING	SPRNG	SPG
FIELD	FIELD	FLD	SPRINGS	SPGS	SPGS
FIELD	FLD	FLD	SPRINGS	SPNGS	SPGS

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
FIELDS	FIELDS	FLDS	SPRINGS	SPRINGS	SPGS
FIELDS	FLDS	FLDS	SPRINGS	SPRNGS	SPGS
FLAT	FLAT	FLT	SPUR	SPUR	SPUR
FLAT	FLT	FLT	SPURS	SPURS	SPUR
FLATS	FLATS	FLTS	SQUARE	SQ	SQ
FLATS	FLTS	FLTS	SQUARE	SQR	SQ
FORD	FORD	FRD	SQUARE	SQRE	SQ
FORD	FRD	FRD	SQUARE	SQU	SQ
FORDS	FORDS	FRDS	SQUARE	SQUARE	SQ
FOREST	FOREST	FRST	SQUARES	SQRS	SQS
FOREST	FORESTS	FRST	SQUARES	SQUARES	SQS
FOREST	FRST	FRST	STATION	STA	STA
FORGE	FORG	FRG	STATION	STATION	STA
FORGE	FORGE	FRG	STATION	STATN	STA
FORGE	FRG	FRG	STATION	STN	STA
FORGES	FORGES	FRGS	STRAVENUE	STRA	STRA
FORK	FORK	FRK	STRAVENUE	STRAV	STRA
FORK	FRK	FRK	STRAVENUE	STRAVE	STRA
FORKS	FORKS	FRKS	STRAVENUE	STRAVEN	STRA
FORKS	FRKS	FRKS	STRAVENUE	STRAVENUE	STRA
FORT	FORT	FT	STRAVENUE	STRAVN	STRA
FORT	FRT	FT	STRAVENUE	STRVN	STRA
FORT	FT	FT	STRAVENUE	STRVNUE	STRA
FREEWAY	FREEWAY	FWY	STREAM	STREAM	STRM
FREEWAY	FREEWY	FWY	STREAM	STREME	STRM
FREEWAY	FRWAY	FWY	STREAM	STRM	STRM
FREEWAY	FRWY	FWY	STREET	ST	ST
FREEWAY	FWY	FWY	STREET	STR	ST
GARDEN	GARDEN	GDN	STREET	STREET	ST
GARDEN	GARDN	GDN	STREET	STRT	ST
GARDEN	GDN	GDN	STREETS	STREETS	STS
GARDEN	GRDEN	GDN	SUMMIT	SMT	SMT
GARDEN	GRDN	GDN	SUMMIT	SUMIT	SMT
GARDENS	GARDENS	GDNS	SUMMIT	SUMITT	SMT
GARDENS	GDNS	GDNS	SUMMIT	SUMMIT	SMT
GARDENS	GRDNS	GDNS	TERRACE	TER	TER
GATEWAY	GATEWAY	GTWY	TERRACE	TERR	TER
GATEWAY	GATEWY	GTWY	TERRACE	TERRACE	TER
GATEWAY	GATWAY	GTWY	THROUGHWAY	THROUGHWAY	TRWY
GATEWAY	GTWAY	GTWY	TRACE	TRACE	TRCE
GATEWAY	GTWY	GTWY	TRACE	TRACES	TRCE

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
GLEN	GLEN	GLN	TRACE	TRCE	TRCE
GLEN	GLN	GLN	TRACK	TRACK	TRAK
GLENS	GLENS	GLNS	TRACK	TRACKS	TRAK
GREEN	GREEN	GRN	TRACK	TRAK	TRAK
GREEN	GRN	GRN	TRACK	TRK	TRAK
GREENS	GREENS	GRNS	TRACK	TRKS	TRAK
GROVE	GROV	GRV	TRAFFICWAY	TRAFFICWAY	TRFY
GROVE	GROVE	GRV	TRAFFICWAY	TRFY	TRFY
GROVE	GRV	GRV	TRAIL	TR	TRL
GROVES	GROVES	GRVS	TRAIL	TRAIL	TRL
HARBOR	HARB	HBR	TRAIL	TRAILS	TRL
HARBOR	HARBOR	HBR	TRAIL	TRL	TRL
HARBOR	HARBR	HBR	TRAIL	TRLS	TRL
HARBOR	HBR	HBR	TUNNEL	TUNEL	TUNL
HARBOR	HRBOR	HBR	TUNNEL	TUNL	TUNL
HARBORS	HARBORS	HBRS	TUNNEL	TUNLS	TUNL
HAVEN	HAVEN	HVN	TUNNEL	TUNNEL	TUNL
HAVEN	HAVN	HVN	TUNNEL	TUNNELS	TUNL
HAVEN	HVN	HVN	TUNNEL	TUNNL	TUNL
HEIGHTS	HEIGHT	HTS	TURNPIKE	TPK	TPKE
HEIGHTS	HEIGHTS	HTS	TURNPIKE	TPKE	TPKE
HEIGHTS	HGTS	HTS	TURNPIKE	TRNPK	TPKE
HEIGHTS	HT	HTS	TURNPIKE	TRPK	TPKE
HEIGHTS	HTS	HTS	TURNPIKE	TURNPIKE	TPKE
HIGHWAY	HIGHWAY	HWY	TURNPIKE	TURNPK	TPKE
HIGHWAY	HIGHWY	HWY	UNDERPASS	UNDERPASS	UPAS
HIGHWAY	HIWAY	HWY	UNION	UN	UN
HIGHWAY	HIWY	HWY	UNION	UNION	UN
HIGHWAY	HWAY	HWY	UNIONS	UNIONS	UNS
HIGHWAY	HWY	HWY	VALLEY	VALLEY	VLY
HILL	HILL	HL	VALLEY	VALLY	VLY
HILL	HL	HL	VALLEY	VLLY	VLY
HILLS	HILLS	HLS	VALLEY	VLY	VLY
HILLS	HLS	HLS	VALLEYS	VALLEYS	VLYS
HOLLOW	HLLW	HOLW	VALLEYS	VLYS	VLYS
HOLLOW	HOLLOW	HOLW	VIADUCT	VDCT	IA
HOLLOW	HOLLOWS	HOLW	VIADUCT	VIA	VIA
HOLLOW	HOLW	HOLW	VIADUCT	VIADCT	VIA
HOLLOW	HOLWS	HOLW	VIADUCT	VIADUCT	VIA
INLET	INLET	INLT	VIEW	VIEW	VW
INLET	INLT	INLT	VIEW	VW	VW

Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation	Primary Street Suffix	Commonly Used Abbreviation	Postal Service Standard Suffix Abbreviation
ISLAND	IS	IS	VIEWS	VIEWS	VWS
ISLAND	ISLAND	IS	VIEWS	VWS	VWS
ISLAND	ISLND	IS	VILLAGE	VILL	VLG
ISLANDS	ISLANDS	SS	VILLAGE	VILLAG	VLG
ISLANDS	I	SLNDS	VILLAGE	VILLAGE	VLG
ISLANDS	ISS	ISS	VILLAGE	VILLG	VLG
ISLE	ISLE	ISLE	VILLAGE	VILLIAGE	VLG
ISLE	ISLES	ISLE	VILLAGE	VLG	VLG
JUNCTION	JCT	JCT	VILLAGES	VILLAGES	VLGS
JUNCTION	JCTION	JCT	VILLAGES	VLGS	VLGS
JUNCTION	JCTN	JCT	VILLE	VILLE	VL
JUNCTION	JUNCTION	JCT	VILLE	VL	VL
JUNCTION	JUNCTN	JCT	VISTA	VIS	VIS
JUNCTION	JUNCTON	JCT	VISTA	VIST	VIS
JUNCTIONS	JCTNS	JCTS	VISTA	VISTA	VIS
JUNCTIONS	JCTS	JCTS	VISTA	VST	VIS
JUNCTIONS	JUNCTIONS	JCTS	VISTA	VSTA	VIS
KEY	KEY	KY	WALK	WALK	WALK
KEY	KY	KY	WALKS	WALKS	WALK
KEYS	KEYS	KYS	WALL	WALL	WALL
KEYS	KYS	KYS	WAY	WAY	WAY
KNOLL	KNL	KNL	WAY	WY	WAY
KNOLL	KNOL	KNL	WAYS	WAYS	WAYS
KNOLL	KNOLL	KNL	WELL	WELL	WL
KNOLLS	KNLS	KNLS	WELLS	WELLS	WLS
KNOLLS	KNOLLS	KNLS	WELLS	WLS	WLS

Appendix C: Generation of Feature Importance Based on Random Forest and Decision Tree Classifier

The weights of demographic variables that were used to model the similarities between any pair of households (in terms of their vehicle ownership) were determined by evaluating the feature importance results of each variable. The variables of number of children, number of adults, household income and ethnicity were first used as input variables to establish the models that predict number of vehicles (target variable), and for each vehicle, prediction models were established iteratively to predict the model year and source type (the variable of number of vehicles was also used as one of the input variables in these iterations). The permutation importance results based on decision tree regression models and the feature importance based on Random Forest models of each variable were then normalized and examined, and the weights were determined by evaluating these results.

The assessment of feature importance started with four initial features (input variables), three of which were numerical (the variables of number of children, number of adults, and household income), and one was categorical (the variable of ethnicity). The initial target variable was number of vehicles, which is also numerical. The team employed one-hot encoding to create a set of binary features representing the various categories of ethnicity, which allowed the categorical variable to be included in the modeling process seamlessly.

The team first assessed the permutation feature importance, which is a model-agnostic method used to estimate the importance of input features by evaluating the decrease in model performance when the values of a specific feature are randomly shuffled (Brière, et al., 2021). This approach works by breaking the relationship between the feature and the target variable, and subsequently measuring the impact on the model's prediction accuracy. The rationale is that if permuting a feature leads to a significant drop in model performance, that feature is considered important for the model's predictions.

The team used a DecisionTreeRegressor model to analyze the feature importance for predicting the initial target variable, and the model's performance was evaluated using the Root Mean Squared Error (RMSE) metric, which measures the average difference (Euclidean distance) between the predicted and actual values of the target variable.

The additional pairs of target variables related to every vehicle were then accounted for as 'Vehicle#*i*SourceType' and 'Vehicle#*i*ModelYear' (where i iterates from 0 to 5), and the vehicles were sorted for each household by source type (passenger cars were ranked before passenger trucks) and then by model year (new vehicles were ranked before old vehicles) so that the iterations represent the vehicle ownership characteristics (lighter and newer vehicles were predicted before heavier and older vehicles). The team iteratively incorporated these pairs into the analysis, using the previous features (number of children, number of adults, household income and number of vehicles) and target variable pairs (the source type and model year of the last vehicle) as inputs.

For each iteration, the feature importance results of the input variables were derived based on the current target variable pair (source type and model year) with a *MultiOutputRegressor* wrapper around the *DecisionTreeRegressor* model (which fits one regressor per target and combines their results to efficiently model the multiple-output problem), and the *permutation_importance* function (from the *scikit-learn* library) was used to provide the permutation importance by input variable. The parameter of *n_repeats* was set to 100, which controls the number of times the model performance is evaluated with the permuted feature values. A higher number of repeats provides a more accurate estimate of feature importance, as it averages the results over multiple permutations. Although increasing *n_repeats* also increases the computational time required for the calculation, it is a choice of balance between the need for accuracy vs. computational efficiency. The value of 10 for n_repeats is widely considered a reasonable choice (Fisher, et al., 2019), as it provides a good trade-off between accuracy and computation time, and the team used a larger value of 100 to provide more stable results. The testing of *n_repeats* parameter from 10 to 100 at intervals of 10 demonstrate almost the same results, and it was indicated that the model's feature importance results are stable and robust (the chosen value of *n_repeats* is large enough to provide reliable estimates of feature importance). The results of the permutation importance assessment are presented in Table 7.

The team also used the Random Forest Regressor to compute the feature importance results, which is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees. It offers several advantages, such as improved accuracy and robustness compared to a single decision tree, and the ability to handle missing data and overfitting (Breiman, 2001). The feature importance in a Random Forest model are calculated by averaging the impurity reduction of each feature across all trees in the forest. Impurity reduction is a measure of how much a feature contributes to improving the prediction when used in a split (Breiman, et al., 1984). This method provides a more stable and reliable estimate of feature importance results compared to a single decision tree, as it accounts for interactions between features and variations in the dataset.

The Random Forest Regressors were employed in the same way with the decision tree regressors (prediction of number of vehicles first, and then iterative prediction of each vehicle), and *MultiOutputRegressor* was also used to handle a pair of target variables at a time. The results of the feature importance assessment by Random Forest are shown in Table 8.

The feature importance results from both regressors were normalized to the scale from 0 to 100, as shown in Table 9, and the normalized importance results were manually examined to assign the weight for pairing the households (Euclidean distances). It was indicated that household income has larger importance when it comes to more vehicles owned, and given that most households do not possess more than 3 vehicles, the weights by input variable were determined based on iterations from #1 to #5, as shown in Table 10.

Table 7. Feature importance of Decision Tree Regressor.

Iteration	Number of Children	Number of Adults	Household Income	Ethnic Group	Number of Vehicles	RMSE
# of Paired Vehicle	0.001	0.046	0.062	0.012	N/A	0.7
1	0.003	0.005	0.028	0.019	0.075	5.5
2	0.006	0.012	0.024	0.021	0.043	5.0
3	-0.001	0.012	0.010	0.000	0.027	5.5
4	0.009	-0.010	-0.014	0.012	0.025	5.5
5	-0.008	-0.002	0.033	0.020	0.029	4.6
6	-0.005	0.001	0.203	0.091	-0.005	2.8

Table 8. Permutation importance of Random Forest Regressor.

Iteration	Number of Children	Number of Adults	Household Income	Ethnic Group	Number of Vehicles	RMSE
# of Paired Vehicle	0.063	0.511	0.355	0.071	N/A	0.7
1	0.038	0.024	0.105	0.038	0.295	5.5
2	0.033	0.033	0.067	0.036	0.044	4.7
3	0.027	0.035	0.081	0.036	0.019	4.5
4	0.017	0.024	0.065	0.028	0.015	4.0
5	0.012	0.016	0.038	0.021	0.013	3.4
6	0.007	0.011	0.039	0.012	0.001	2.3

Table 9. Normalized importance results by iteration.

Regressor	Iteration	Number of Children	Number of Adults	Household Income	Number of Vehicles	Ethnic Group
Random Forest	Paired Vehicle	12.4	100.0	69.6	N/A	13.8
Random Forest	1	12.8	8.2	35.5	100.0	13.0
Random Forest	2	49.5	49.2	100.0	65.0	54.0
Random Forest	3	32.8	42.9	100.0	23.7	44.1
Random Forest	4	26.5	37.3	100.0	22.5	42.7
Random Forest	5	31.3	42.4	100.0	35.3	55.8
Random Forest	6	16.8	27.5	100.0	1.3	29.3
Decision Tree	Paired Vehicle	2.3	74.3	100.0	N/A	19.1
Decision Tree	1	3.9	6.5	36.5	100.0	25.2
Decision Tree	2	13.9	29.2	57.2	100.0	48.5
Decision Tree	3	5.2	45.7	37.5	100.0	0.4
Decision Tree	4	34.3	39.0	55.0	100.0	47.4
Decision Tree	5	24.9	6.6	100.0	87.4	58.6
Decision Tree	6	2.2	0.5	100.0	2.6	44.8

Table 10. Average importance results and weights assigned.

Variable	Average Importance	Weight	Note
Number of Children	20.8	0.1	Average for the iterations targeting the number of paired vehicles and the iterations from #1 to #5.
Number of Adults	40.1	0.2	Average for the iterations targeting the number of paired vehicles and the iterations from #1 to #5.
Household Income	74.3	0.35	Average for the iterations targeting the number of paired vehicles and the iterations from #1 to #5.
Number of Vehicles	73.4	0.35	Average for the iterations from #1 to #5.

Appendix D: Relative Mileage Accumulative Rates

This research adopted the RMAR from the MOVES2014 model (U.S. EPA, 2016), which was derived from the VMT distributions from the NHTSA 2001 survey (NHTSA). The regressed model of VMT prediction by vehicle age for passenger cars and for light duty trucks were extracted, and the mileage for ages of 26 to 30 for passenger cars were extrapolated.

The linear relationship between VMT and vehicle age for passenger cars are shown in Equation (5) and Equation (6).

$$VMT_{Passenger\ cars} = \mathrm{A}_{PC} \times (\mathrm{Age})^3 + \mathrm{B}_{PC} \times (\mathrm{Age})^2 + \mathrm{C}_{PC} \times \mathrm{Age} + D_{PC} \quad (5)$$

$$VMT_{Light\ trucks} = \begin{cases} \mathrm{A}_{LT} \times (\mathrm{Age})^3 + \mathrm{B}_{LT} \times (\mathrm{Age})^2 + \mathrm{C}_{LT} \times \mathrm{Age} + D_{LT} & Age \leq 27 \\ 6{,}648 & 27 \leq Age \leq 36 \end{cases} \quad (6)$$

where the coefficients are as follows in Table 11.

Table 11. Coefficients for VMT estimates from NHTSA report (NHTSA, 2006).

Scenario	Number of MOVES-Matrix Links
A_{PC}	0.3672131
B_{PC}	-13.21949
C_{PC}	-232.8491
D_{PC}	14476.36
A_{LT}	0.6806403
B_{LT}	-22.84481
C_{LT}	-238.5518
D_{LT}	16345.32

Note the NHTSA vehicle age starts from 1 instead of 0, and the estimated VMT were adjusted accordingly to be converted to MOVES vehicle age. The RMAR for passenger cars were further adjusted to 88.5 percent to accommodate the lower mileage for new passenger cars. The RMAR used in this research is shown in Table 3 in the context. Please note the team did not find the RMAR values in the MOVES2014 and MOVES3 database, and the RMAR values were derived based on the methodology provided in EPA's vehicle population and activity report (U.S. EPA, 2016).

Appendix E: Synthetic Fleet Ownership (Vehicle Counts)

This appendix provides the number of vehicles owned by all synthetic households across the 1,000 primary iterations, as shown in Table 12.

Table 12. Synthetic fleet ownership results.

Source Type	Vehicle Age ID	Average Count	2.5 Percentile of Counts	97.5 Percentile of Counts
21	9	161,480	161,029	161,890
21	10	155,253	154,824	155,652
21	11	135,469	135,080	135,831
21	12	143,599	143,224	143,989
21	13	120,033	119,718	120,415
21	14	148,970	148,549	149,366
21	15	153,830	153,402	154,237
21	16	140,473	140,087	140,862
21	17	123,469	123,094	123,822
21	18	101,340	101,025	101,686
21	19	99,830	99,499	100,150
21	20	103,006	102,658	103,336
21	21	86,493	86,189	86,790
21	22	87,113	86,802	87,413
21	23	70,182	69,905	70,444
21	24	60,335	60,094	60,570
21	25	57,647	57,410	57,900
21	26	46,509	46,293	46,732
21	27	50,476	50,251	50,710
21	28	32,161	31,966	32,359
21	29	28,695	28,519	28,869
21	30	84,858	84,568	85,169
31	9	106,384	106,072	106,709
31	10	104,047	103,721	104,381
31	11	103,118	102,782	103,432
31	12	82,171	81,880	82,462
31	13	58,830	58,605	59,071
31	14	117,314	116,991	117,653
31	15	131,837	131,482	132,193
31	16	117,863	117,506	118,196
31	17	118,230	117,883	118,589
31	18	113,260	112,887	113,598
31	19	108,795	108,469	109,121
31	20	111,283	110,944	111,621
31	21	68,151	67,875	68,419
31	22	64,190	63,940	64,454
31	23	55,113	54,890	55,336
31	24	42,542	42,351	42,747
31	25	35,936	35,745	36,125

Source Type	Vehicle Age ID	Average Count	2.5 Percentile of Counts	97.5 Percentile of Counts
31	26	28,486	28,318	28,663
31	27	21,545	21,394	21,703
31	28	16,091	15,971	16,213
31	29	11,119	11,007	11,230
31	30	29,076	28,893	29,244
Total	All	3,836,602	3,836,204	3,836,965

Appendix F: Link-by-link Emission Differences by Hour

This appendix presents the link-by-link results of the emission differences (between the synthetic fleet vs. the county-based approach) by hour. The average differences are presented in Figure 19 through Figure 42, the CIs are presented in Figure 43 through Figure 66, and the biased links (with both CI bounds negative or positive) are shown in Figure 67 through Figure 90.

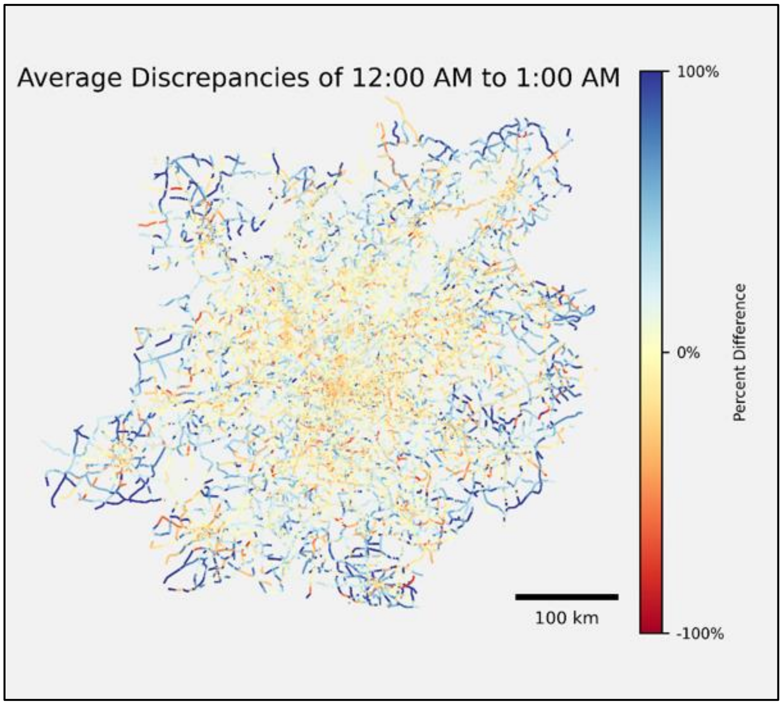

Figure 19. Link-by-link average differences between the scenarios, 12 AM to 1AM.

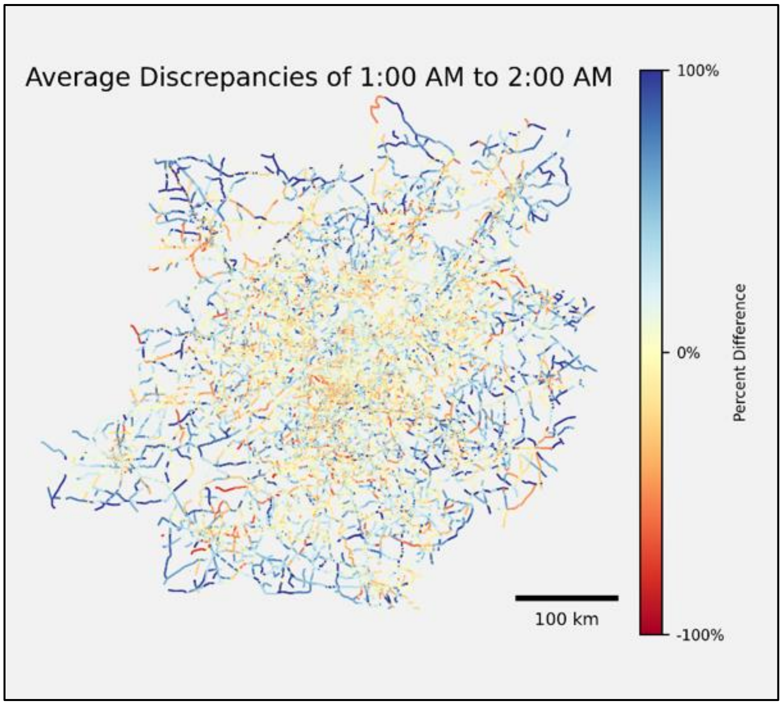

Figure 20. Link-by-link average differences between the scenarios, 1 AM to 2 AM.

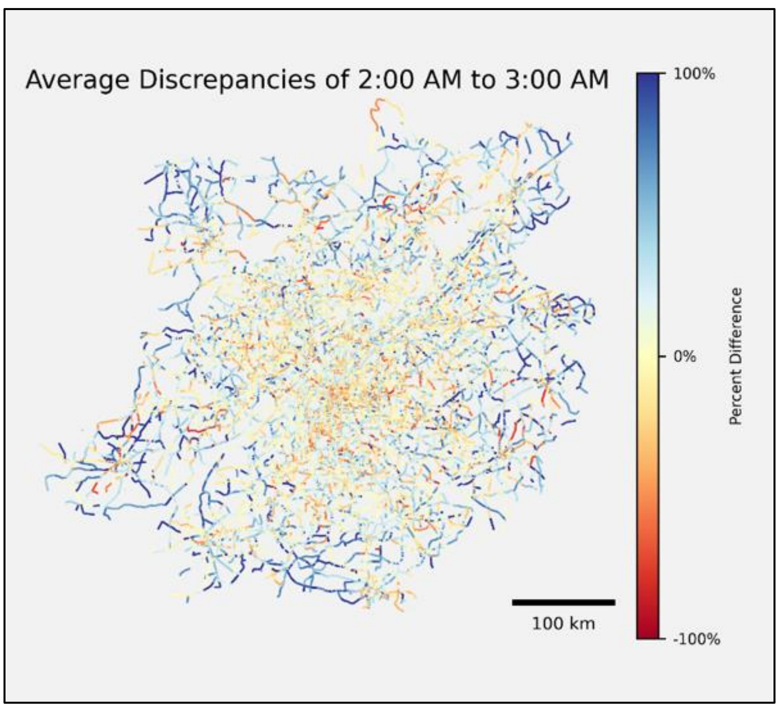

Figure 21. Link-by-link average differences between the scenarios, 2 AM to 3 AM.

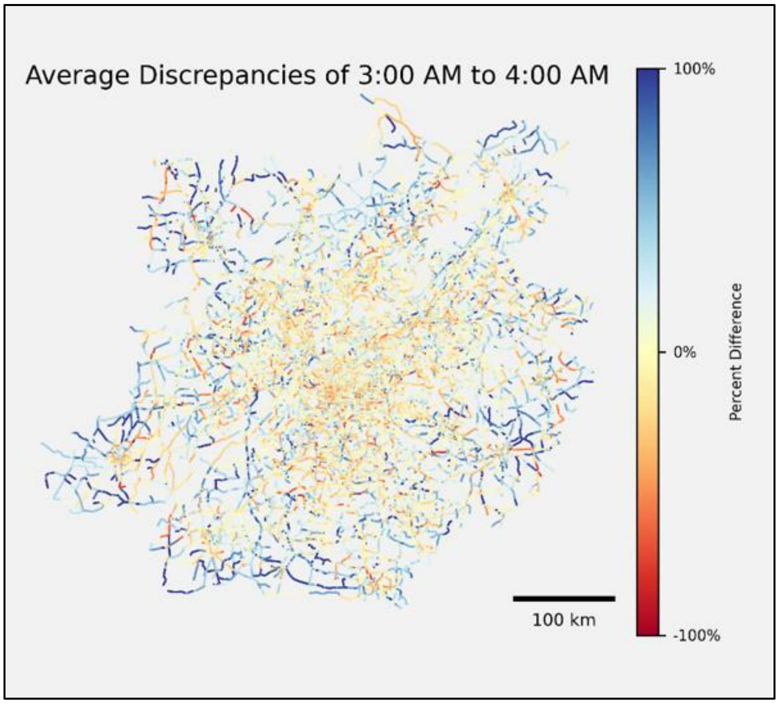

Figure 22. Link-by-link average differences between the scenarios, 3 AM to 4 AM.

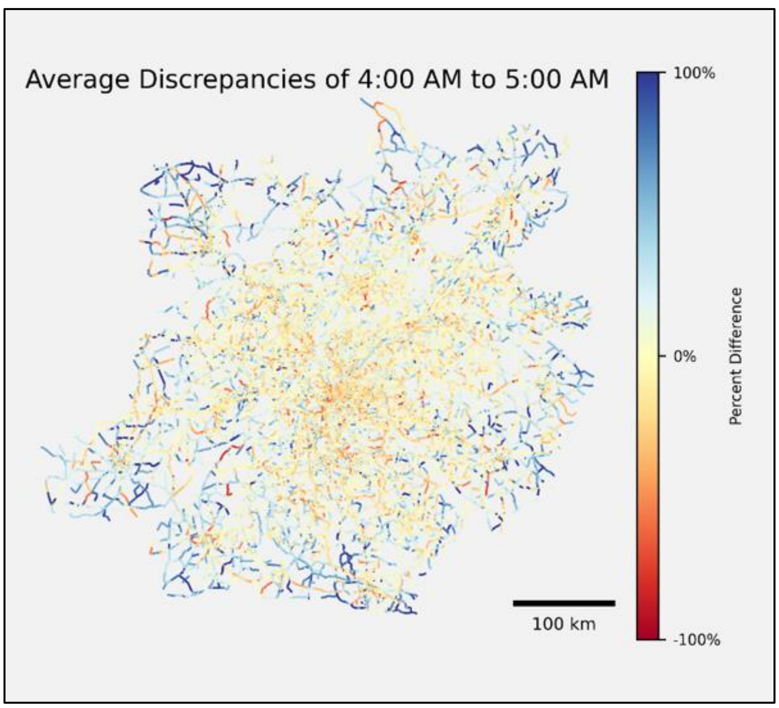

Figure 23. Link-by-link average differences between the scenarios, 4 AM to 5 AM.

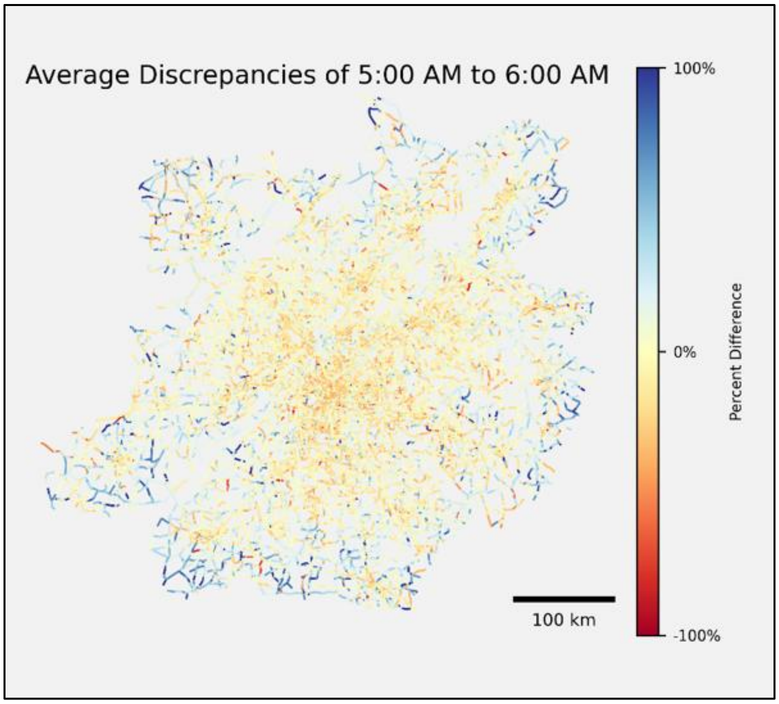

Figure 24. Link-by-link average differences between the scenarios, 5 AM to 6 AM.

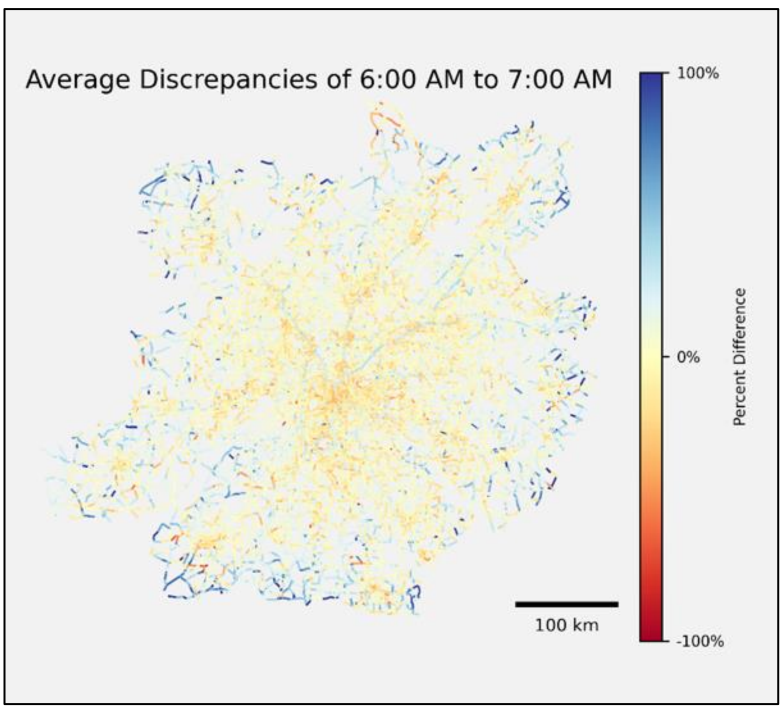

Figure 25. Link-by-link average differences between the scenarios, 6 AM to 7 AM.

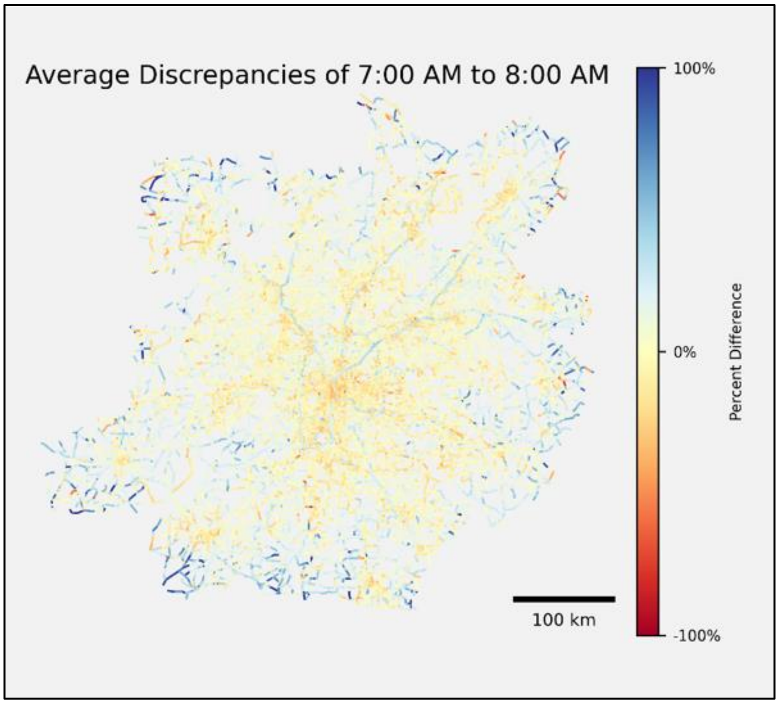

Figure 26. Link-by-link average differences between the scenarios, 7 AM to 8 AM.

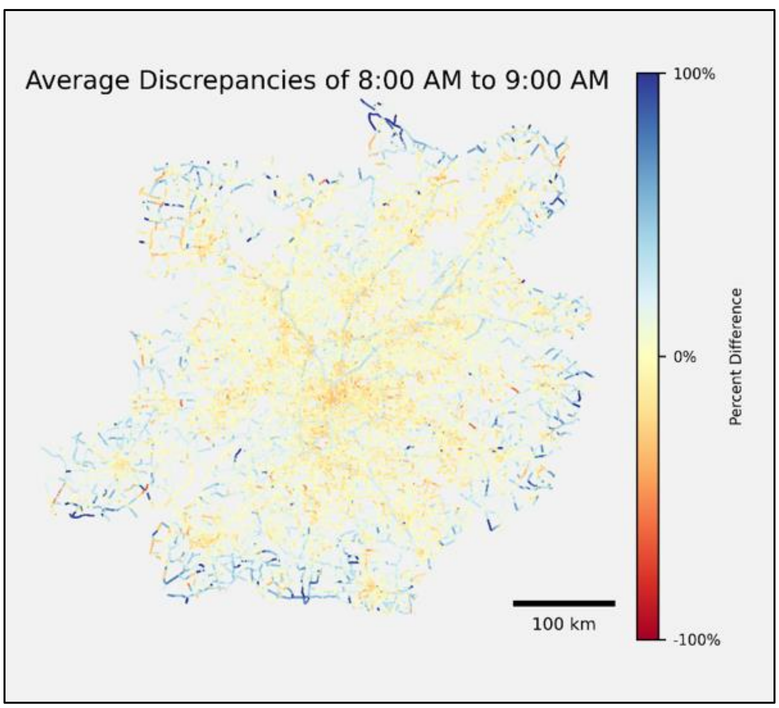

Figure 27. Link-by-link average differences between the scenarios, 8 AM to 9 AM.

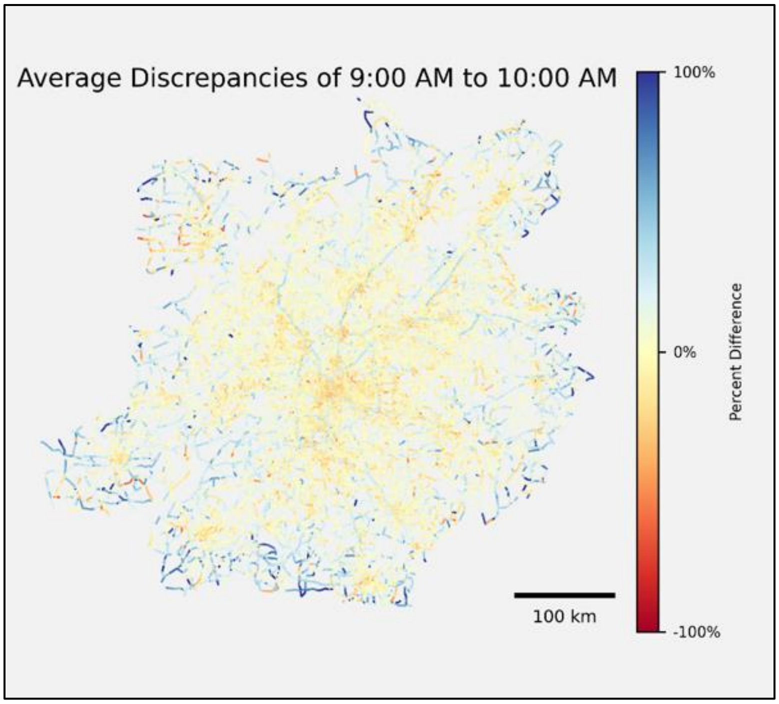

Figure 28. Link-by-link average differences between the scenarios, 9 AM to 10 AM.

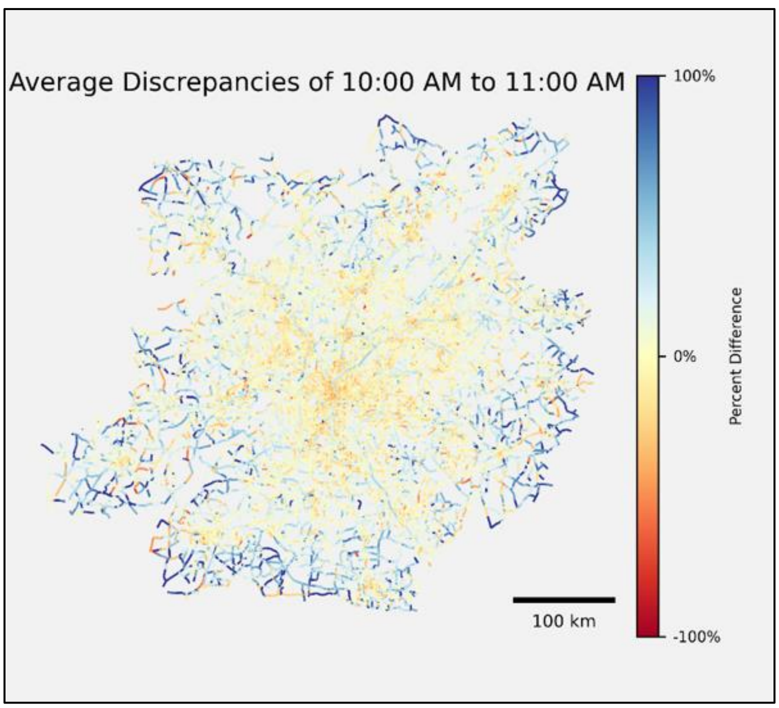

Figure 29. Link-by-link average differences between the scenarios, 10 AM to 11 AM.

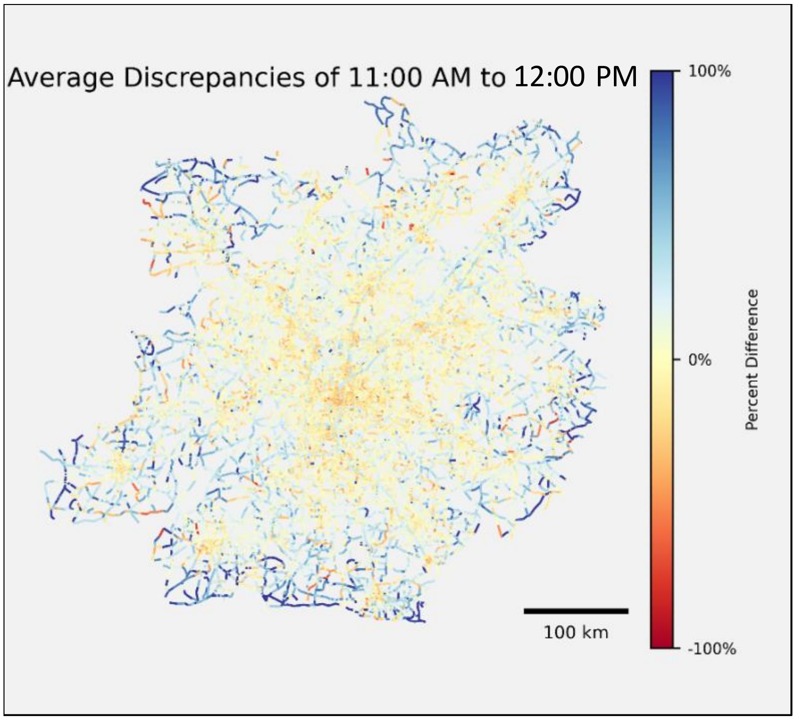

Figure 30. Link-by-link average differences between the scenarios, 11 AM to 12 PM.

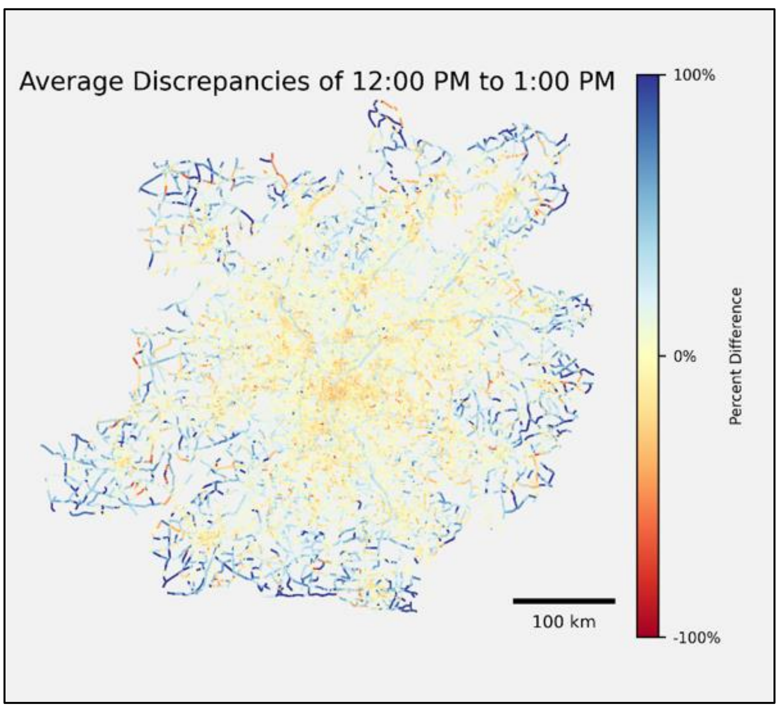

Figure 31. Link-by-link average differences between the scenarios, 12 PM to 1 PM.

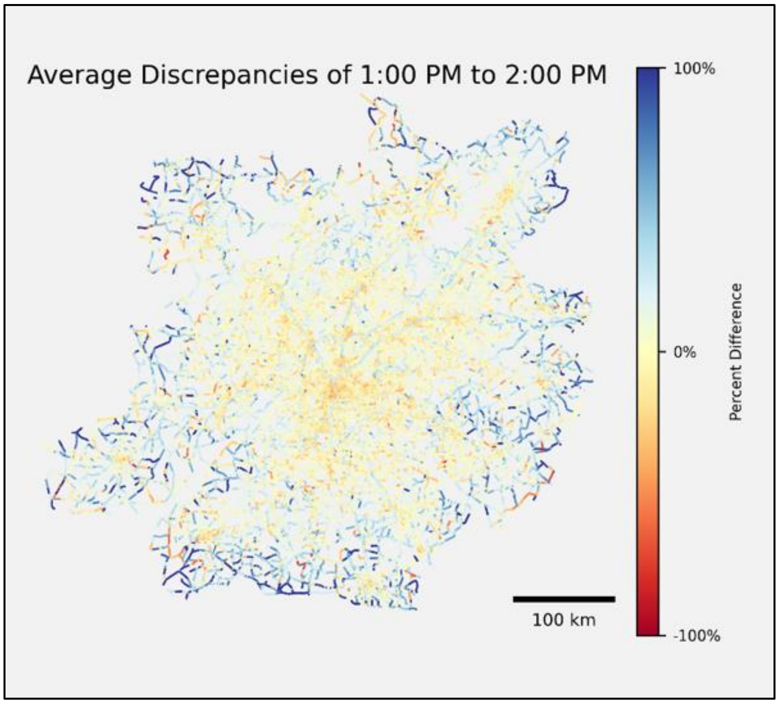

Figure 32. Link-by-link average differences between the scenarios, 1 PM to 2 PM.

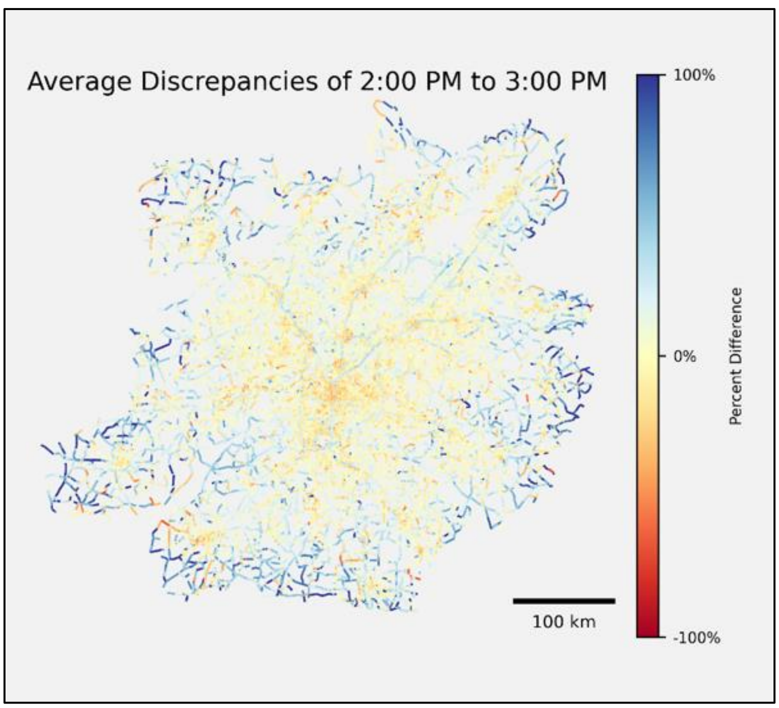

Figure 33. Link-by-link average differences between the scenarios, 2 PM to 3 PM.

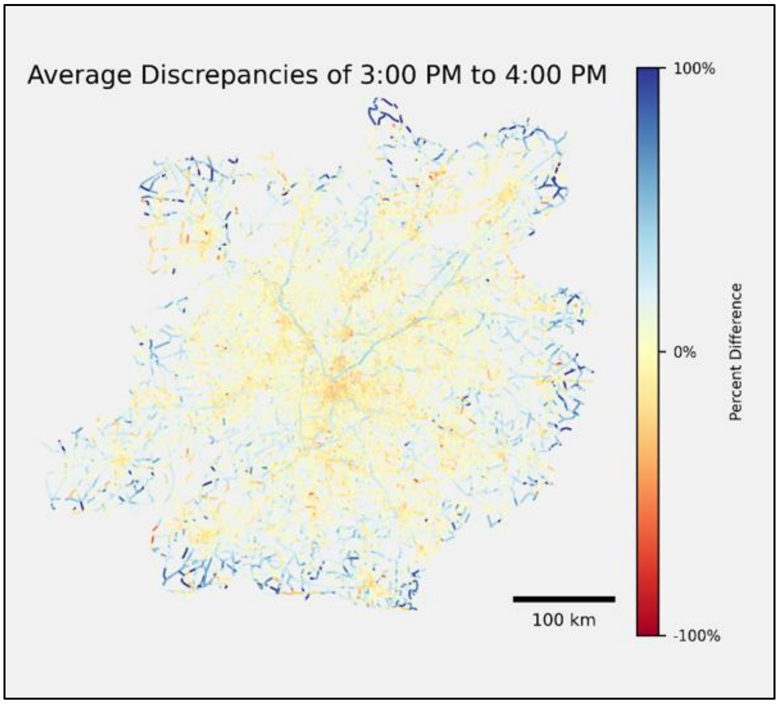

Figure 34. Link-by-link average differences between the scenarios, 3 PM to 4 PM.

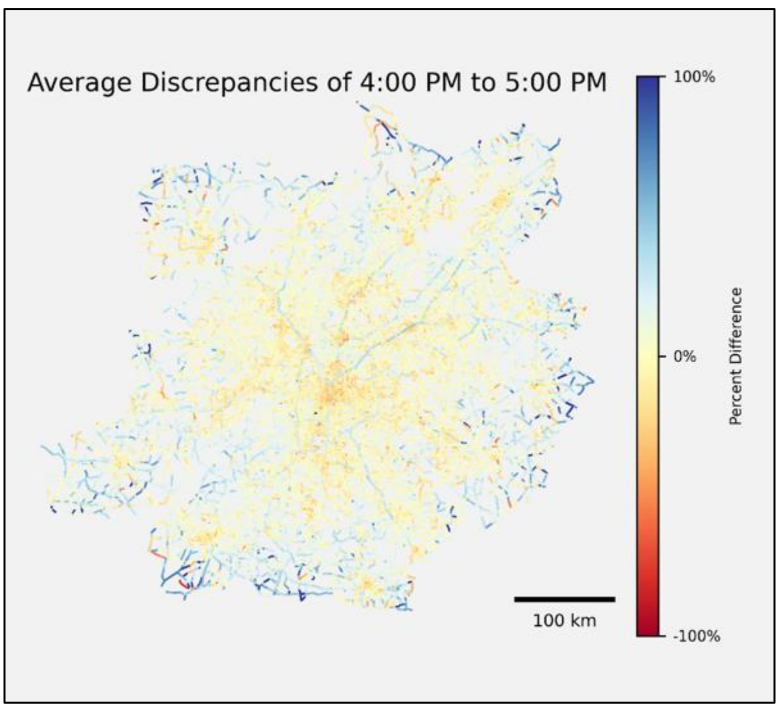

Figure 35. Link-by-link average differences between the scenarios, 4 PM to 5 PM.

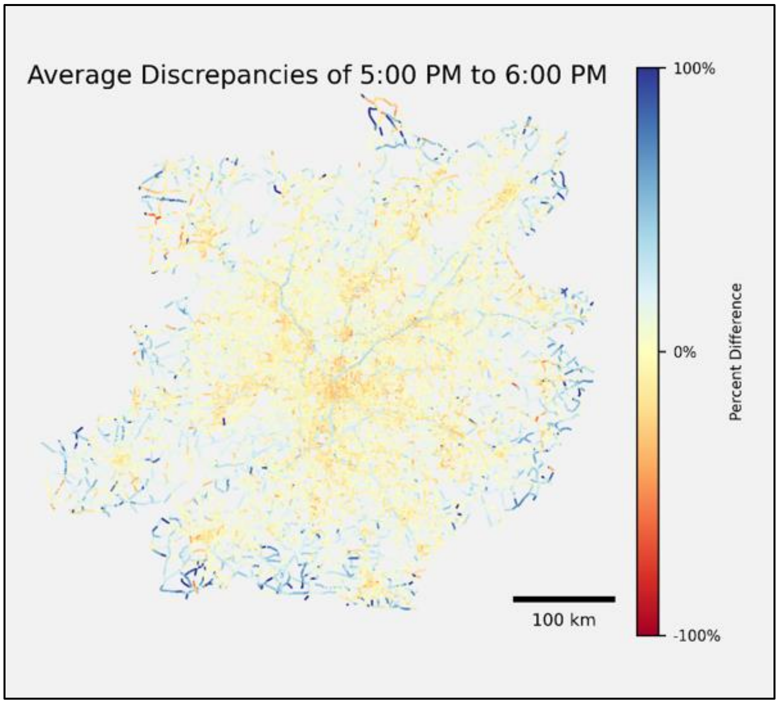

Figure 36. Link-by-link average differences between the scenarios, 5 PM to 6 PM.

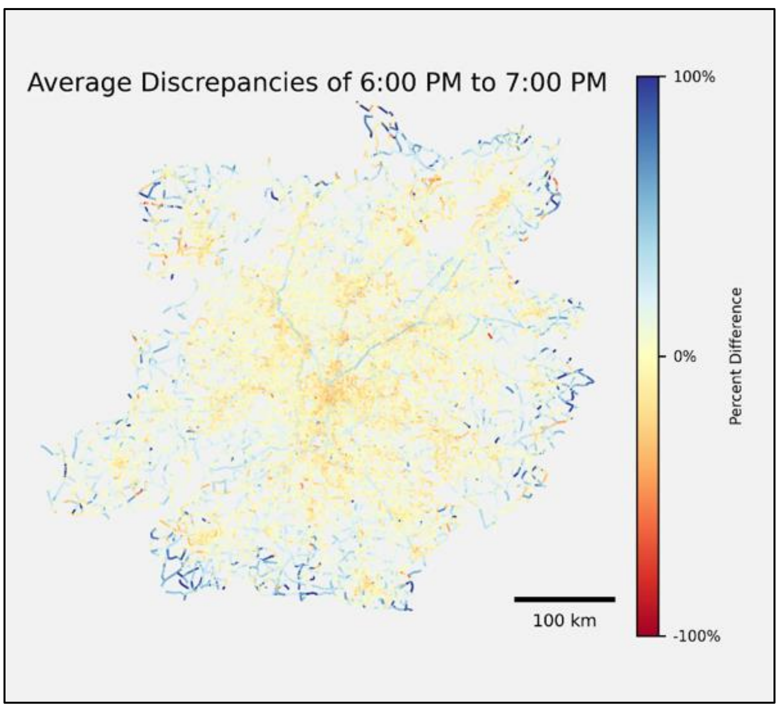

Figure 37. Link-by-link average differences between the scenarios, 6 PM to 7 PM.

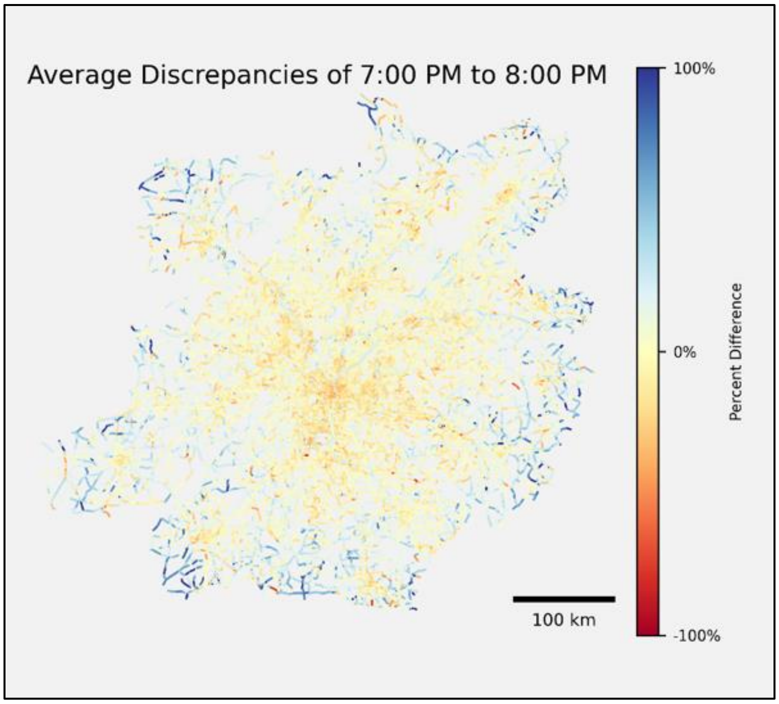

Figure 38. Link-by-link average differences between the scenarios, 7 PM to 8 PM.

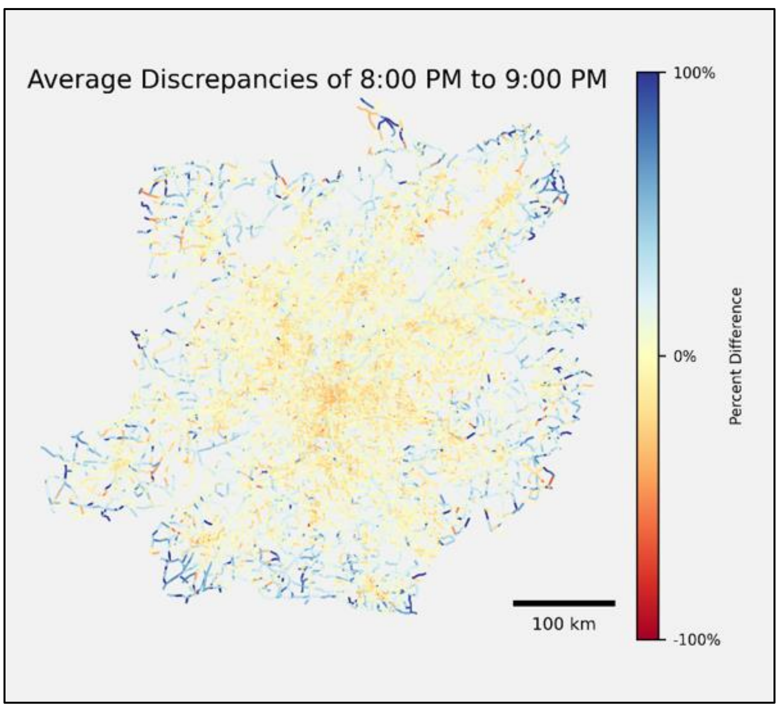

Figure 39. Link-by-link average differences between the scenarios, 8 PM to 9 PM.

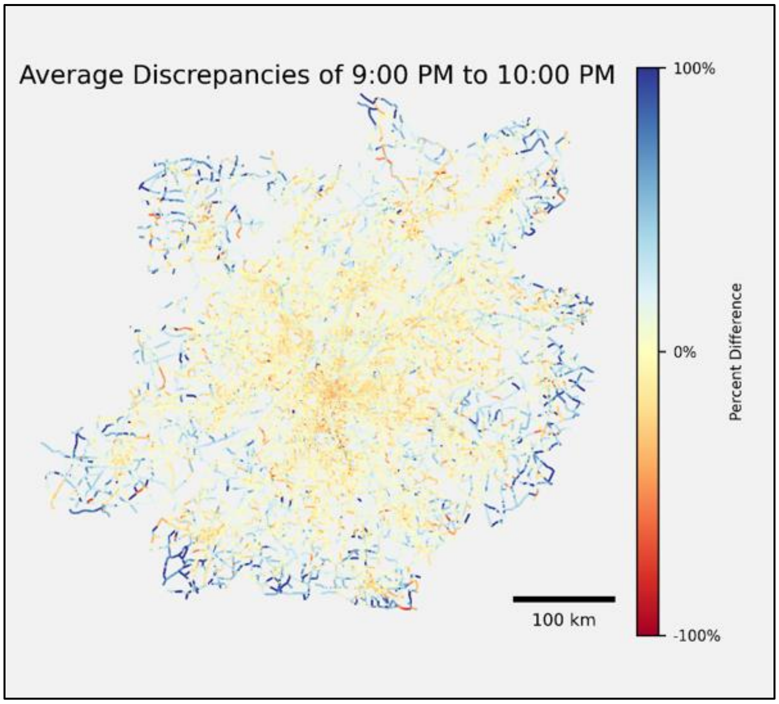

Figure 40. Link-by-link average differences between the scenarios, 9 PM to 10 PM.

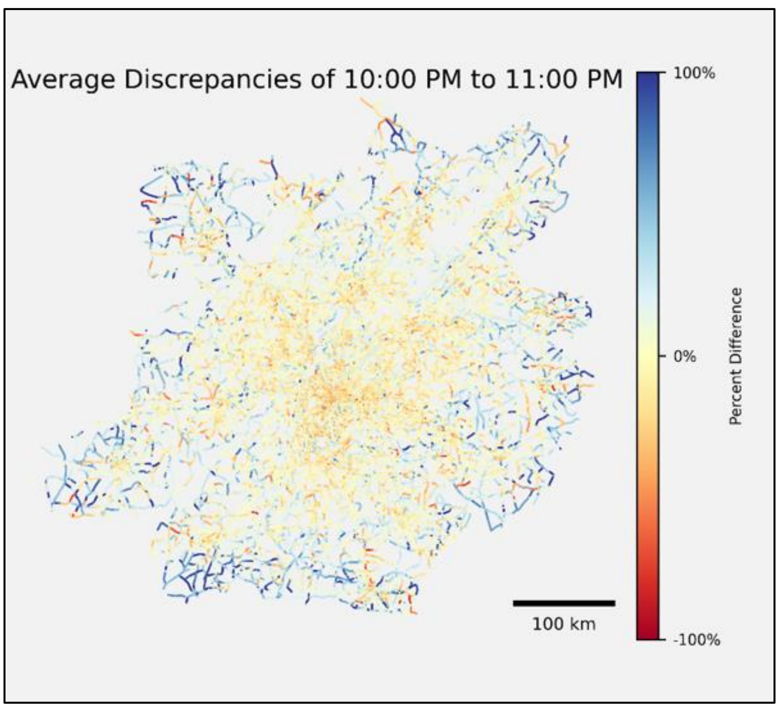

Figure 41. Link-by-link average differences between the scenarios, 10 PM to 11 PM.

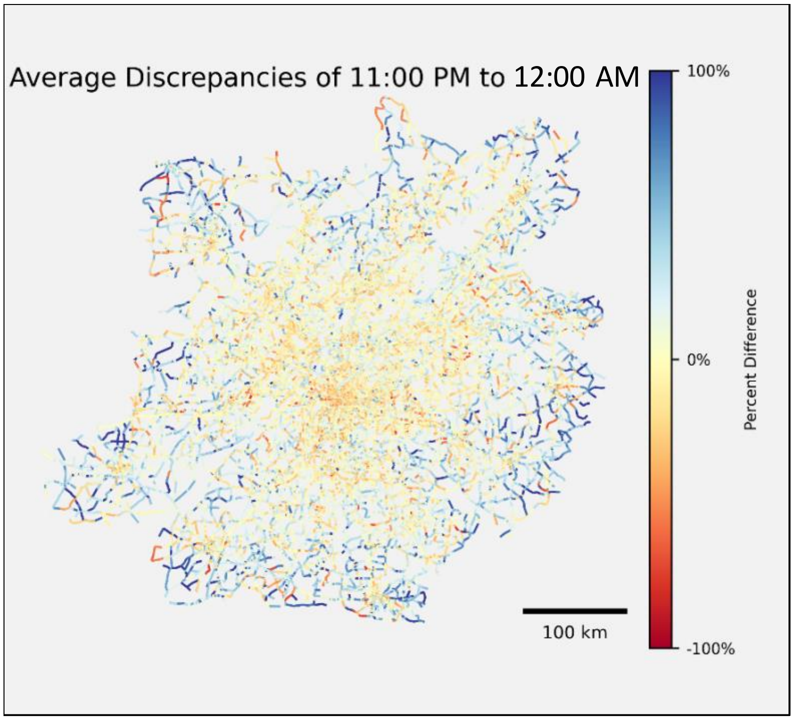

Figure 42. Link-by-link average differences between the scenarios, 11 PM to 12 AM.

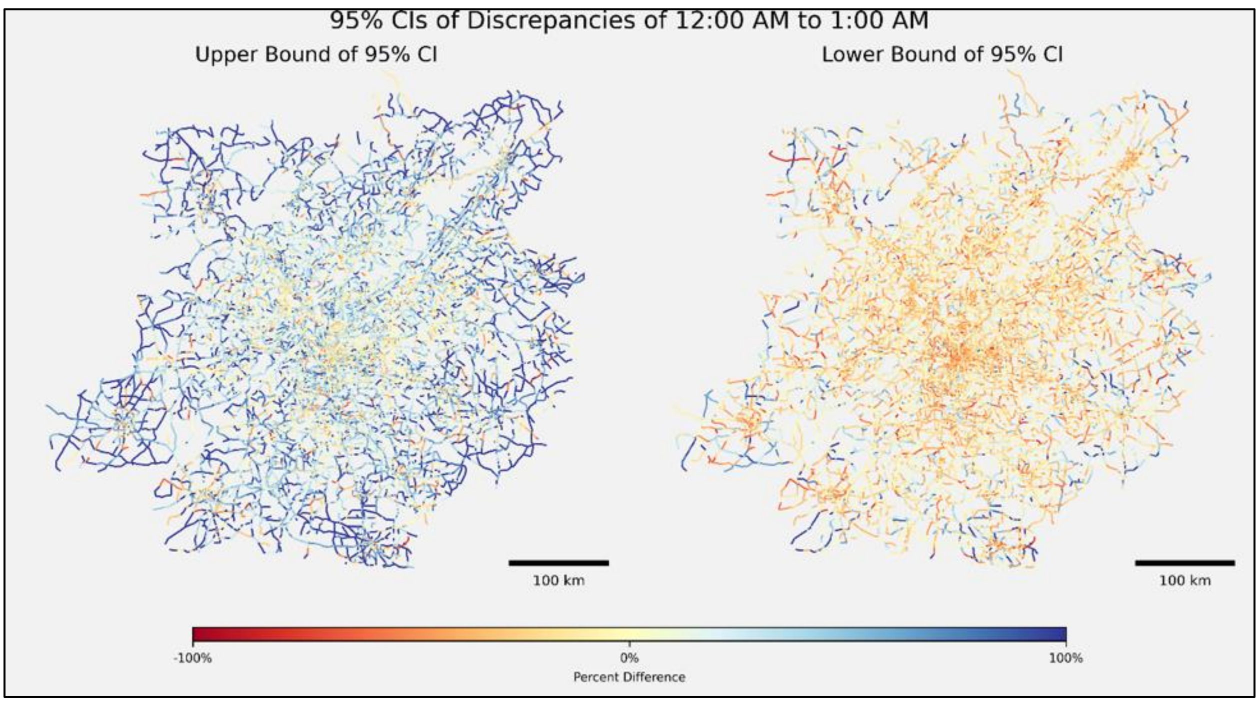

Figure 43. Link-by-link CIs of difference between the scenarios, 12 AM to 1 AM.

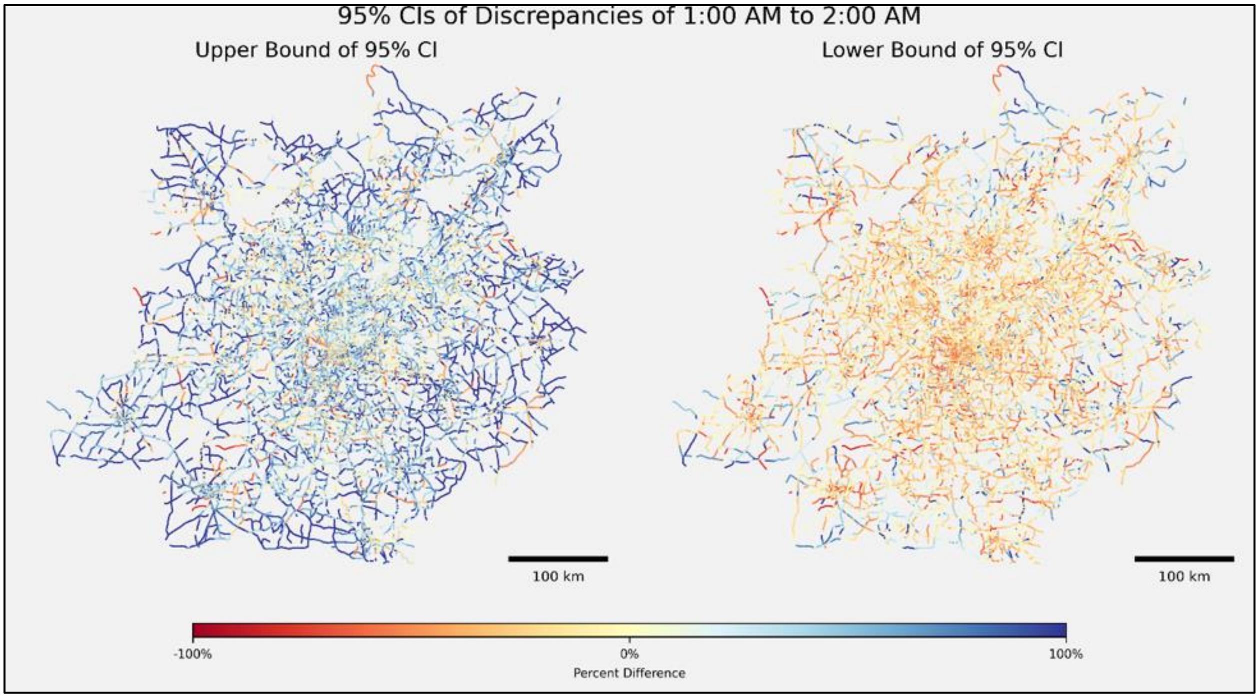

Figure 44. Link-by-link CIs of difference between the scenarios, 1 AM to 2 AM.

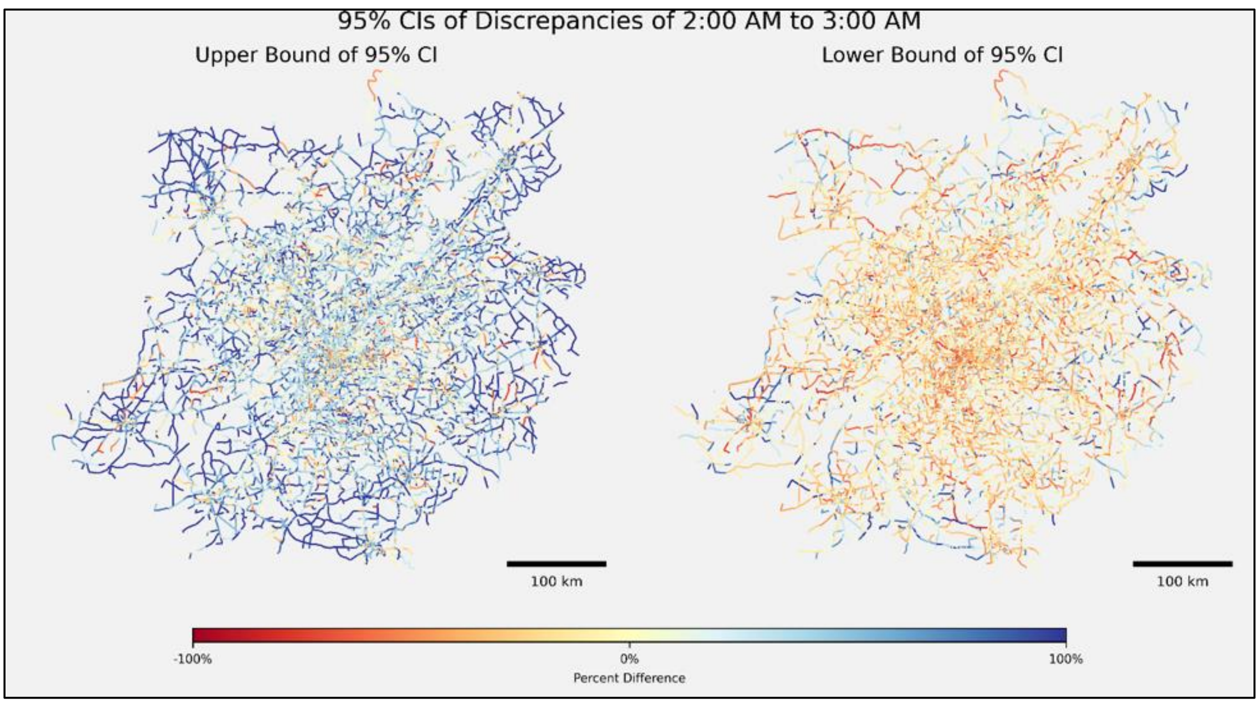

Figure 45. Link-by-link CIs of difference between the scenarios, 2 AM to 3 AM.

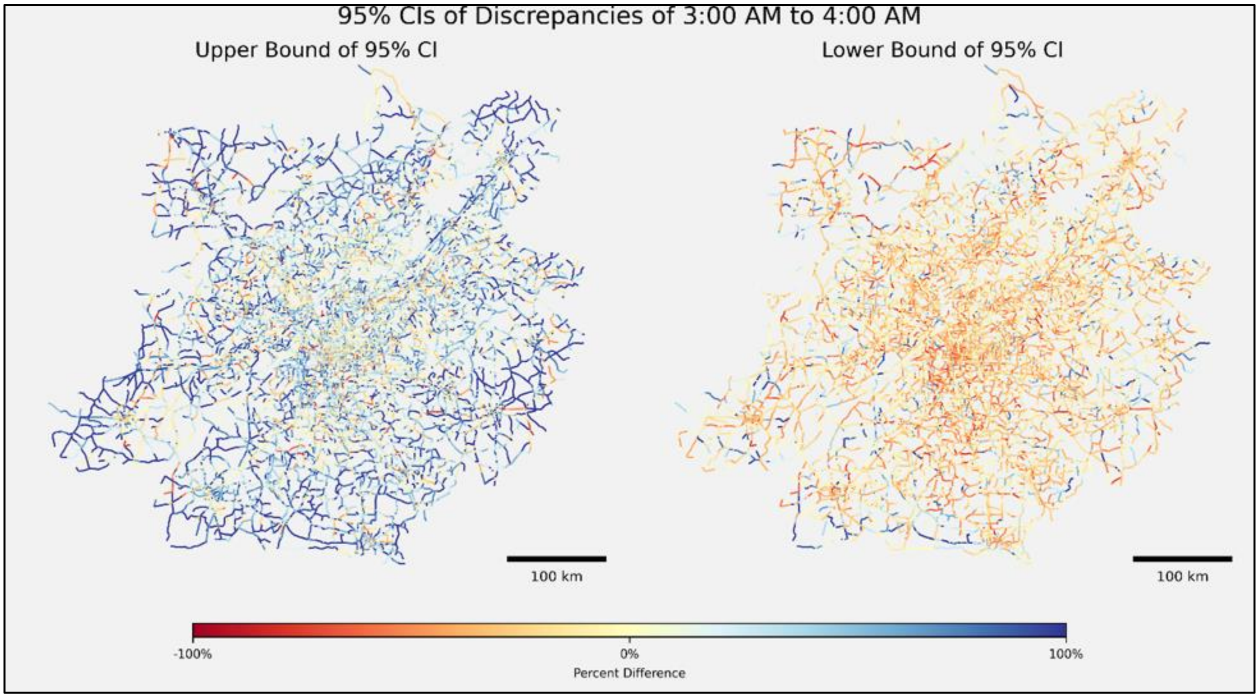

Figure 46. Link-by-link CIs of difference between the scenarios, 3 AM to 4 AM.

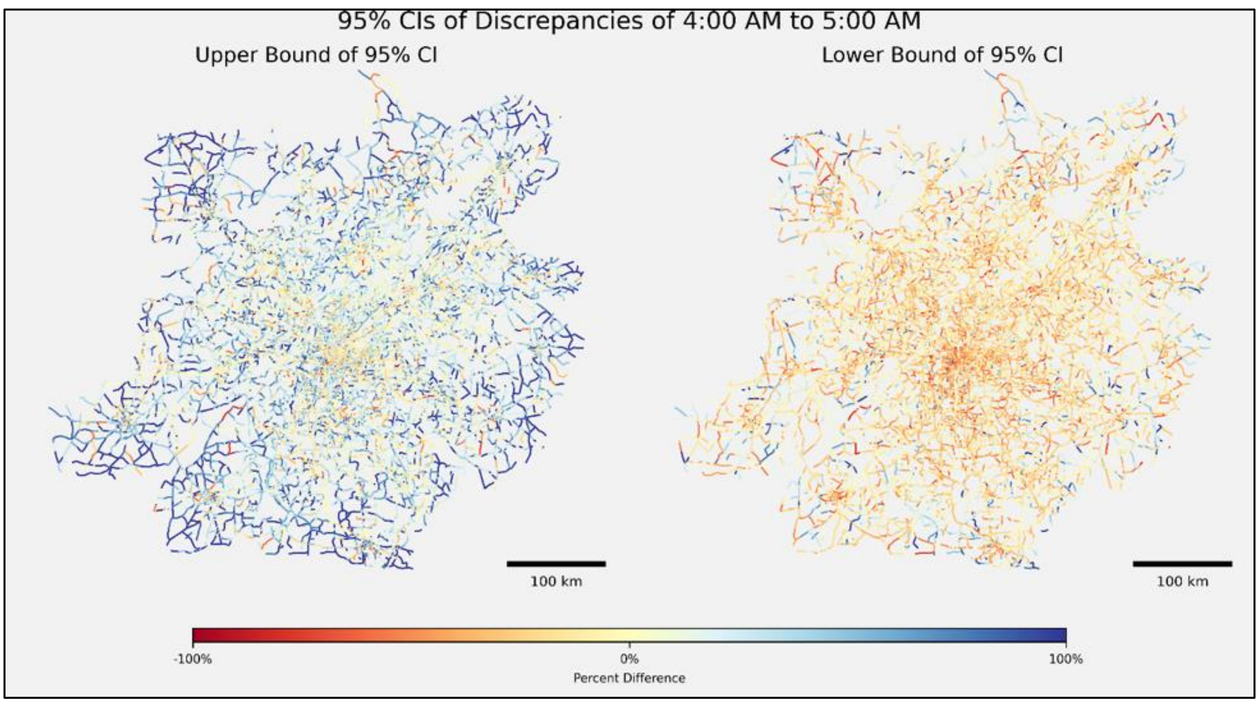

Figure 47. Link-by-link CIs of difference between the scenarios, 4 AM to 5 AM.

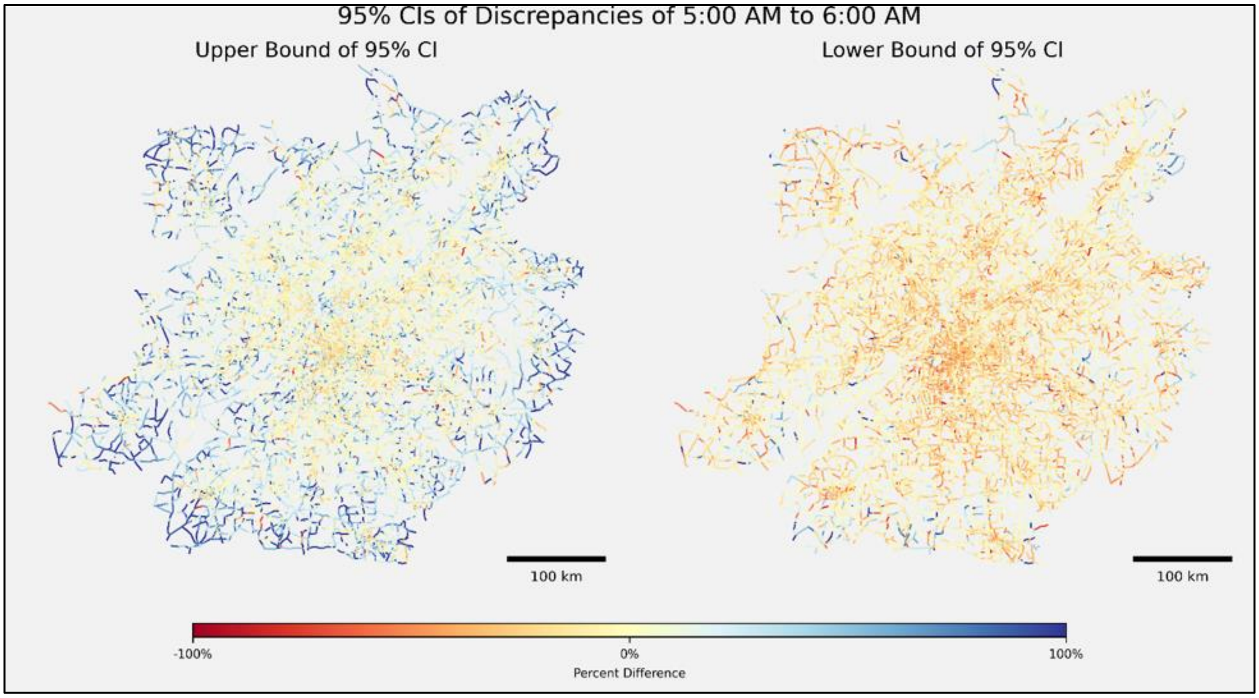

Figure 48. Link-by-link CIs of difference between the scenarios, 5 AM to 6 AM.

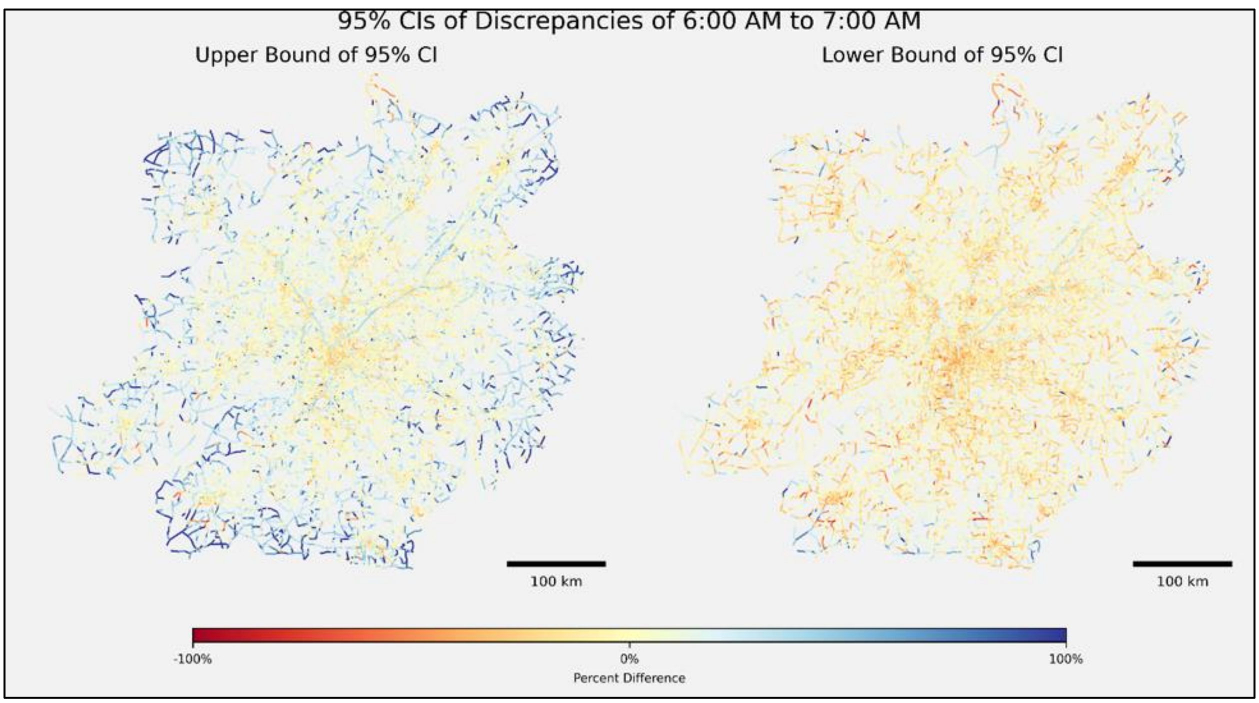

Figure 49. Link-by-link CIs of difference between the scenarios, 6 AM to 7 AM.

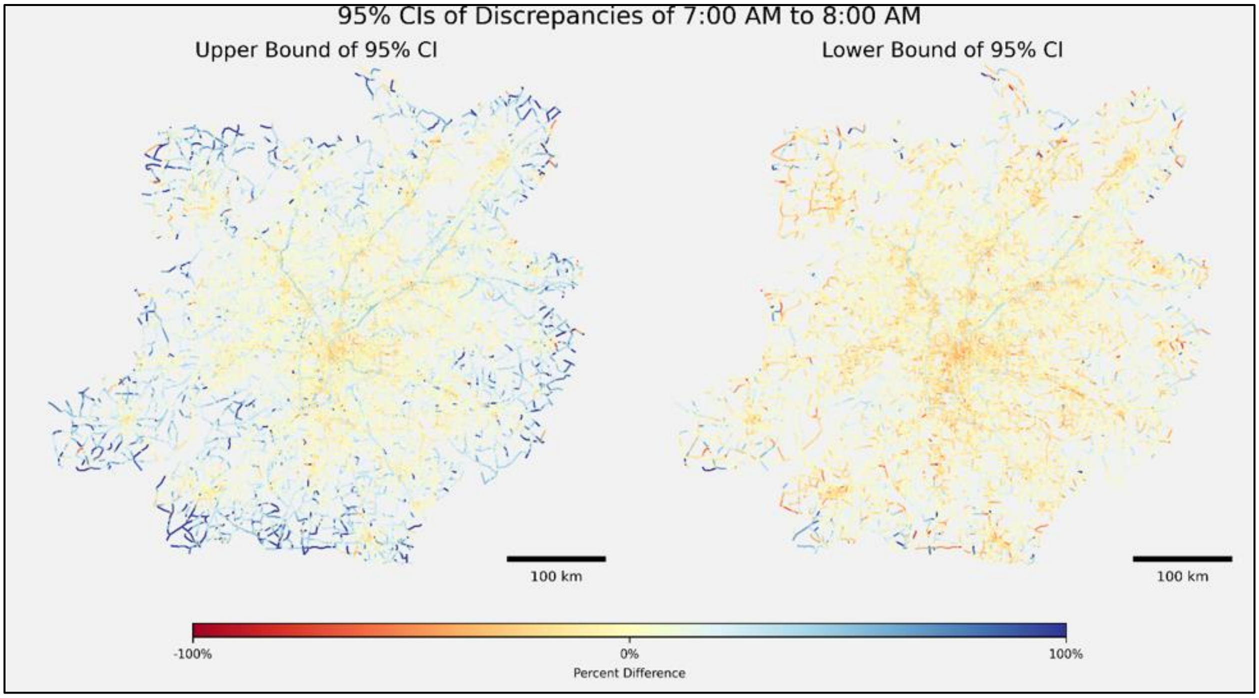

Figure 50. Link-by-link CIs of difference between the scenarios, 7 AM to 8 AM.

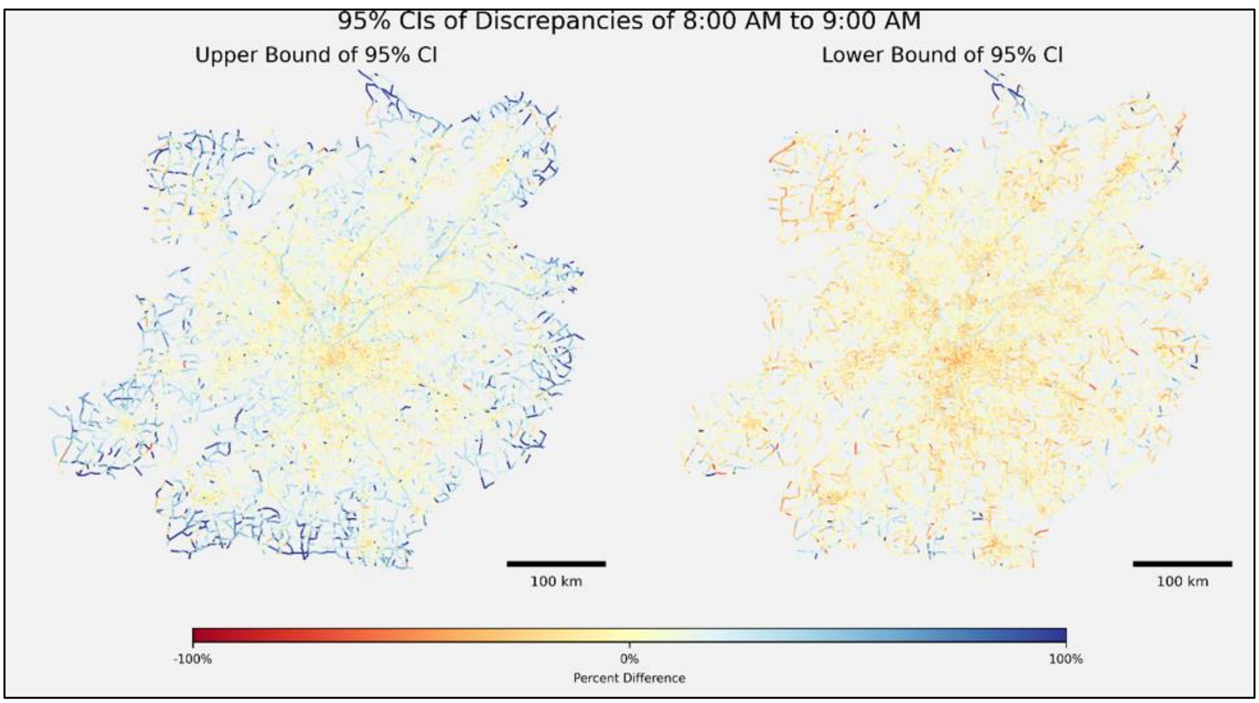

Figure 51. Link-by-link CIs of difference between the scenarios, 8 AM to 9 AM.

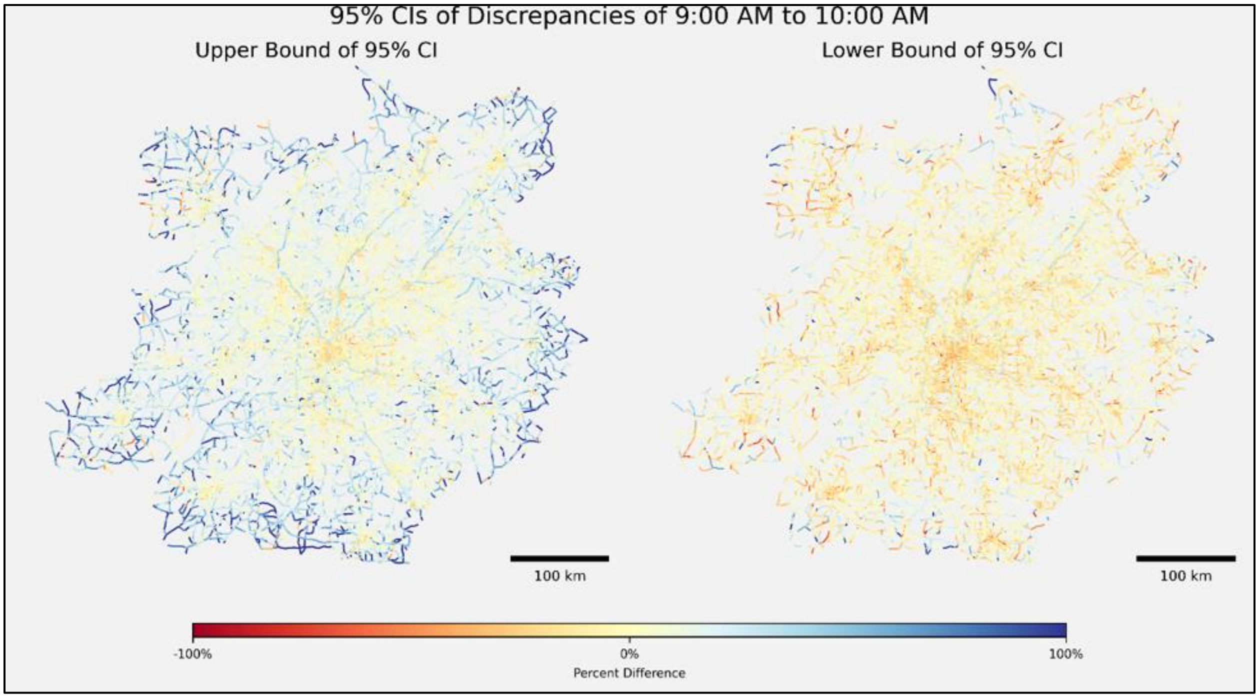

Figure 52. Link-by-link CIs of difference between the scenarios, 9 AM to 10 AM.

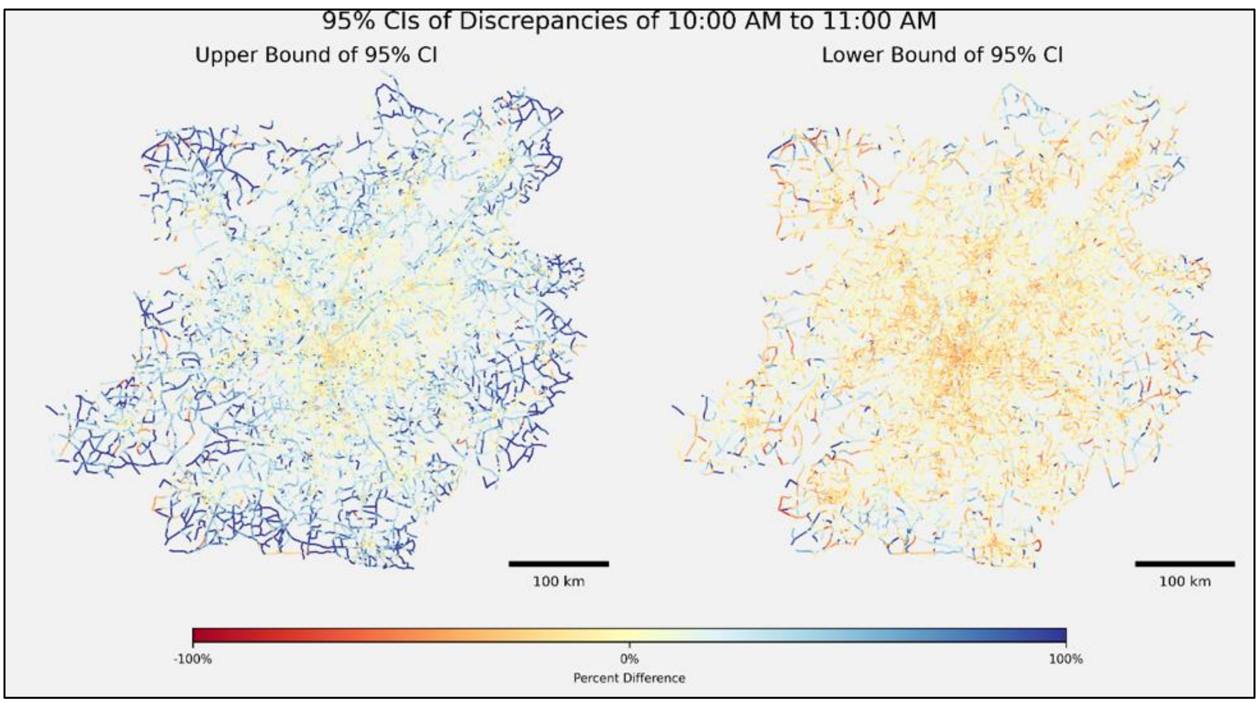

Figure 53. Link-by-link CIs of difference between the scenarios, 10 AM to 11 AM.

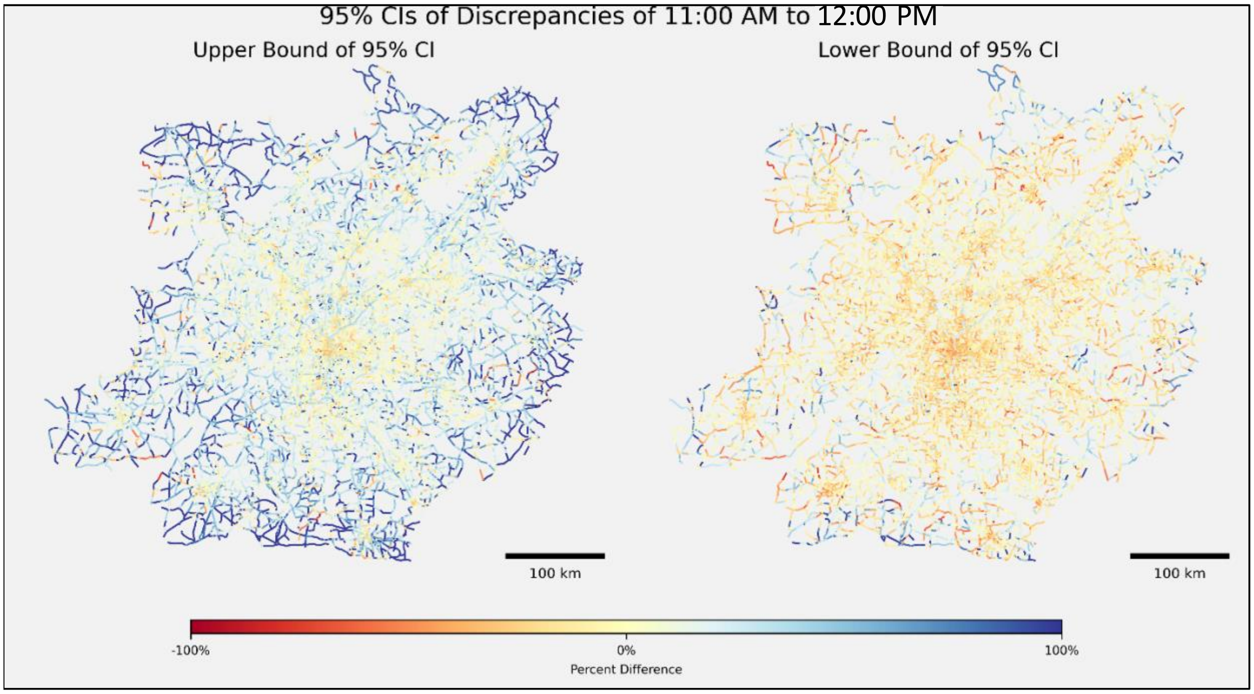

Figure 54. Link-by-link CIs of difference between the scenarios, 11 AM to 12 PM.

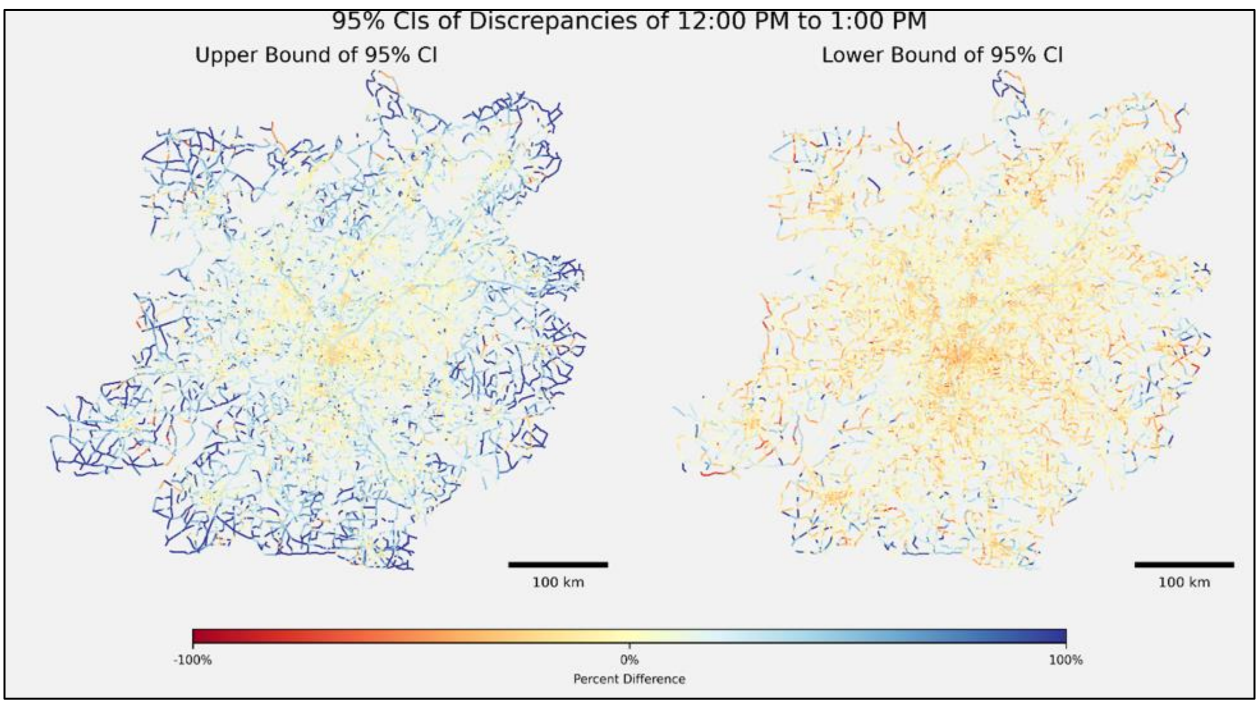

Figure 55. Link-by-link CIs of difference between the scenarios, 12 PM to 1 PM.

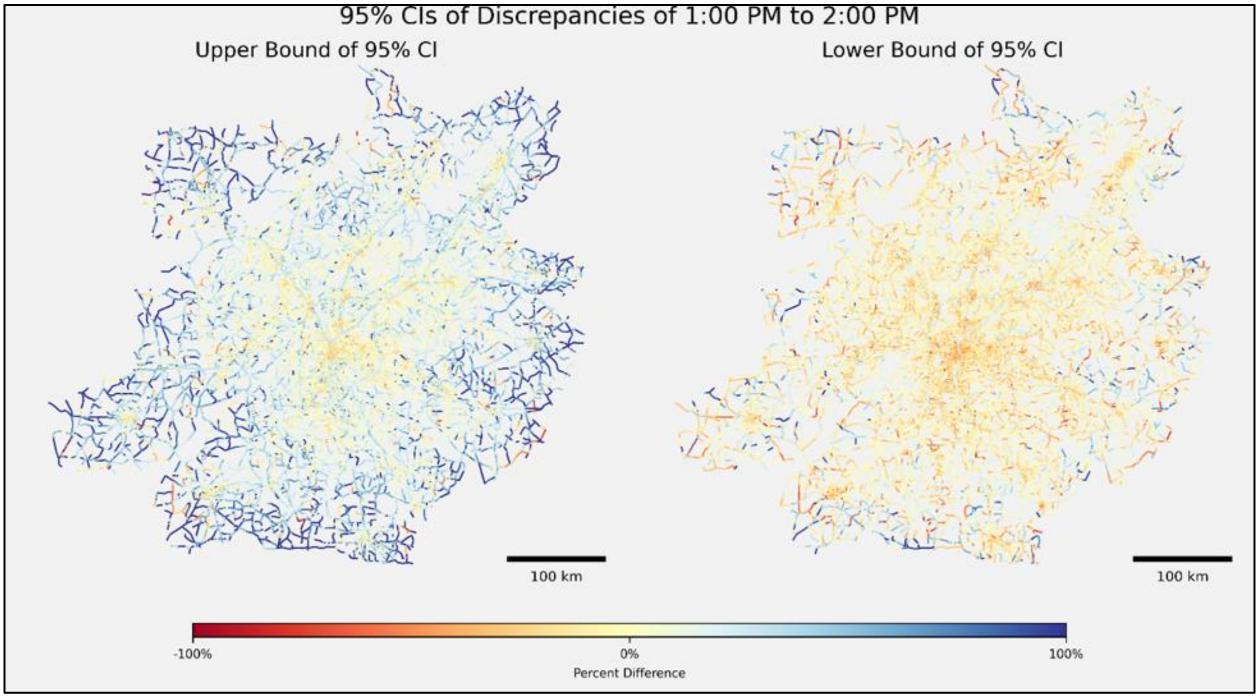

Figure 56. Link-by-link CIs of difference between the scenarios, 1 PM to 2 PM.

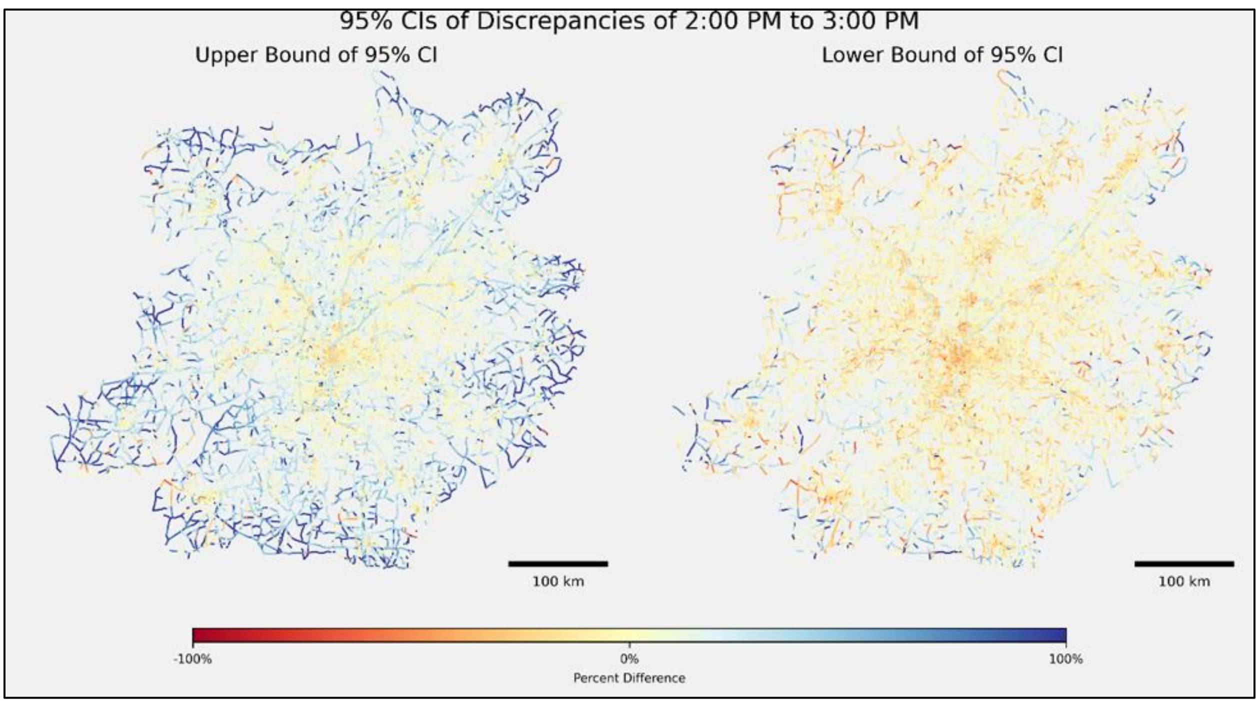

Figure 57. Link-by-link CIs of difference between the scenarios, 2 PM to 3 PM.

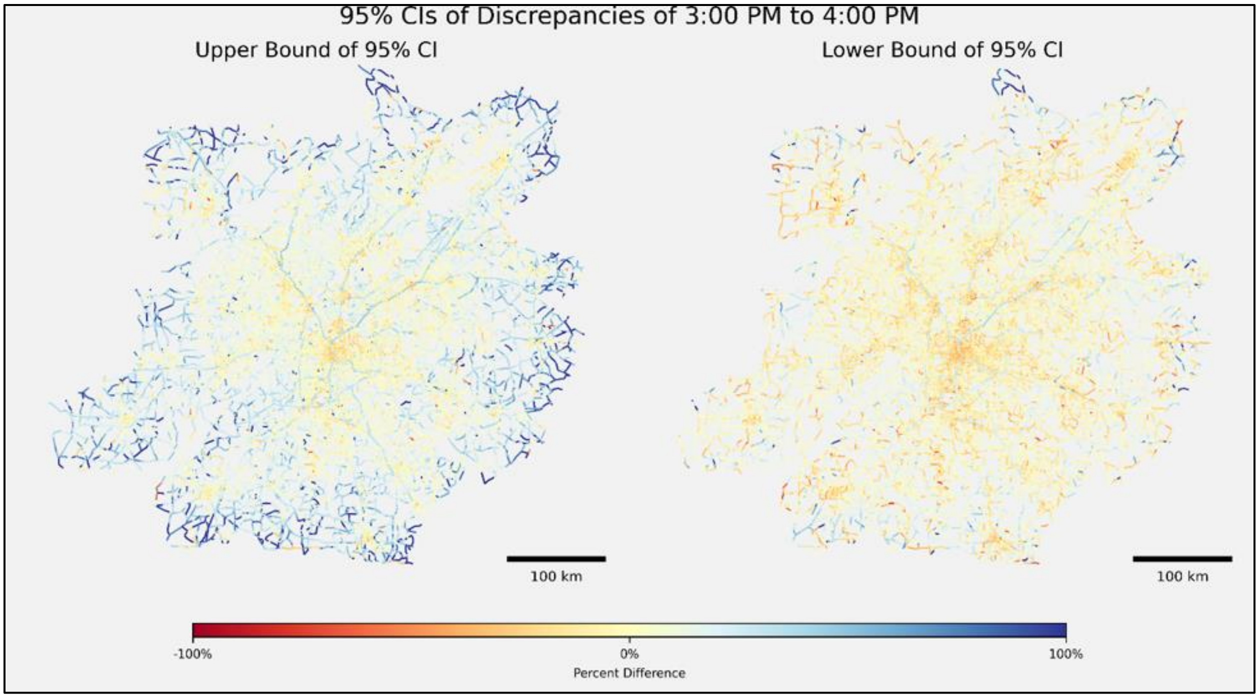

Figure 58. Link-by-link CIs of difference between the scenarios, 3 PM to 4 PM.

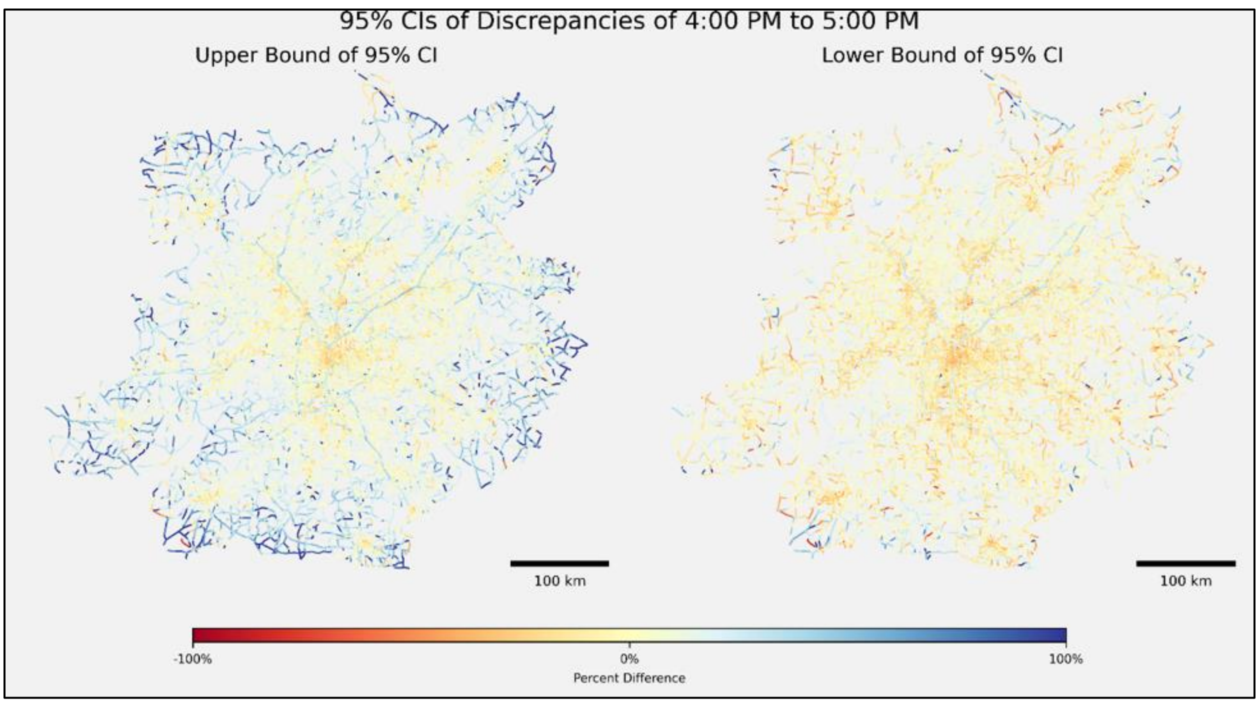

Figure 59. Link-by-link CIs of difference between the scenarios, 4 PM to 5 PM.

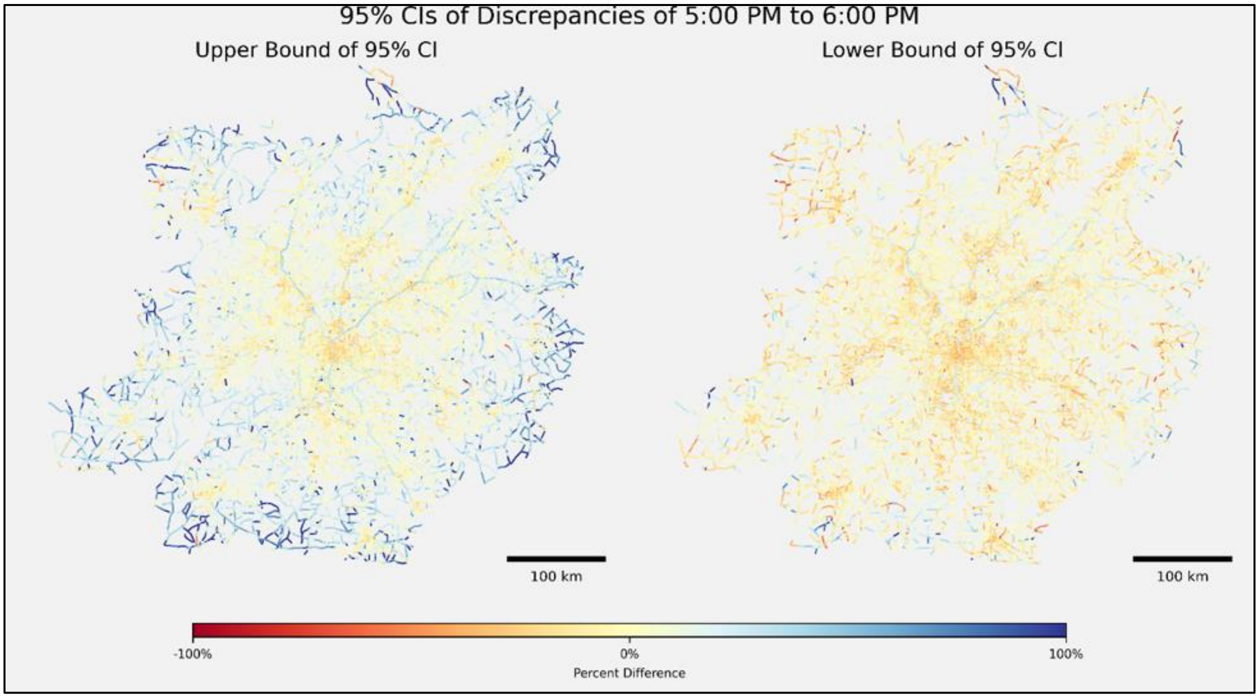

Figure 60. Link-by-link CIs of difference between the scenarios, 5 PM to 6 PM.

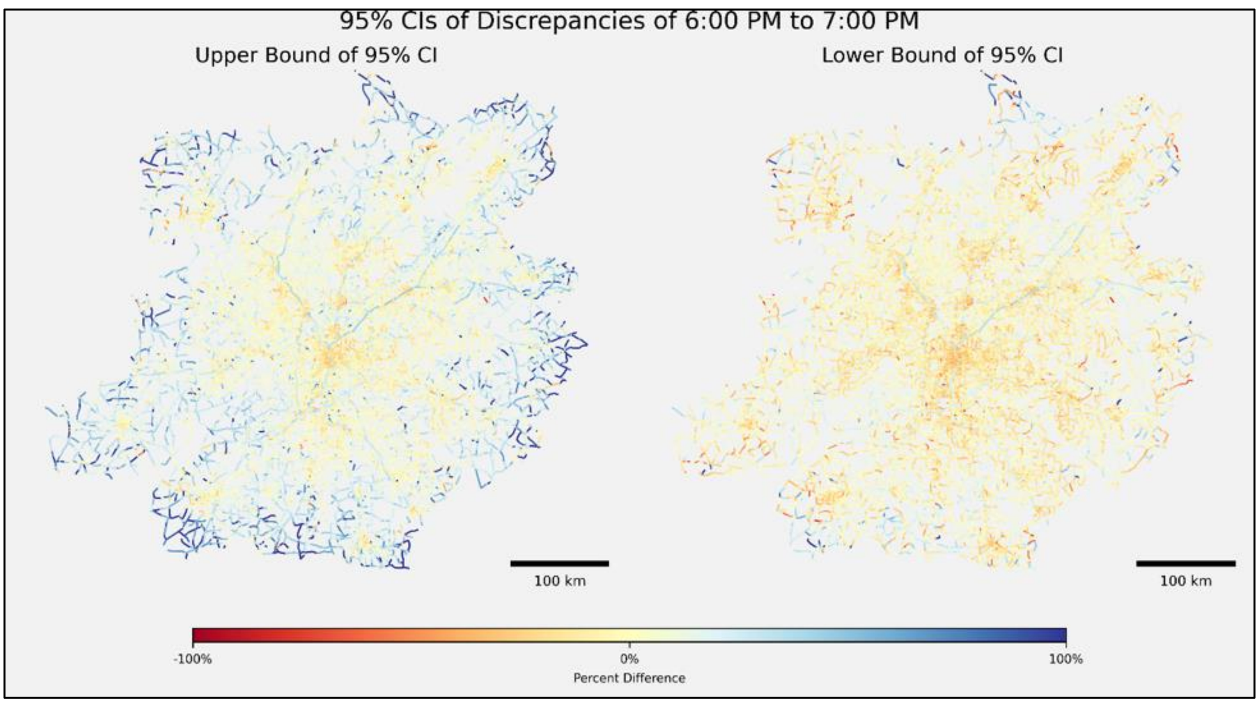

Figure 61. Link-by-link CIs of difference between the scenarios, 6 PM to 7 PM.

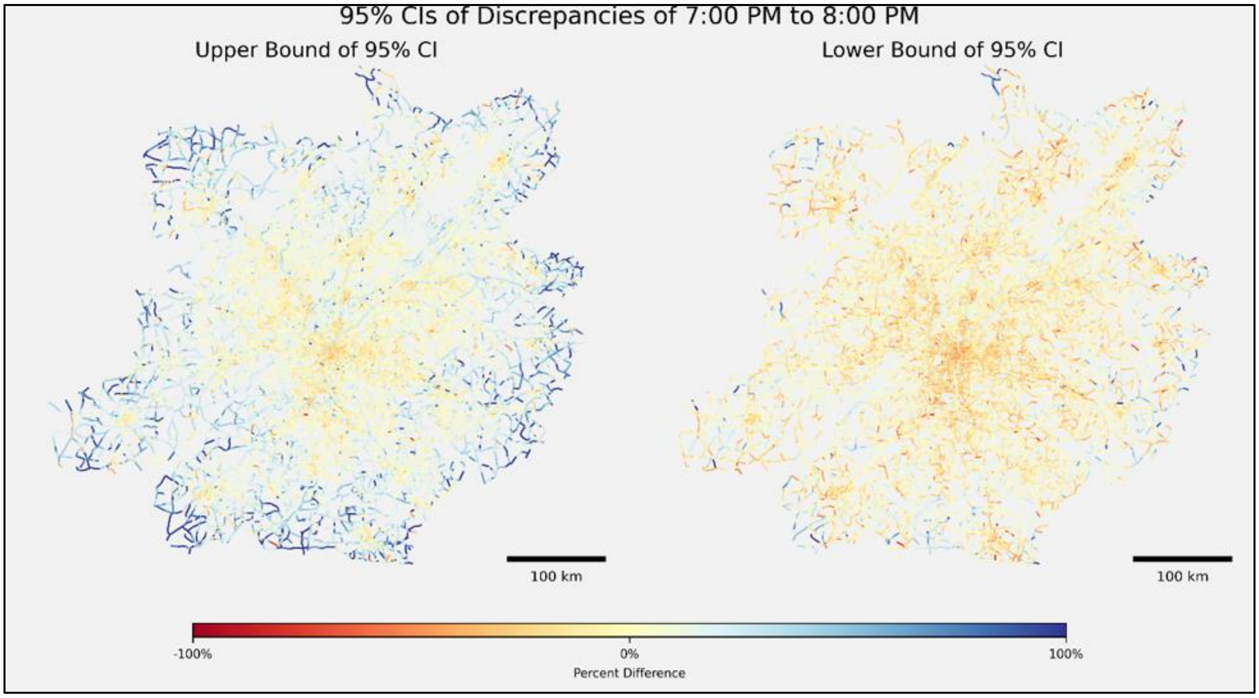

Figure 62. Link-by-link CIs of difference between the scenarios, 7 PM to 8 PM.

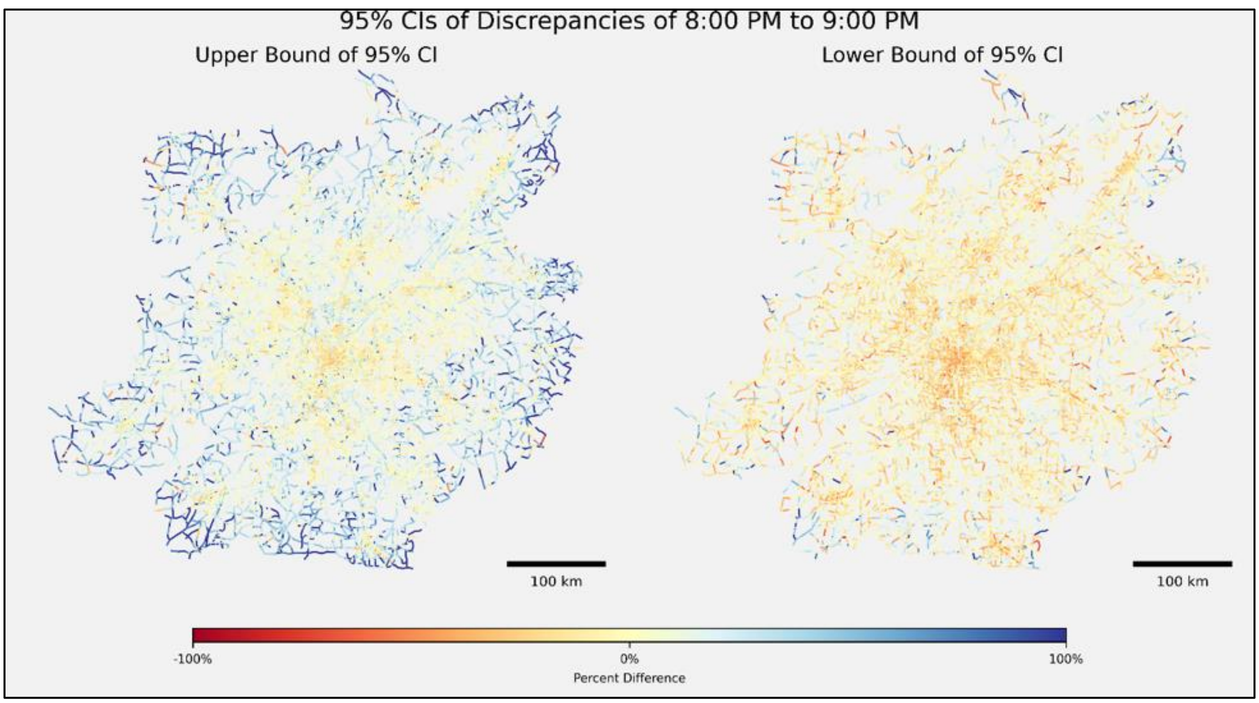

Figure 63. Link-by-link CIs of difference between the scenarios, 8 PM to 9 PM.

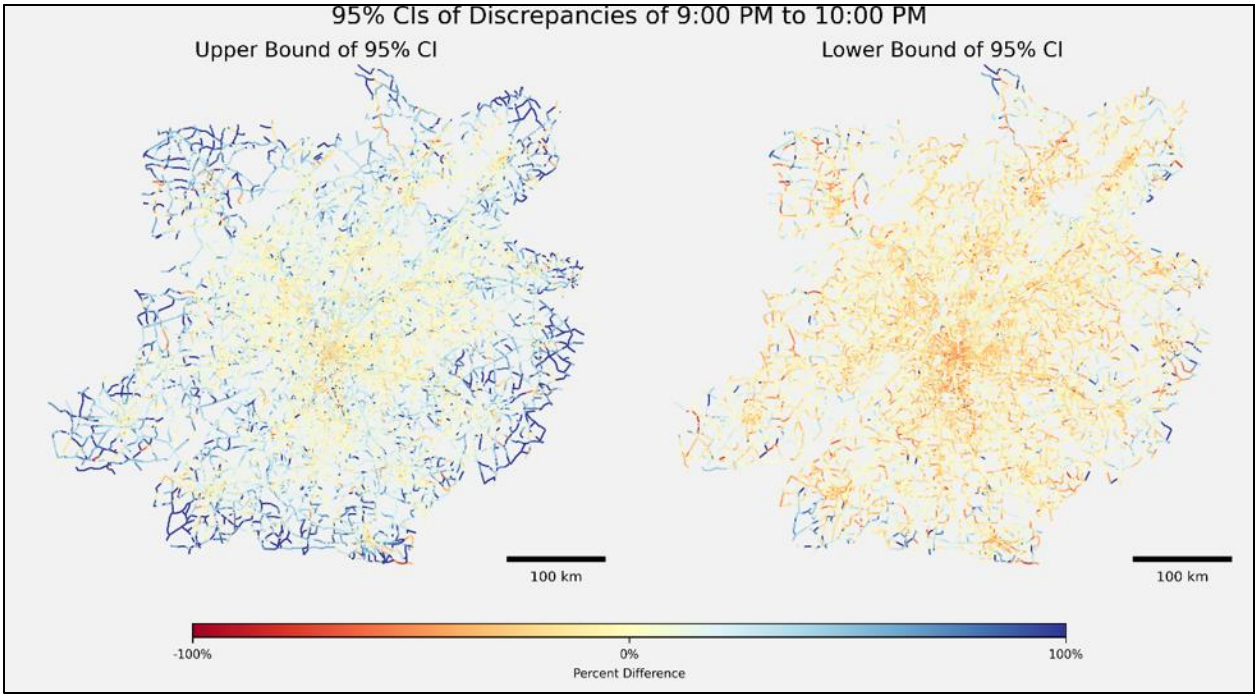

Figure 64. Link-by-link CIs of difference between the scenarios, 9 PM to 10 PM.

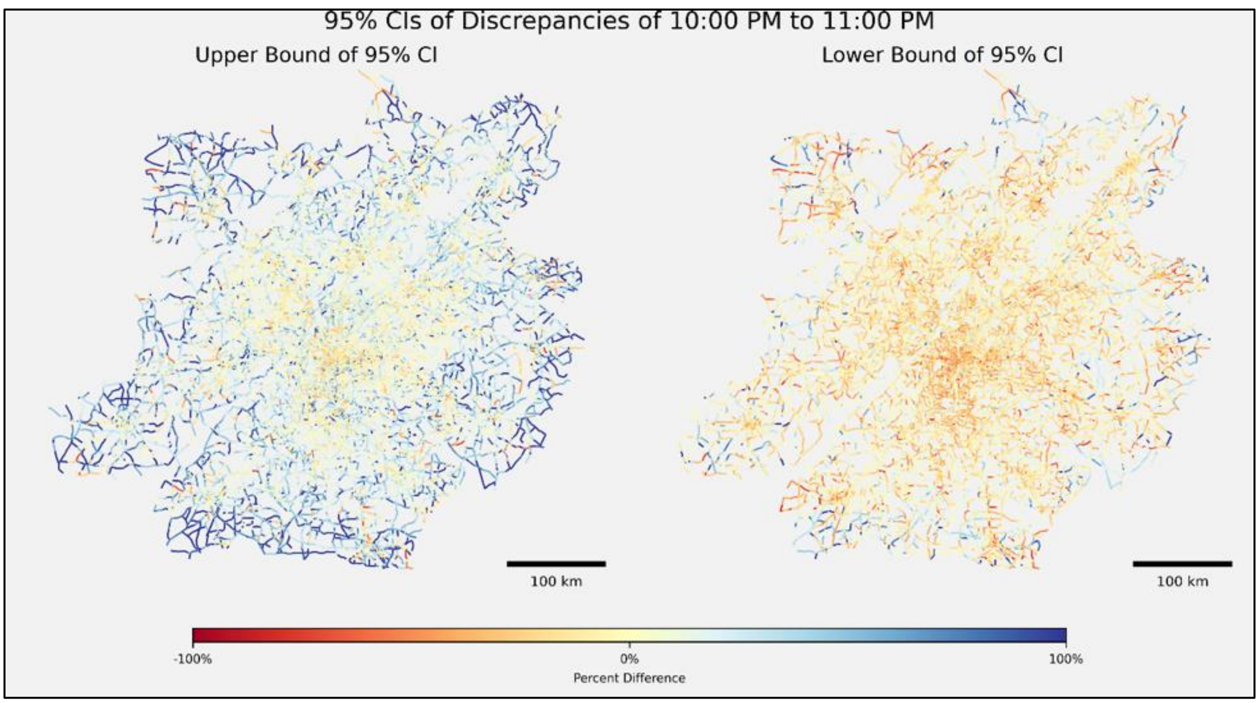

Figure 65. Link-by-link CIs of difference between the scenarios, 10 PM to 11 PM.

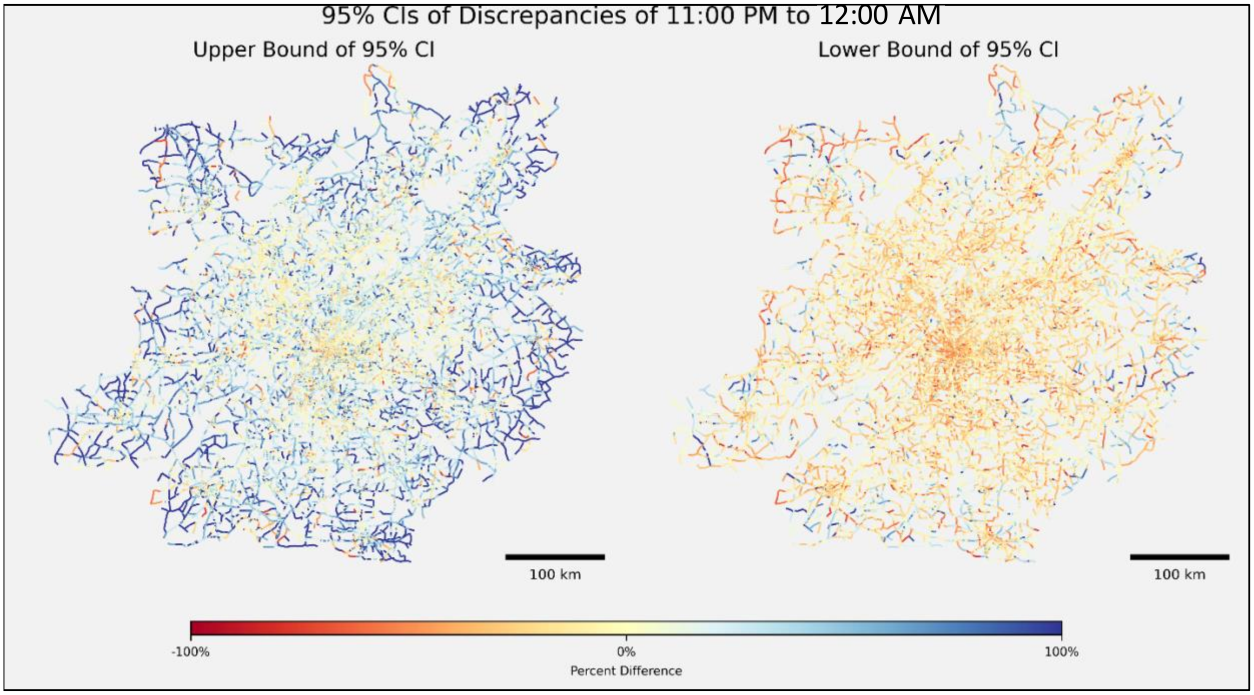

Figure 66. Link-by-link CIs of difference between the scenarios, 11 PM to 12 AM.

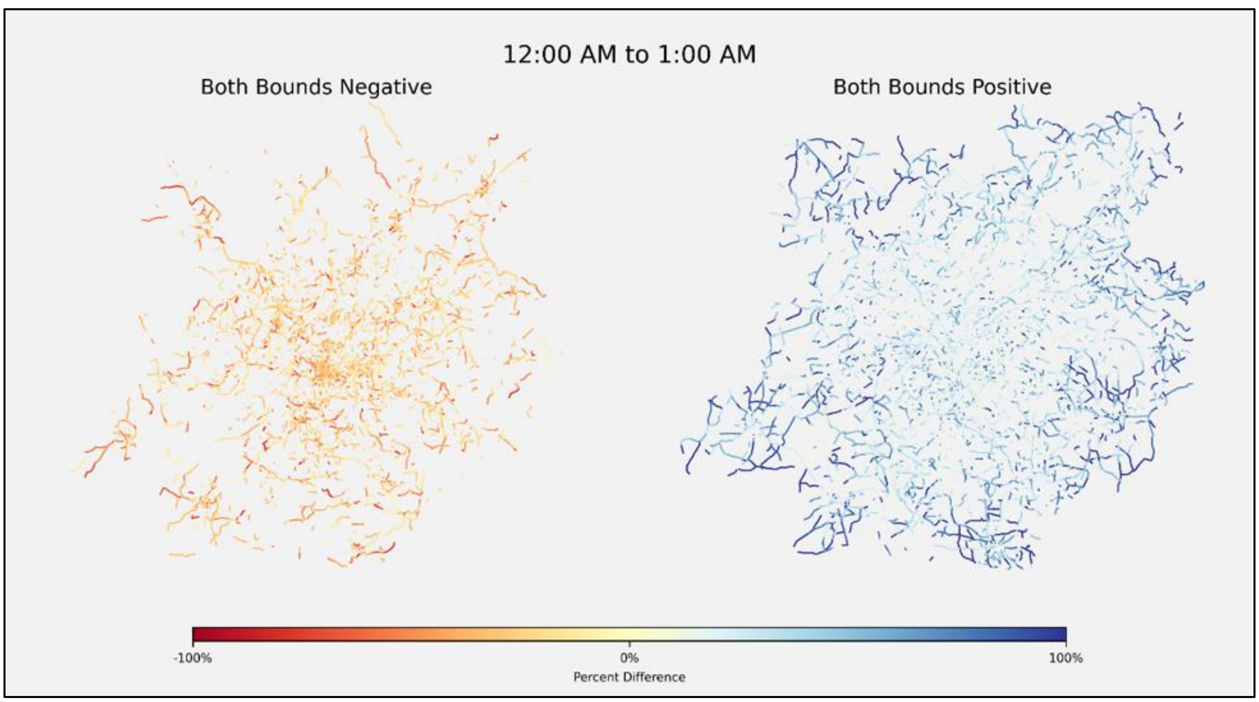

Figure 67. Links with both CI bounds negative or positive, 12 AM to 1 AM.

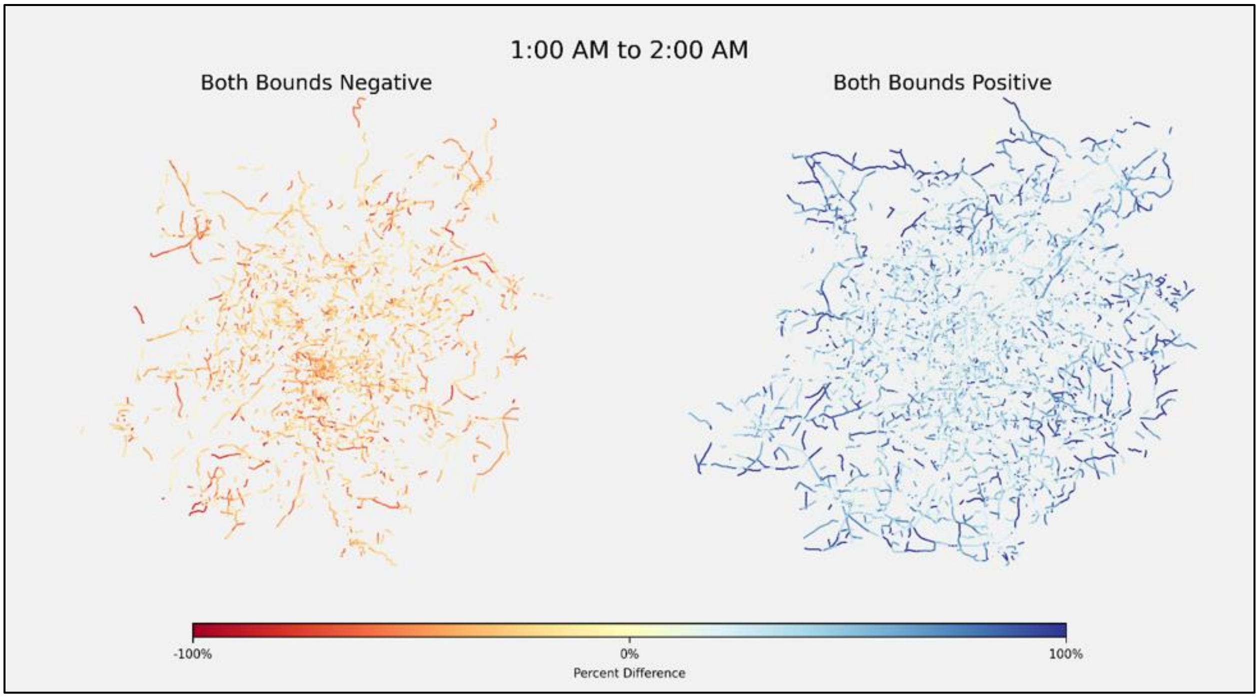

Figure 68. Links with both CI bounds negative or positive, 1 AM to 2 AM.

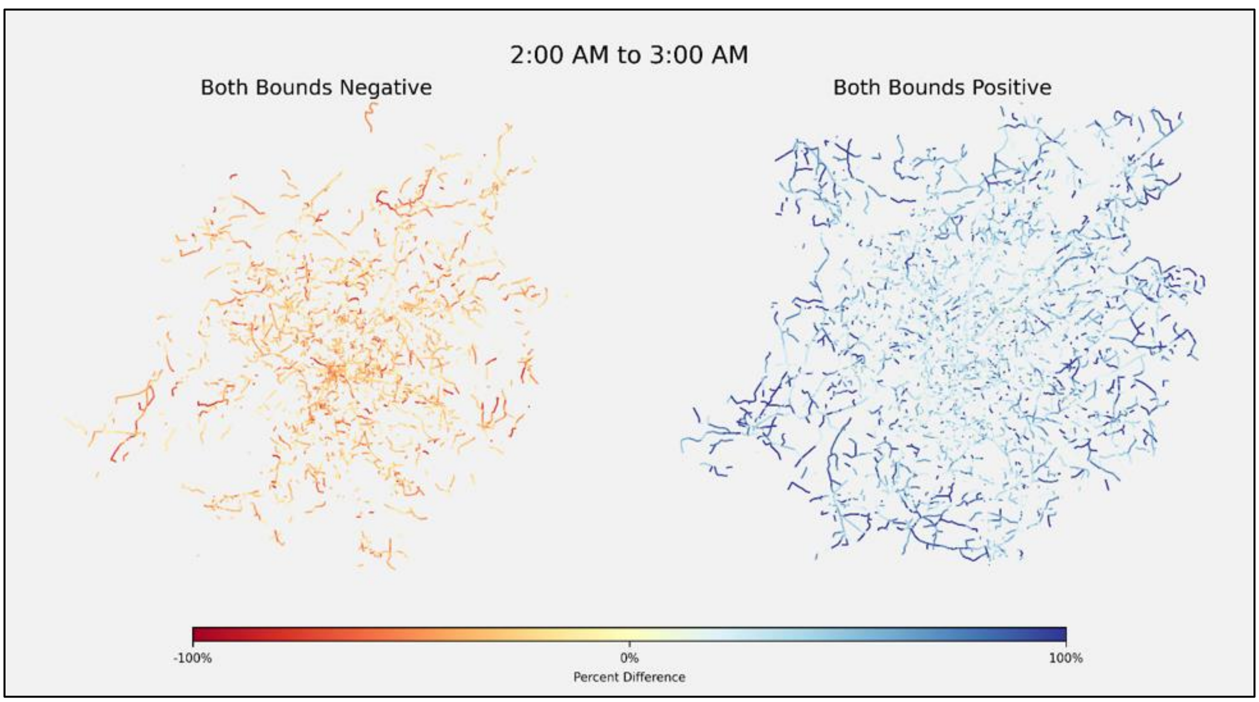

Figure 69. Links with both CI bounds negative or positive, 2 AM to 3 AM.

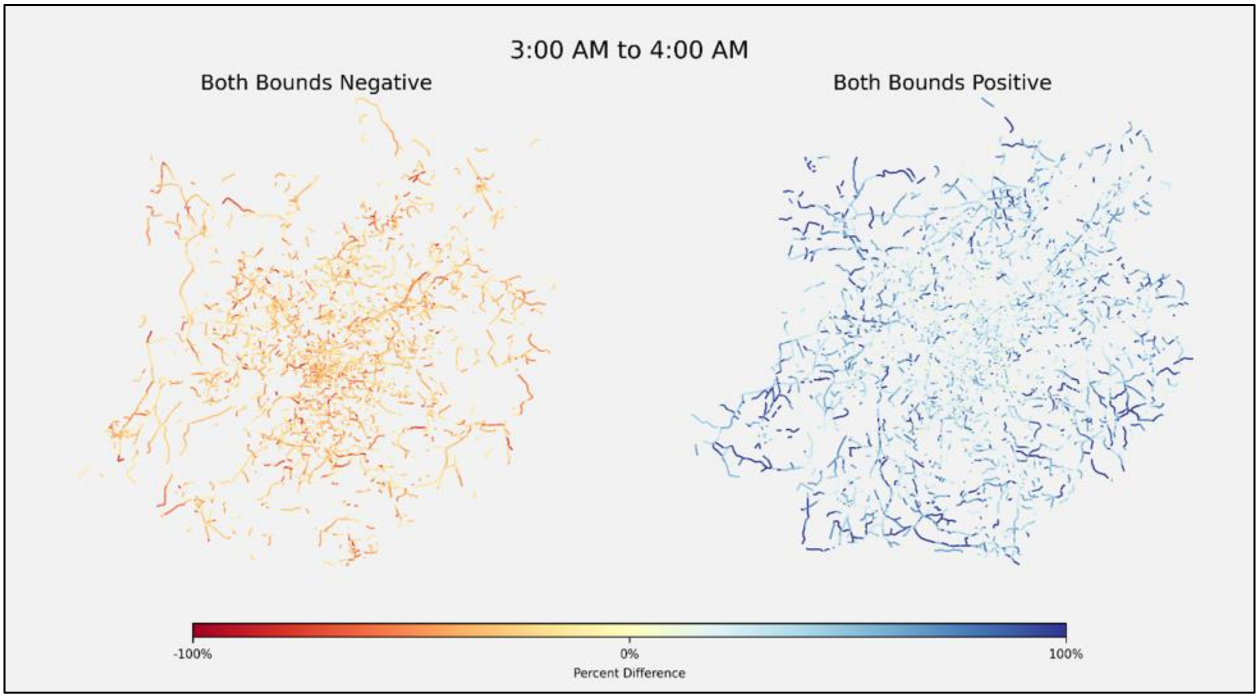

Figure 70. Links with both CI bounds negative or positive, 3 AM to 4 AM.

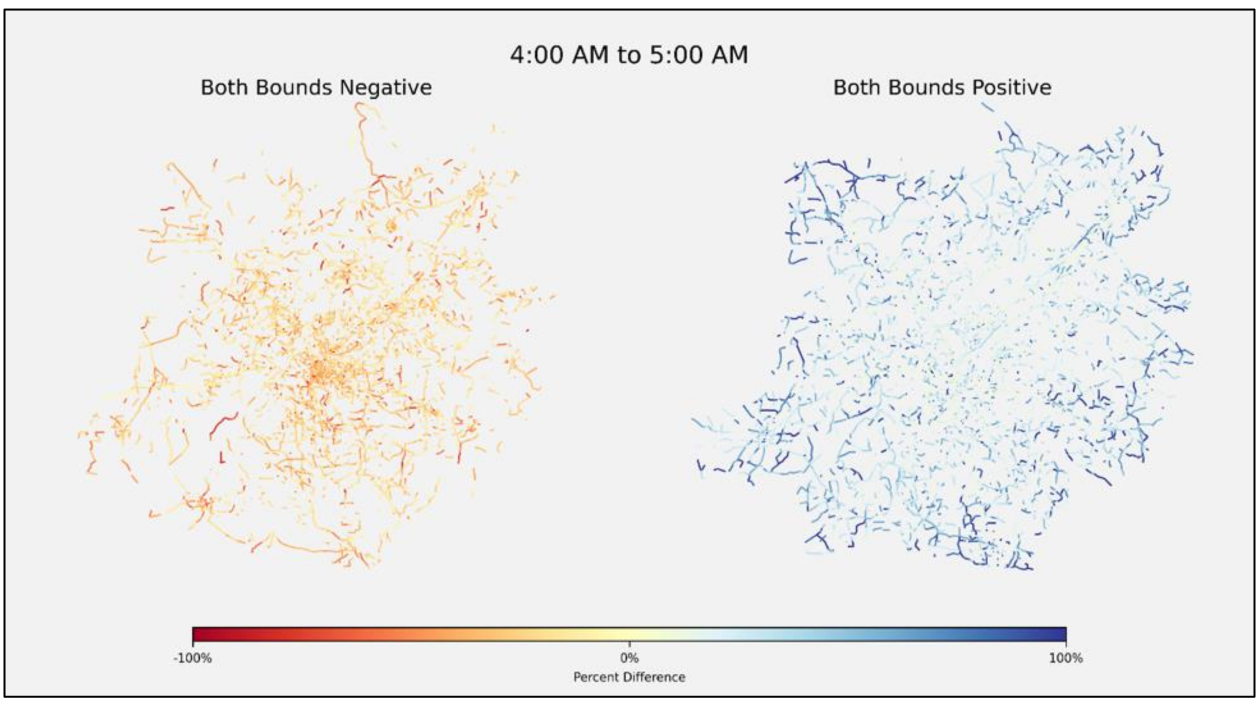

Figure 71. Links with both CI bounds negative or positive, 4 AM to 5 AM.

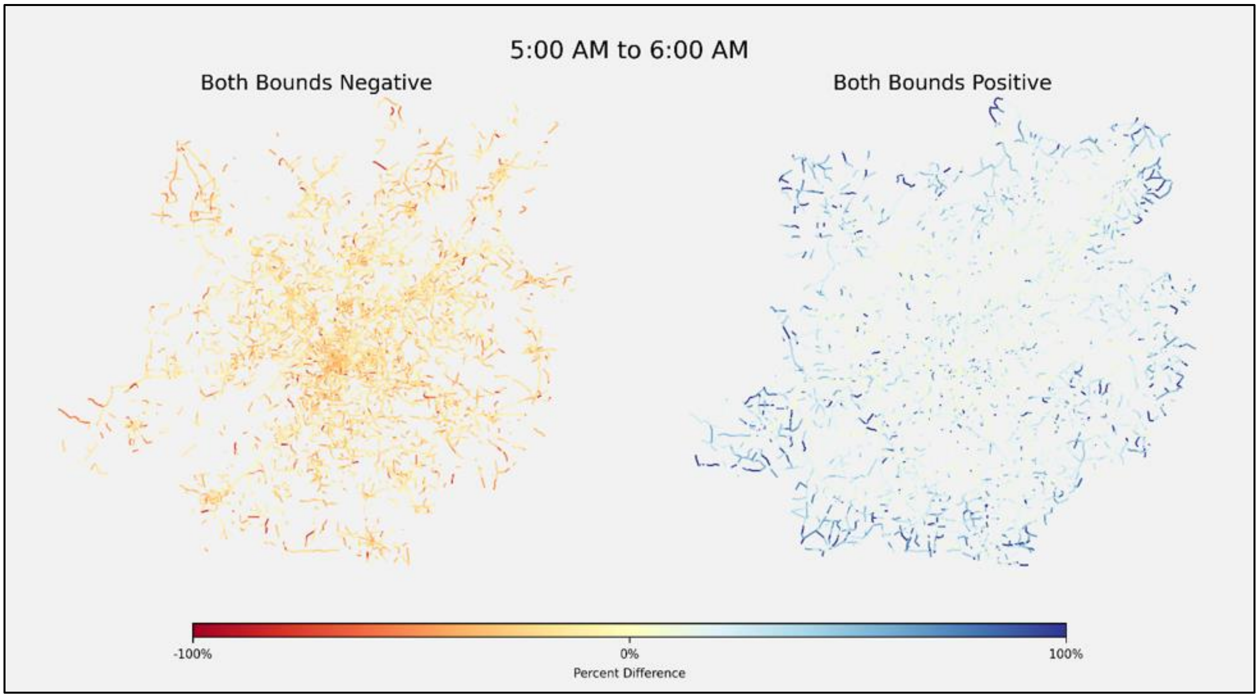

Figure 72. Links with both CI bounds negative or positive, 5 AM to 6 AM.

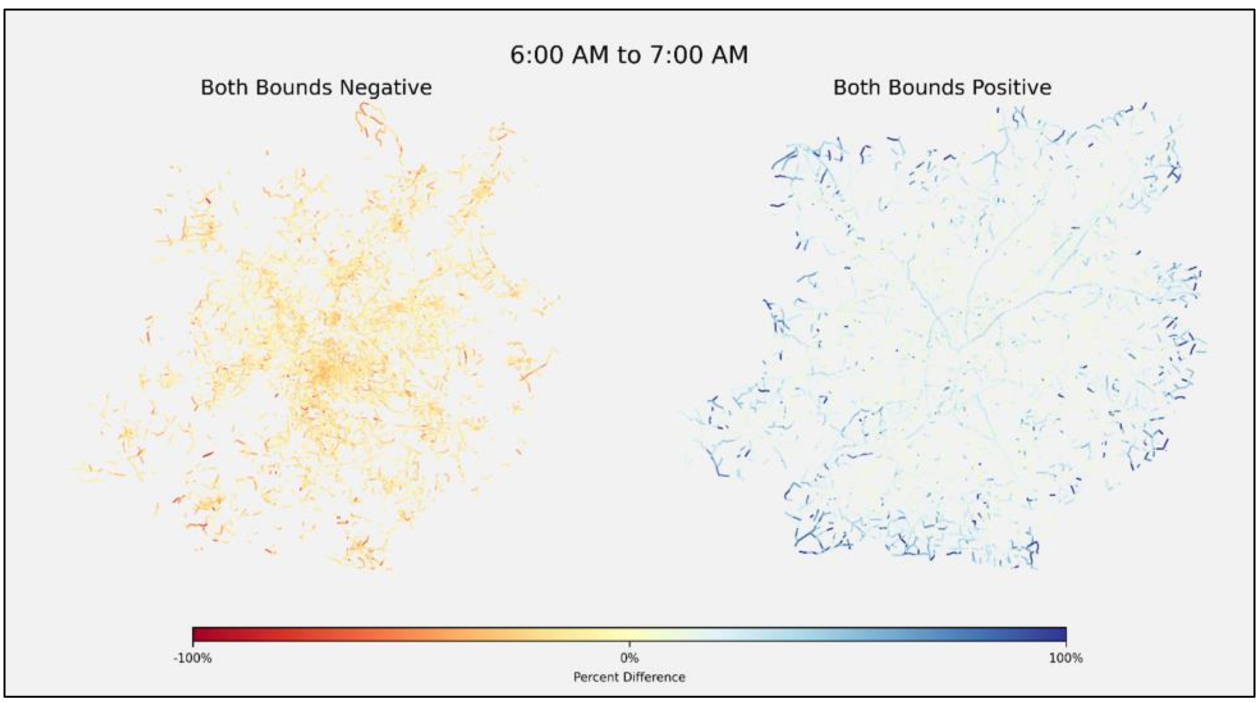

Figure 73. Links with both CI bounds negative or positive, 6 AM to 7 AM.

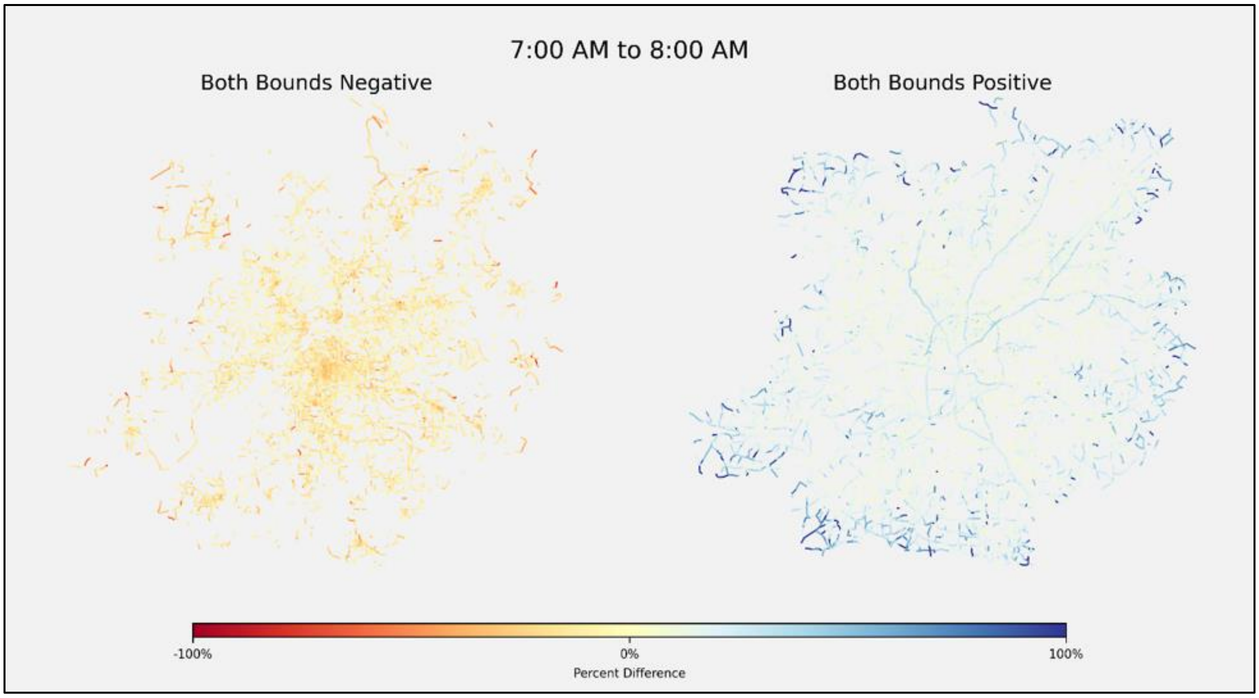

Figure 74. Links with both CI bounds negative or positive, 7 AM to 8 AM.

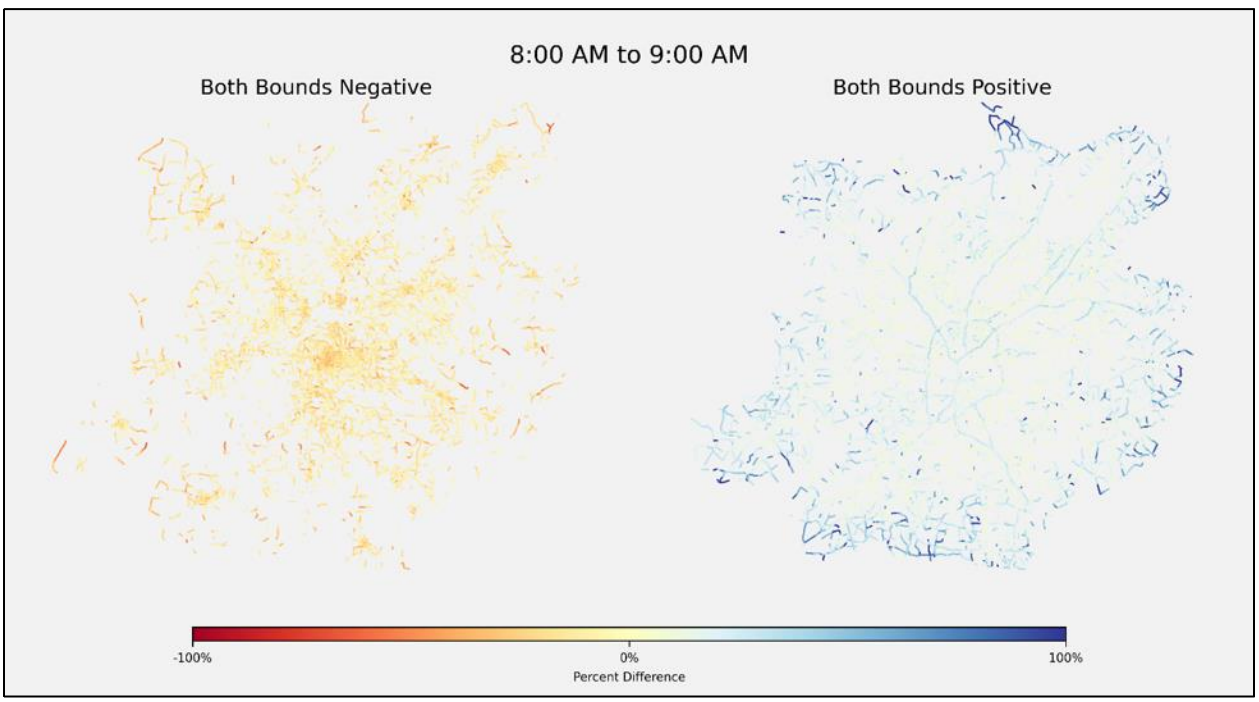

Figure 75. Links with both CI bounds negative or positive, 8 AM to 9 AM.

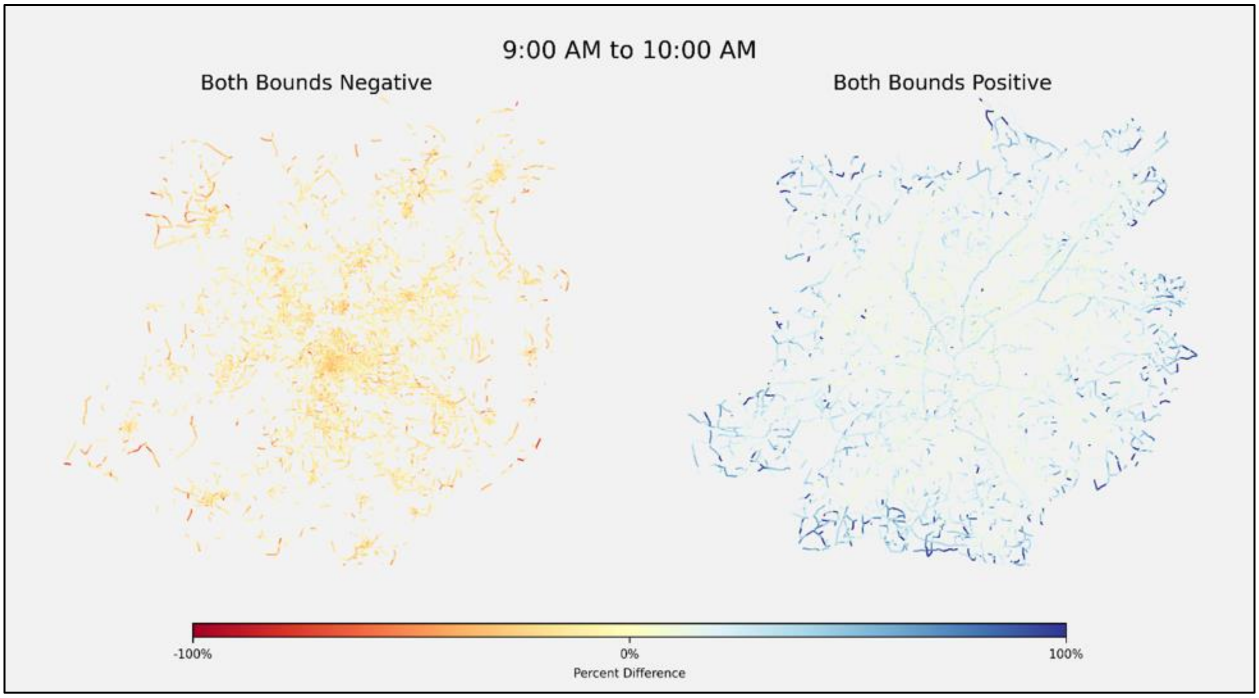

Figure 76. Links with both CI bounds negative or positive, 9 AM to 10 AM.

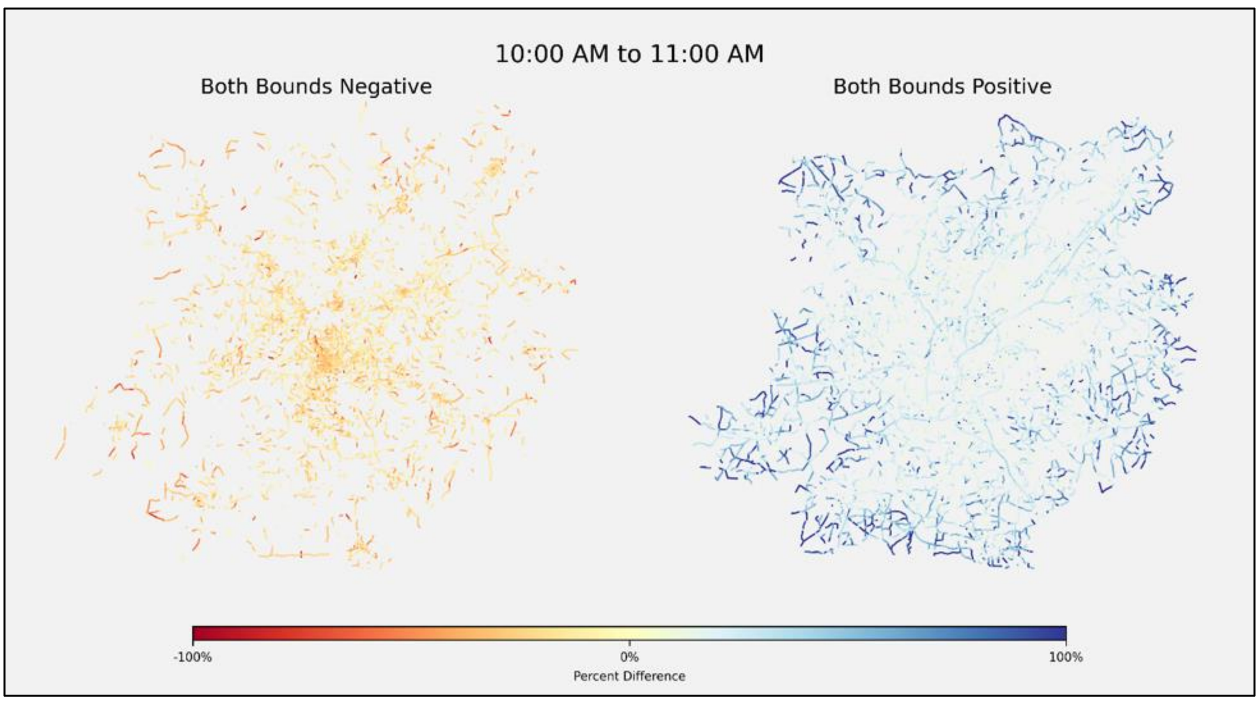

Figure 77. Links with both CI bounds negative or positive, 10 AM to 11 AM.

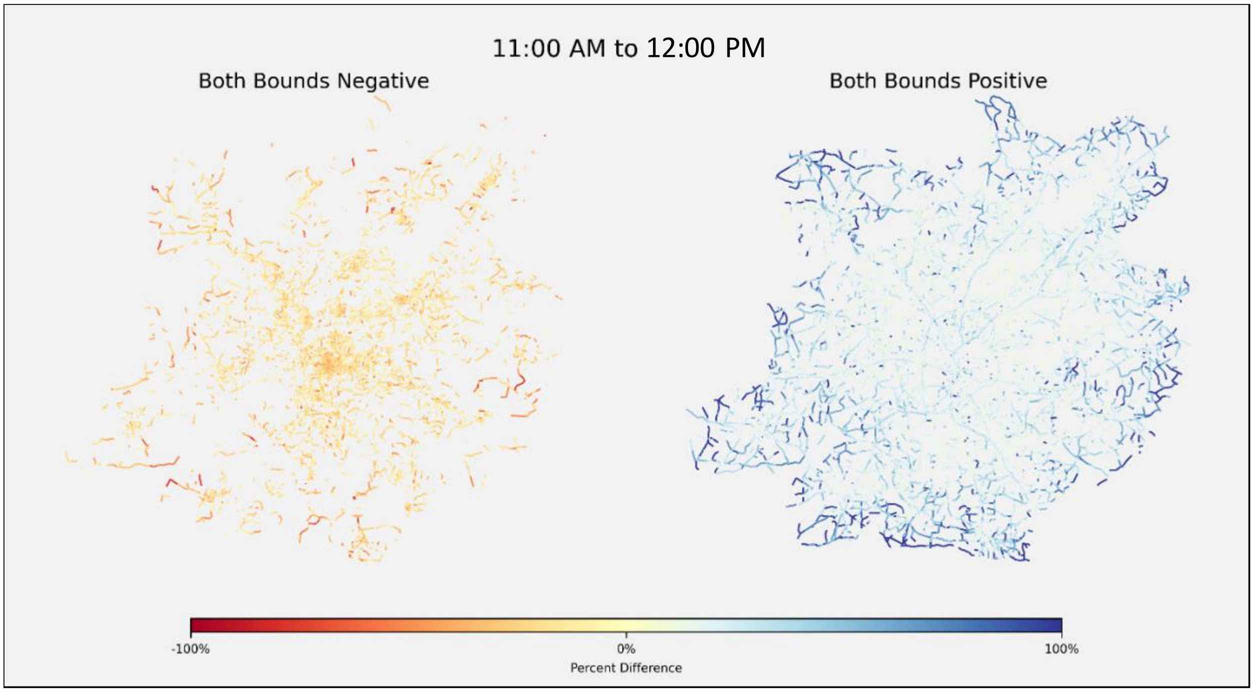

Figure 78. Links with both CI bounds negative or positive, 11 AM to 12 PM.

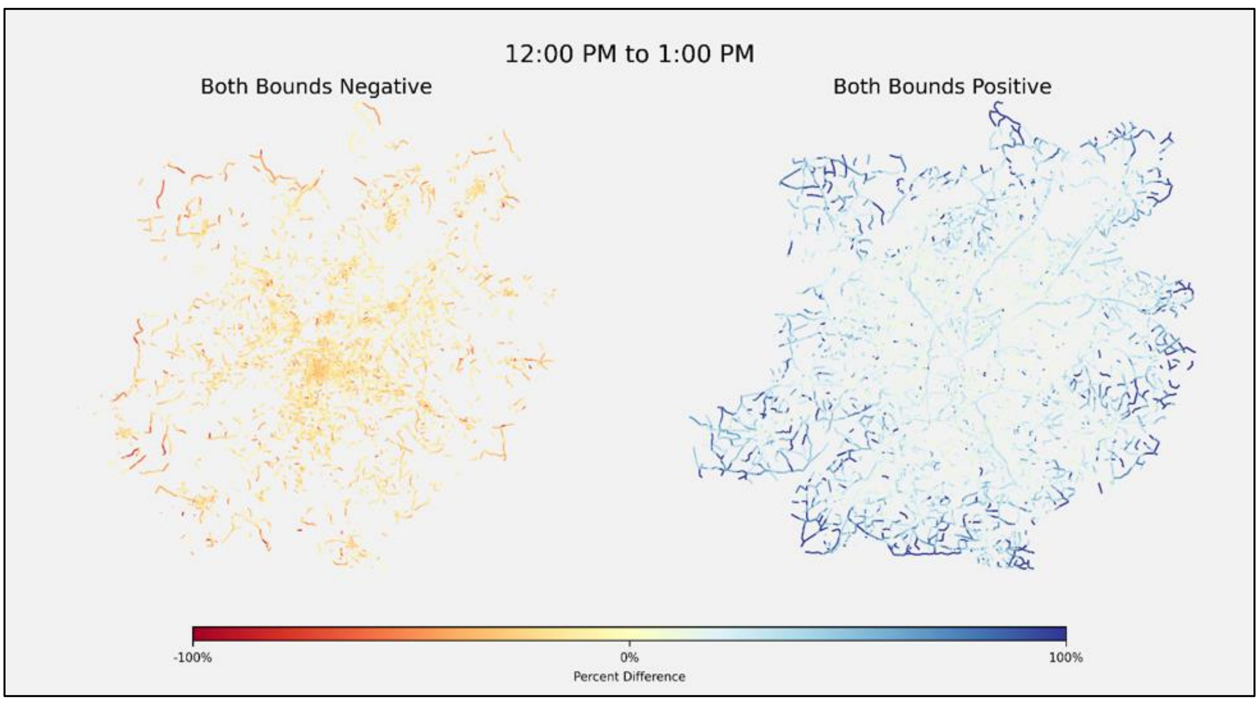

Figure 79. Links with both CI bounds negative or positive, 12 PM to 1 PM.

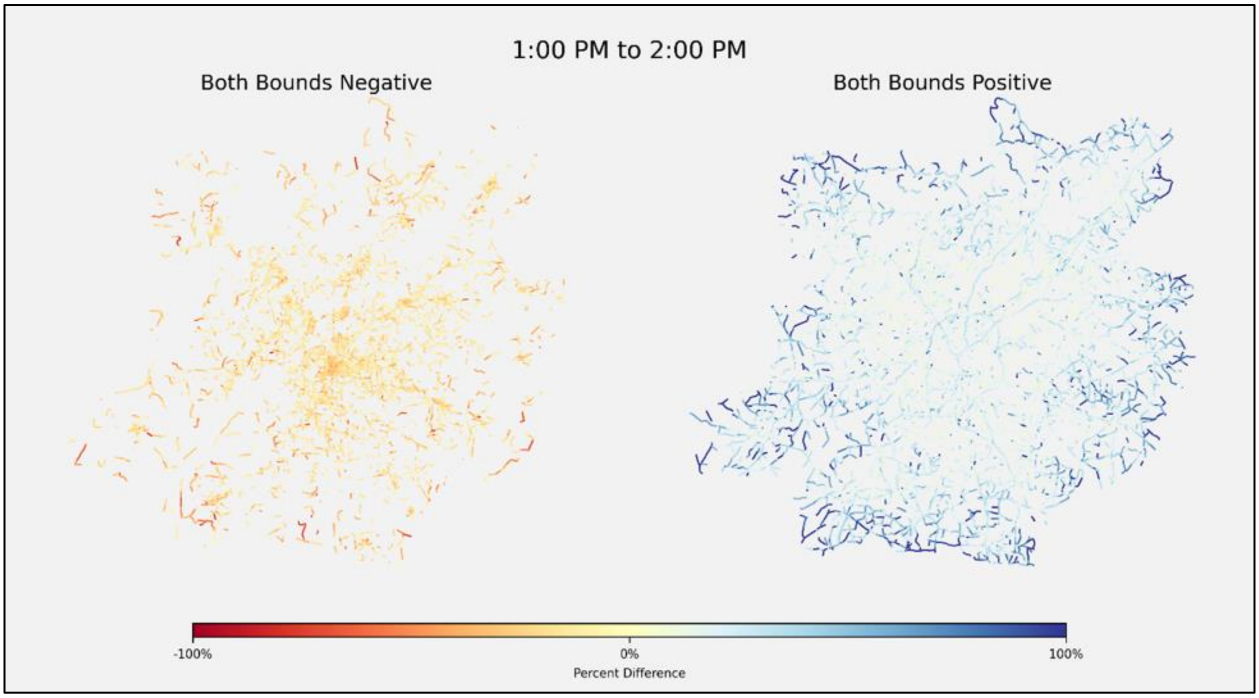

Figure 80. Links with both CI bounds negative or positive, 1 PM to 2 PM.

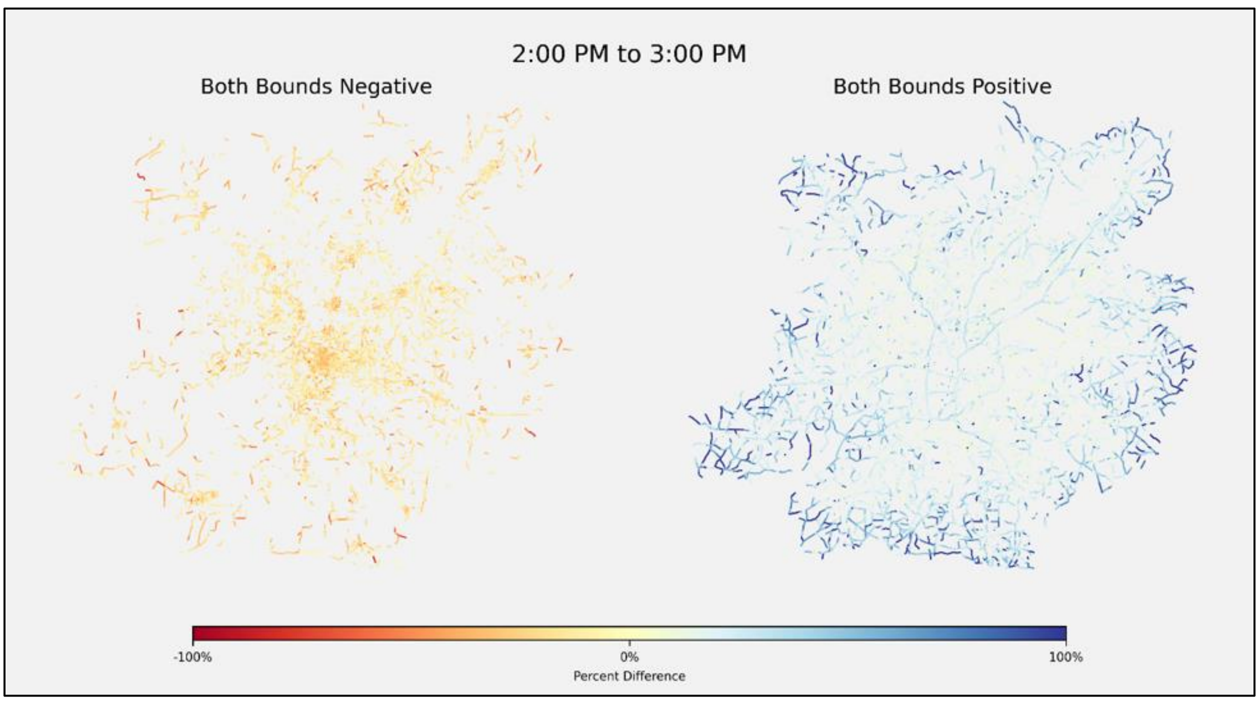

Figure 81. Links with both CI bounds negative or positive, 2 PM to 3 PM.

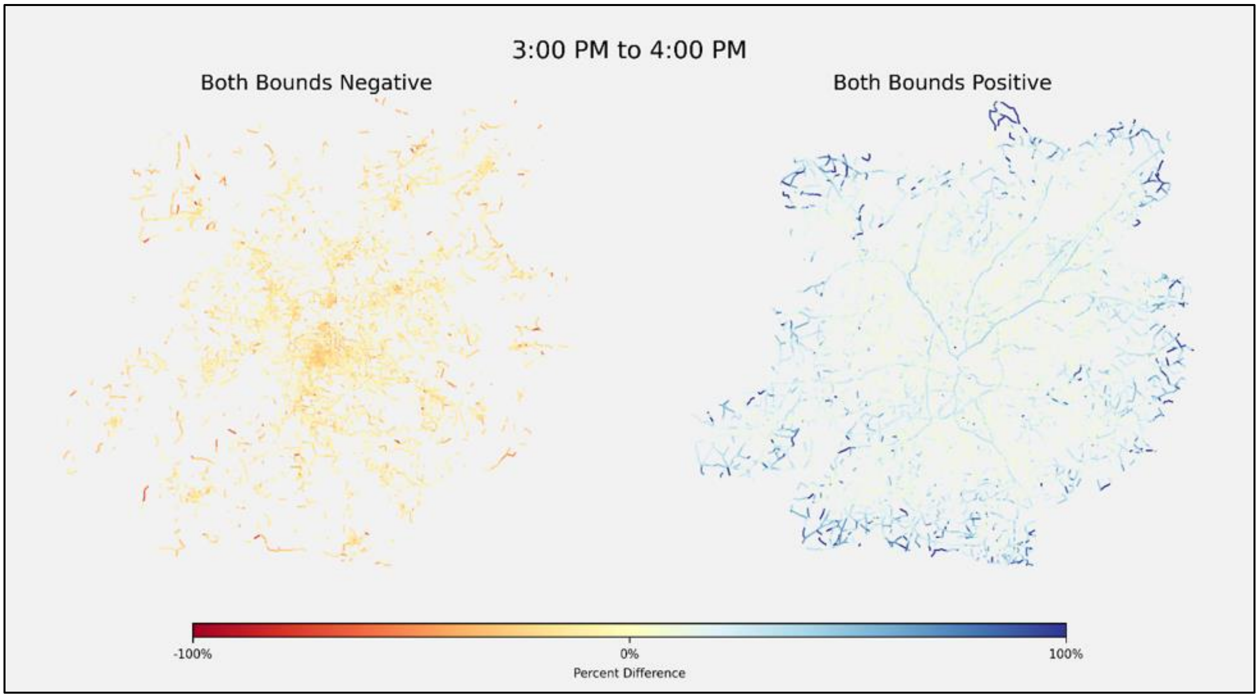

Figure 82. Links with both CI bounds negative or positive, 3 PM to 4 PM.

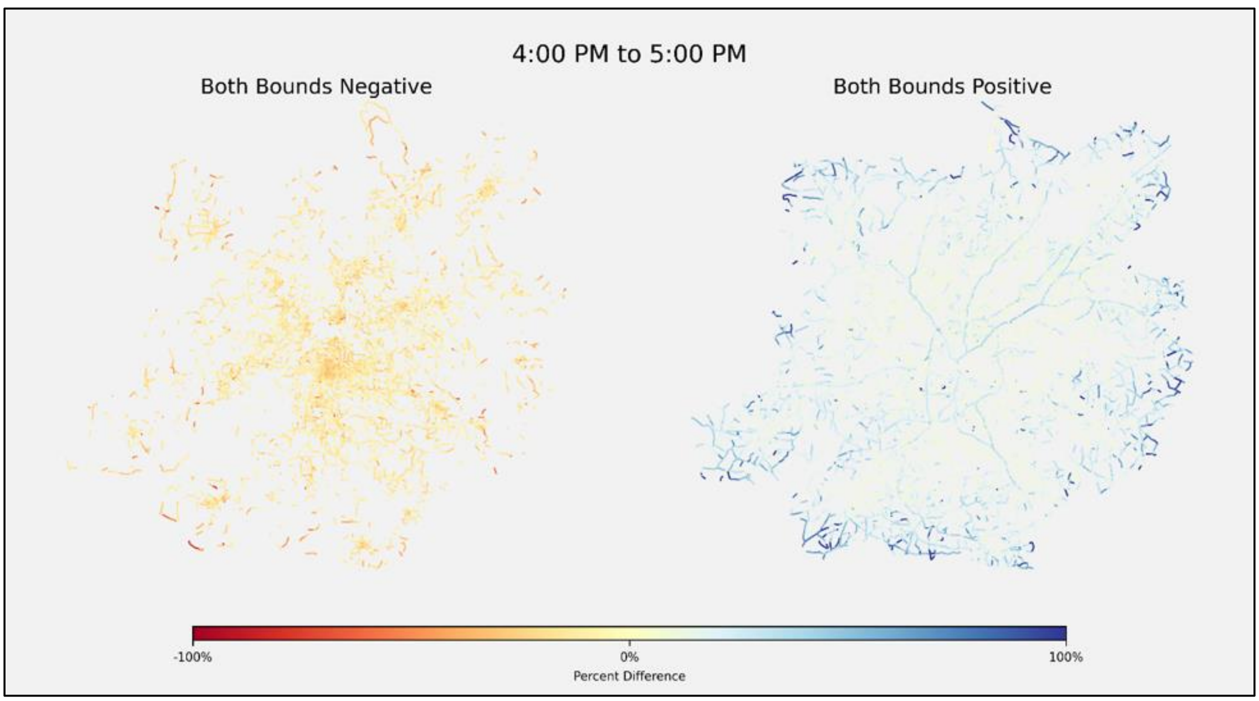

Figure 83. Links with both CI bounds negative or positive, 4 PM to 5 PM.

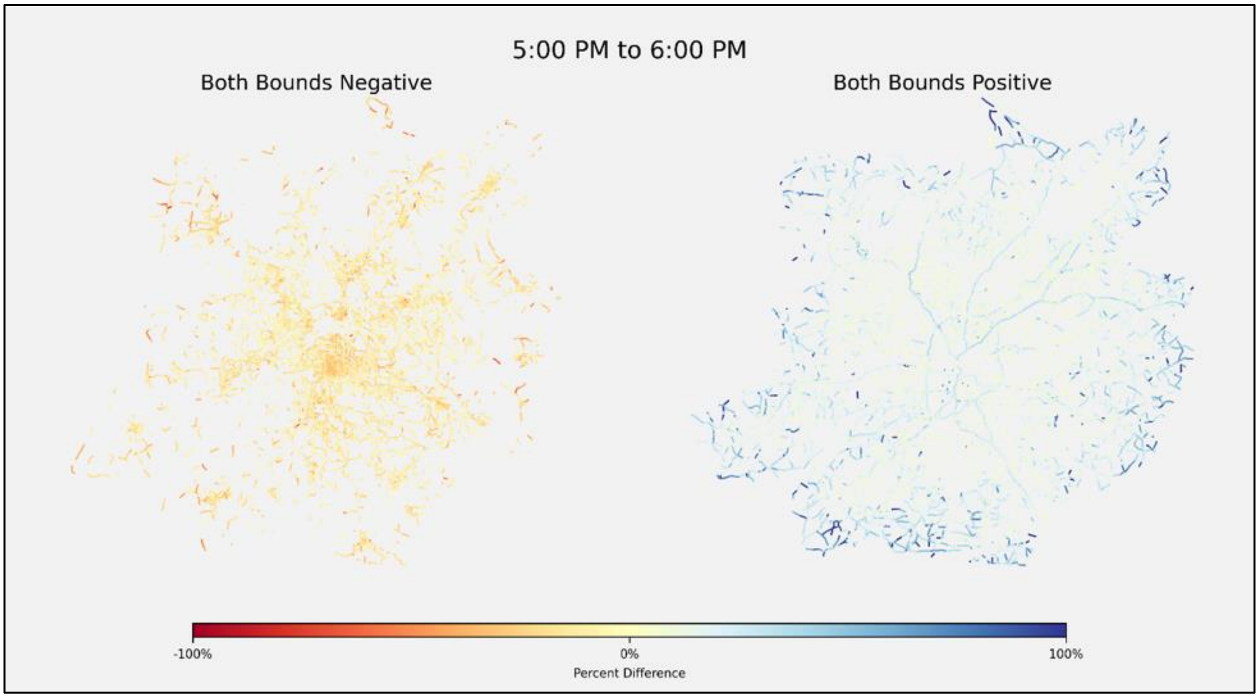

Figure 84. Links with both CI bounds negative or positive, 5 PM to 6 PM.

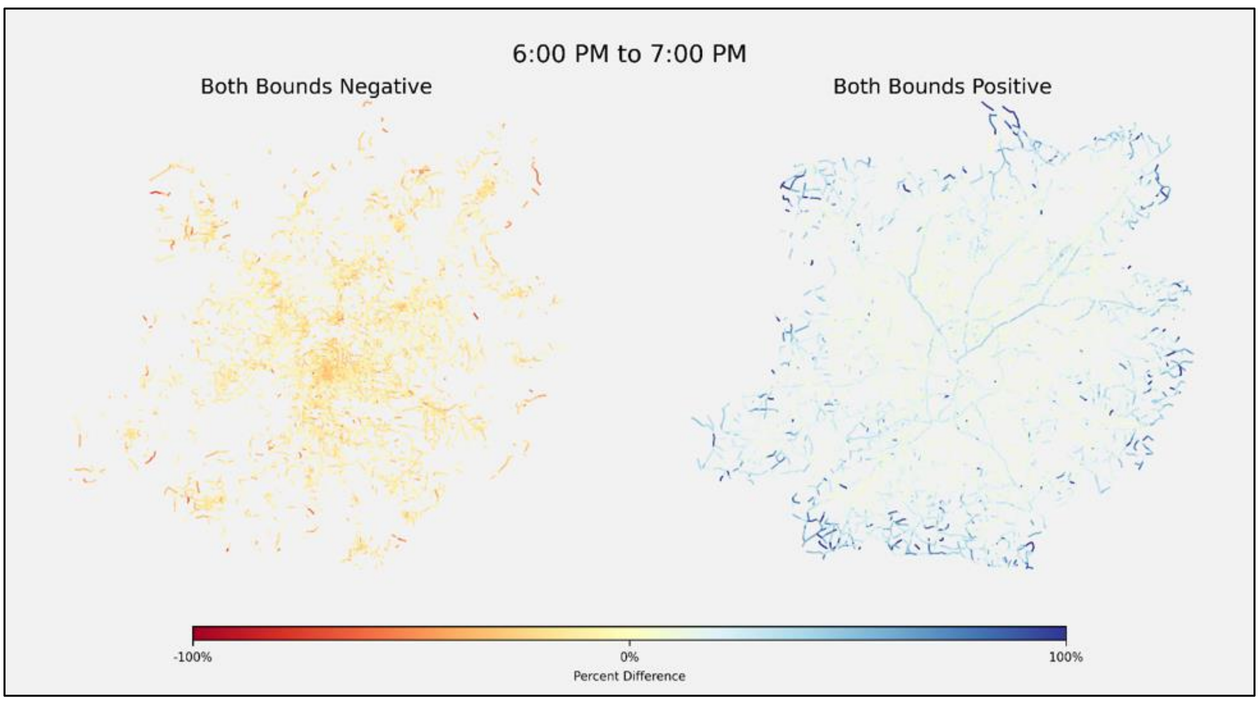

Figure 85. Links with both CI bounds negative or positive, 6 PM to 7 PM.

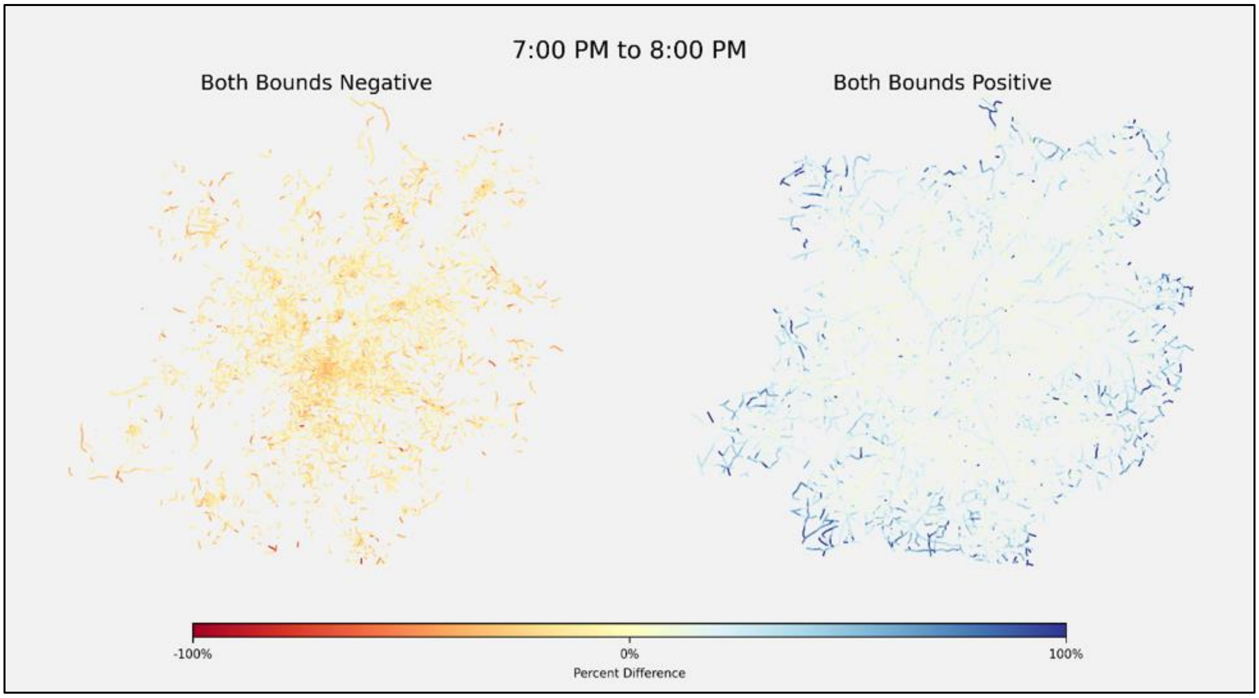

Figure 86. Links with both CI bounds negative or positive, 7 PM to 8 PM.

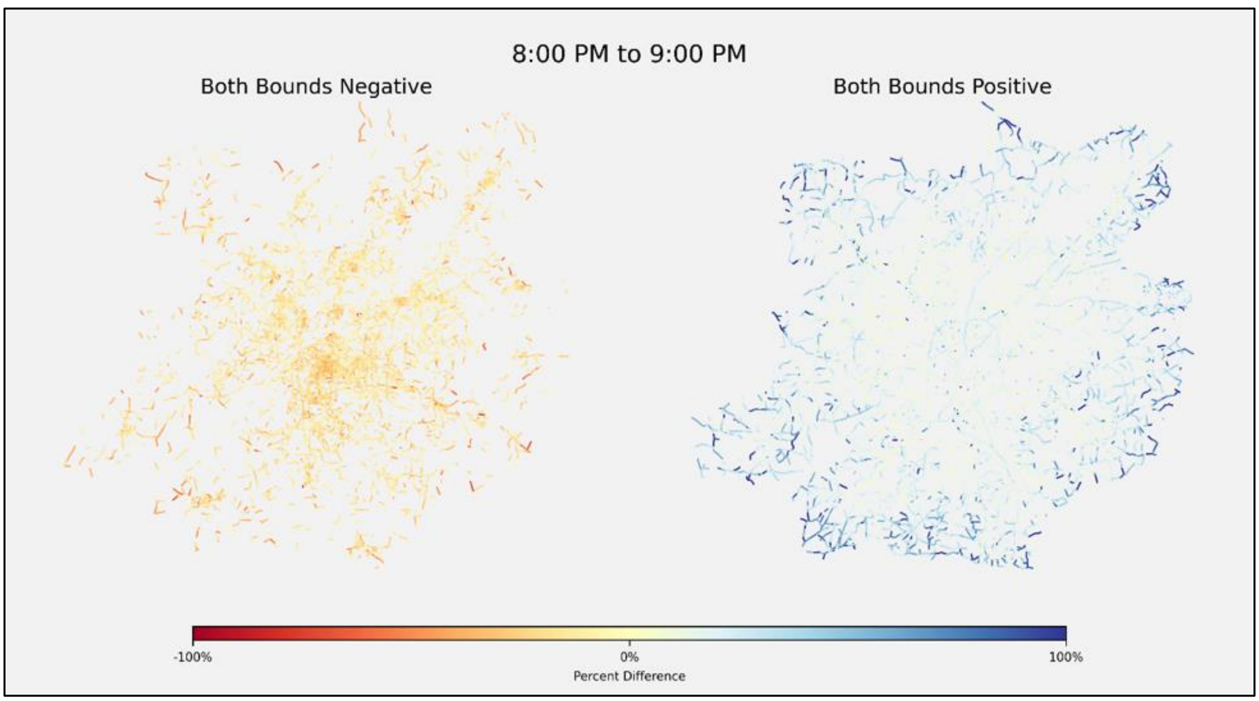

Figure 87. Links with both CI bounds negative or positive, 8 PM to 9 PM.

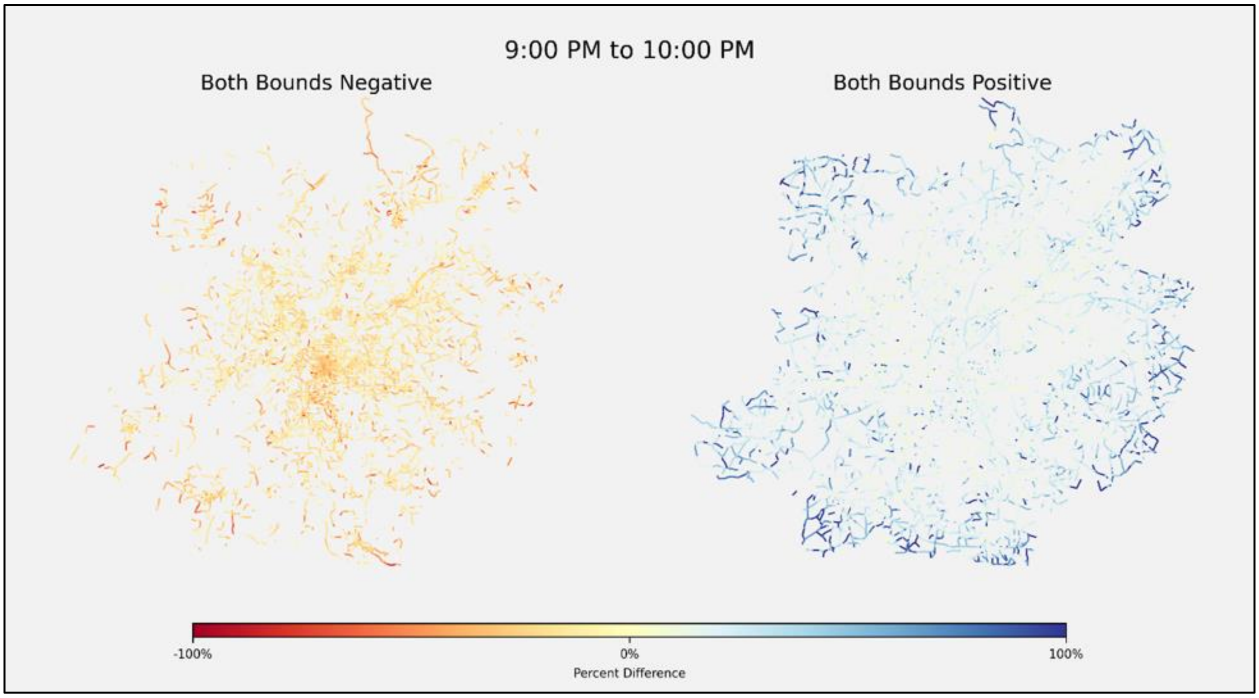

Figure 88. Links with both CI bounds negative or positive, 9 PM to 10 PM.

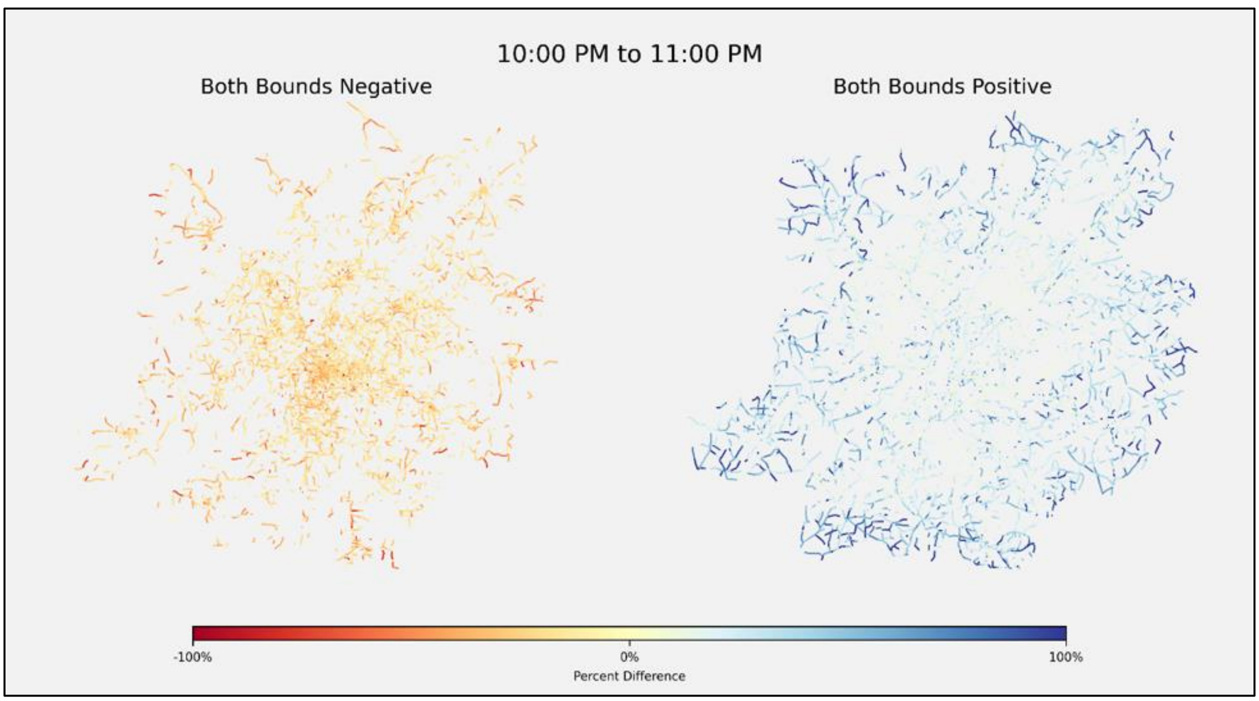

Figure 89. Links with both CI bounds negative or positive, 10 PM to 11 PM.

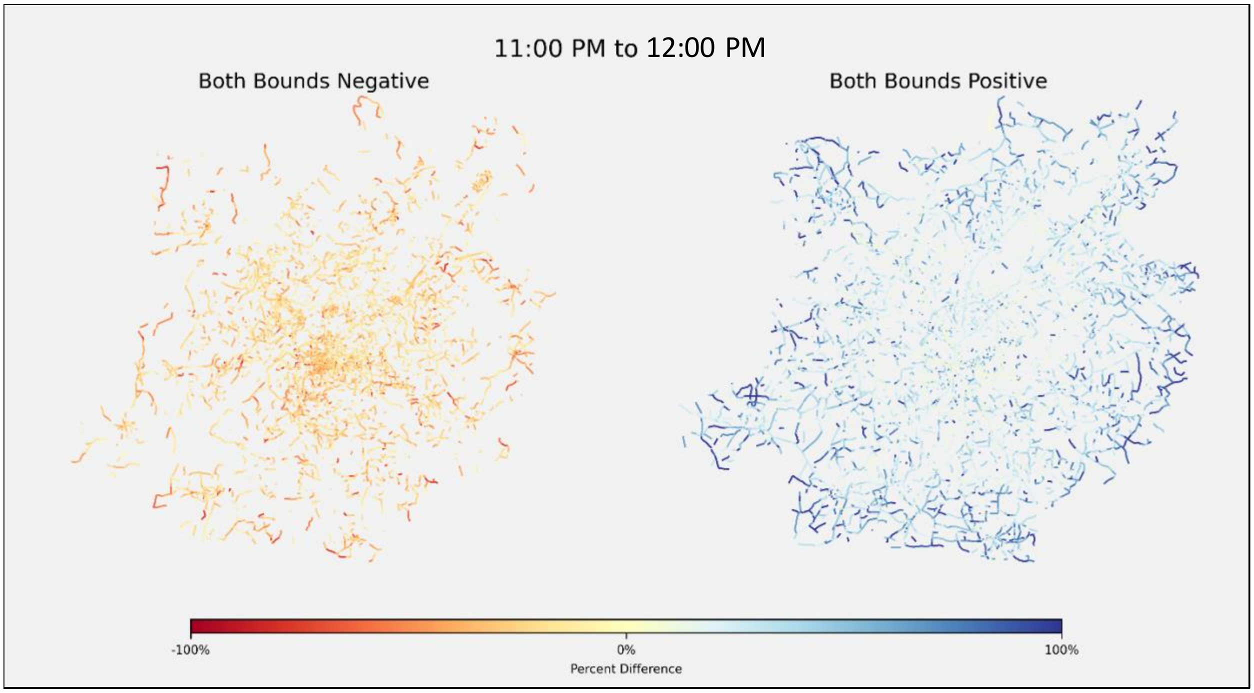

Figure 90. Links with both CI bounds negative or positive, 11 PM to 12 AM.