

ABSTRACT

Title of Dissertation: DATA-DRIVEN DECISION-MAKING FOR
SOCIETAL SYSTEMS: AUDITING
MOBILITY, CONTEXTUAL DATA
AND MODELS FOR HIGH-STAKES APPLICATIONS

Saad Mohammad Abrar
Doctor of Philosophy, 2026

Dissertation Directed by: Professor Vanessa Frias-Martinez
College of Information Studies and UMIACS

High-stakes decisions in domains such as public health and economic policy are increasingly mediated by behavioral data and predictive models, yet the phenomena they target often evolve faster than conventional measurement systems can track. This has driven growing reliance on human mobility and contextual data as real-time proxies and inputs to predictive models, a dependence made especially visible during the COVID-19 pandemic. Yet these data, and the models built on top of them, are rarely subjected to the kinds of systematic evaluation needed to establish whether they can be trusted, for whom their benefits hold, and under what conditions their measurements and predictions fail to hold. This dissertation argues that mobility, contextual data, and the associated downstream models must be audited along three dimensions before being trusted in high-stakes settings: *utility*, *fairness*, and *validity*. It further argues that the same scrutiny should be applied to emerging *alternatives* for direct behavioral traces, such as large language models used to infer or simulate human mobility when observational data are sparse or unavailable.

The first study examines the *utility* of mobility data for COVID-19 case prediction, asking not whether it improves performance on average, but under what conditions across counties, datasets, and modeling choices its value can be reliably demonstrated. The second study conducts a *fairness* audit of the U.S. COVID-19 Forecast Hub, examining whether the case prediction models used to inform CDC communications performed equitably across racial, ethnic, and urbanization dimensions. The third study evaluates the *validity* of foot traffic as a proxy for economic activity, assessing whether visit-spend relationships are stable across business categories, time periods, and neighborhood demographic contexts. The fourth study evaluates whether large language models can serve as an *alternative* for behavioral traces by asking whether they encode empirically meaningful knowledge about how neighborhood context relates to movement, or merely produce plausible but unverified outputs.

Together, these four studies demonstrate that the trustworthiness of mobility and contextual data is not a fixed property but a conditional one, dependent on the task, the population, and the context in which these data are used. The dissertation contributes a unifying auditing framework and a set of empirical evaluations that together make the case for treating utility, fairness, and validity as fundamental criteria in any data-driven system where predictions shape real-world decisions.

DATA-DRIVEN DECISION-MAKING FOR SOCIETAL SYSTEMS: AUDITING
MOBILITY, CONTEXTUAL DATA AND MODELS FOR HIGH-STAKES
APPLICATIONS

by

Saad Mohammad Abrar

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:

Professor Vanessa Frias-Martinez, Chair/Advisor
Professor Kunpeng Zhang, Dean's representative
Professor Hal Daume
Professor Ashok Agrawala
Professor Louiqa Raschid
Professor Zhicheng Liu

© Copyright by
Saad Mohammad Abrar
2026

Dedication

To my family.

Acknowledgments

I am deeply grateful for all the support I have received throughout my PhD.

First and foremost, I would like to thank my advisor, Dr. Vanessa Frias-Martinez. Since I joined her lab after moving to UMD in 2020, she has been a constant source of support, guidance, and encouragement. I have learned so much from her over the years, far more than I can put into words. She was always patient with me, always approachable, and always willing to help whenever I was stuck. I am especially grateful for how involved and supportive she was throughout my PhD, not only in my research but also in helping me grow professionally. I truly feel lucky to have had such an incredible advisor.

I would also like to thank my labmates, especially Naman and Daniel. I worked closely with both of them for most of my PhD, and I learned so much from our collaborations, conversations, and shared struggles through different projects. I am incredibly grateful to both of them. I also want to thank my other labmates, Jiahui, Nikki, and Dea, for making the lab such a warm and supportive place to be. I was also fortunate to work with several undergraduate students in the lab, and I would especially like to mention Eesha and Arnav for helping me with my final project.

I am very thankful to my committee members: Dr. Hal Daumé, Dr. Louiqa Raschid, Dr. Ashok Agrawala, Dr. Kunpeng Zhang, and Dr. Zhicheng Liu. I especially want to thank Hal, Louiqa, and Ashok for helping me think more deeply about the bigger picture of my research during my proposal stage. Their questions and feedback pushed me to think beyond the immediate technical contributions of my work. I am also especially grateful to Louiqa for her thoughtful feedback and encouragement throughout my PhD.

I also want to thank my friends in College Park, who were there for me through all the ups and downs of this journey. PhD life can be isolating at times, and their support, friendship, and presence

meant a lot to me.

I want to thank my family for all their love and support, especially my mother. I would not be where I am today without her sacrifices, strength, and endless care. So much of what I have been able to do is because of her. I am also grateful to my family as a whole for always standing by me.

And finally, Zinat, I do not think I would have made it this far without you. You were there for me in my hardest moments, and also in my happiest ones. Thank you for being my source of strength, comfort, and joy throughout this journey.

Table of Contents

Dedication	ii
Acknowledgements	iii
Table of Contents	v
List of Tables	viii
List of Figures	xii
Chapter 1: Introduction	1
1.1 Motivation & Background	1
1.2 Research Objectives and Questions	6
1.3 Dissertation Organization	9
Chapter 2: Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models	11
2.1 Introduction	11
2.2 Related Work	14
2.2.1 Human mobility data and COVID-19 case predictions	14
2.2.2 COVID-19 prediction models	16
2.3 Datasets	17
2.4 Methodology	19
2.4.1 Prediction Models	20
2.4.2 Training-Testing Approaches	22
2.4.3 Model performance	24
2.4.4 Adjusting Predictive Models	24
2.4.5 Evaluation approach	25
2.5 Results and Analysis	26
2.5.1 Does mobility data help, and by how much?	26
2.5.2 What types of counties benefit from mobility data? A mobility data bias analysis	31
2.5.3 Do correlation improvements change across training approaches?	35
2.5.4 Do correlation improvements change across mobility datasets?	38
2.5.5 Do correlation improvements change across predictive models?	40
2.6 Discussion	42
Chapter 3: Auditing the Fairness of COVID-19 Forecast Hub's Case Prediction Models	45
3.1 Introduction	45
3.1.1 The US COVID-19 Forecast Hub: Development and Context	45
3.1.2 Fairness in COVID-19 Prediction Models: A Critical Gap	46
3.2 Materials and Methods	50
3.2.1 Data Sources and Variables	50

3.2.2	Analytical Approach	55
3.2.3	Model Selection and Evaluation	61
3.3	Results	62
3.3.1	Fairness of COVID-19 case predictions across race and ethnicity (GLM-1) . . .	62
3.3.2	Fairness of COVID-19 case predictions across urbanization levels (GLM-2) . .	64
3.3.3	Fairness of COVID-19 case predictions across race, ethnicity and model-data characteristics	66
3.3.4	Fairness of COVID-19 case predictions across urbanicity and model-data characteristics	74
3.4	Dashboard	80
3.5	Discussion	86
Chapter 4:	Assessing Ethnoracial Disparities in COVID-19’s Impact on Small Businesses in US Metro Areas Using Mobility Data	90
4.1	Introduction	90
4.2	Data	92
4.2.1	Mobility Data	92
4.2.2	ACS Data	93
4.2.3	Why CBGs as a proxy measure of localized impact?	95
4.3	Methodological Approach	96
4.4	Results	98
Chapter 5:	Do Footfalls Reliably Mirror Economic Activity?	102
5.1	Introduction	102
5.2	Related Work	105
5.2.1	Mobility Data as a proxy for downstream tasks	105
5.2.2	Evaluation of mobility data as a proxy	106
5.3	Material and Methods	107
5.3.1	Data Sources and Preprocessing	109
5.3.2	RQ1: Proxy Validity and Temporal Stability Assessment	112
5.3.3	RQ2: Demographic Moderation of Elasticities	116
5.4	Analysis of Results	118
5.4.1	RQ1: Proxy Validity and Temporal Stability Assessment	118
5.4.2	RQ2: Do neighborhood-level demographic data moderate the visit-spend elasticities?	126
5.5	Discussion	131
Chapter 6:	Evaluating LLM Priors About Neighborhood-Level Mobility	135
6.1	Introduction	135
6.2	Related Work	138
6.2.1	LLMs for Mobility Generation, Prediction, and Urban Tasks	138
6.2.2	Evaluating Geospatial and Mobility Knowledge in LLMs	140
6.3	Method	141
6.3.1	Problem Formulation	141
6.3.2	Mobility Outcome Construction	143

6.3.3	Mobility Outcome Prediction Framework	148
6.3.4	Directional Alignment Framework	151
6.4	Experiments	153
6.4.1	Experimental Setting	153
6.4.2	RQ1: Predictive Performance	157
6.4.3	RQ2: Directional Alignment Results	160
Chapter 7:	Conclusion	164
7.1	Conclusion and Summary of Contribution	164
7.1.1	Study 1: Evaluating the Impact of Mobility Data on COVID-19 Case Prediction	164
7.1.2	Study 2: Fairness Audit of the U.S. COVID-19 Forecast Hub	165
7.1.3	Study 3: Validity of Foot Traffic as an Economic Proxy	166
7.1.4	Study 4: LLMs as Priors for Urban Mobility	166
7.2	Implications	167
Appendix A:	Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models	169
A.1	Appendix	169
Appendix B:	Auditing the Fairness of COVID-19 Forecast Hub’s Case Prediction Models	184
B.1	Supporting Information for Data	184
B.1.1	Exclusion of some counties and certain minority races	184
B.1.2	Description of the model type classification	187
B.1.3	Description of Phases	189
B.2	Supporting Information for Preprocessing & Model selection	191
B.2.1	Data Preprocessing	191
B.2.2	Multicollinearity Analysis and Variable Selection	193
B.2.3	Model Diagnostics	195
Appendix C:	Do Footfalls Reliably Mirror Economic Activity?	226
C.1	Visit Attribution using Cuebiq and Safegraph Places	226
C.2	Representation check of SafeGraph Spend	227
Appendix D:	Evaluating LLM Priors About Neighborhood-Level Mobility	237
D.1	System Prompts	237
D.1.1	Appendix: Prompt Used for Zero-Shot Mobility Classification	237
D.1.2	Prompt Used for Directional Alignment Analysis	239

List of Tables

1.1	Broader gaps motivating this dissertation under the umbrella of auditing.	6
2.1	Number of counties considered for each mobility dataset.	19
2.2	Mann-Whitney U test between the STW and LTW distributions of the percentage of counties that benefit from adding mobility data across the ten datasets and two models for each training approach.	35
2.3	Mann-Whitney U test per lookahead between the STW and LTW distributions of the correlation improvements across the ten datasets and two models for each training approach.	36
2.4	Improvement statistics per mobility dataset across training approaches and predictive models. The percentage of counties column represents the average percentage of counties whose COVID-19 case prediction models benefit from adding mobility data; while the correlation improvement (ci) column quantifies the average median <i>ci</i> improvement over the non-mobility baseline.	38
2.5	Mann-Whitney U test between the ARIMAX and linear distributions of the percentage of counties that benefit from adding mobility data across the ten datasets and two training approaches for each predictive model.	40
2.6	Mann-Whitney U test per lookahead between the ARIMAX and linear distributions of the correlation improvements across the ten datasets and two training approaches for each model type.	41
3.1	County-Level Demographic Distribution Statistics	52
3.2	Summary of Generalized Linear Models and Their Specifications	60
3.3	Summary of Regression Results: GLM-1	63
3.4	Summary of Regression Results: GLM-2	65
3.5	GLM-1a: Race $\times$ Lookahead Effects Relative to White Reference Group	67
3.6	GLM-1b: Race $\times$ Phase Effects Relative to White Reference Group	69
3.7	GLM-1c: Race $\times$ Model Type Effects Relative to White Reference Group	71
3.8	GLM-1d: Race $\times$ Mobility Data Inclusion Effects Relative to White	73
3.9	GLM-2a: Urbanicity $\times$ Lookahead Effects Relative to LM Reference Group	75
3.10	GLM-2b: Urbanicity $\times$ Phase Effects Relative to LM Areas	76
3.11	GLM-2c: Urbanicity $\times$ Model Type Effects Relative to LM Areas	77
3.12	GLM-2d: Urbanicity $\times$ Mobility Usage Effects Relative to LM	79
4.1	Average percentage change and the standard deviation (std) in visitation patterns per ethnorace over all the time period under study as well as for each phase.	101
4.2	Two-sample Kolmogorov-Smirnov tests between each pair of ethnoracial average percentage change time series. For each pair of distributions we report the ks statistic and its significance (p-value).	101

5.1	Sample Size Distribution by MSA and Business Category after quality filters. Table shows POI-week observations with number of unique POIs in parentheses. Sample covers 2019-2024 period across <i>six</i> major metropolitan areas and <i>twelve</i> business categories grouped by economic function.	112
6.1	Summary of CBG coverage and device counts by core-based statistical area (CBSA).	154
6.2	Coverage statistics across the study CBSAs. The table reports the number of CBGs covered, the total number of devices, and the distribution of devices per CBG.	154
A.1	Average p_{corr} value for each regularized method. The best value for each window and each lookahead is bolded.	169
A.2	Regression coefficient analysis for Arimax LTW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '****' 0.001, '***' 0.01, '**' 0.05	170
A.3	Regression coefficient analysis for Arimax STW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '****' 0.001, '***' 0.01, '**' 0.05	171
A.4	Regression coefficient analysis for Ridge Regression LTW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '****' 0.001, '***' 0.01, '**' 0.05	172
A.5	Regression coefficient analysis for ElasticNet Regression STW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '****' 0.001, '***' 0.01, '**' 0.05	173
A.6	Non-parametric Friedman test analysis with distributions containing the percentage of counties where mobility data improved COVID-19 case prediction models without mobility data across datasets, lookaheads, training approaches and models.	173
A.7	Non-parametric Friedman test analysis with distributions containing the median correlation improvement values of counties where mobility data improved COVID-19 case prediction models without mobility data across datasets, lookaheads, training approaches and models. Each test is run with one or a few mobility datasets excluded.	174
A.8	Summary of the combinations of p, d, q values identified for the ARIMAX models using grid search and AIC. The table represents the number of counties whose p, d, q , values are the ones listed on the left hand side column.	183
B.1	List of the 36 COVID-19 forecast models analyzed in this paper, together with their model type and whether they used mobility data or not, including the source of that mobility data. Information about model type and the use of mobility data was manually extracted from the paper(s) for each model by the authors of this paper.	186
B.2	Distribution of Pinball Loss Norm Values	193
B.3	Iterative removal of Health Outcome (HO) variables based on adjusted GVIF < 2	193

B.4	GVIF Assessment for GLM-2 (Urbanization Main Effects): Initial Assessment satisfies multicollinearity assumption	194
B.5	State Fixed Effects for GLM-1 (Reference State: NY)	201
B.6	State Fixed Effects for GLM-2 (Reference State: NY)	202
B.7	Main Effects for GLM-1a (Race × Lookahead Interactions)	205
B.8	State Fixed Effects for GLM-1a (Reference State: NY)	206
B.9	Main Effects for GLM-1b (Race × Phase Interactions)	207
B.10	State Fixed Effects for GLM-1b (Reference State: NY)	208
B.11	Main Effects for GLM-1c (Race × Model Type Interactions)	209
B.12	State Fixed Effects for GLM-1c (Reference State: NY)	210
B.13	Main Effects for GLM-1d (Race × Mobility Interactions)	211
B.14	State Fixed Effects for GLM-1d (Reference State: NY)	212
B.15	Main Effects for GLM-2a (Urbanicity × Lookahead Interactions)	213
B.16	State Fixed Effects for GLM-2a (Reference State: NY)	214
B.17	Main Effects for GLM-2b (Urbanicity × Phase Interactions)	215
B.18	State Fixed Effects for GLM-2b (Reference State: NY)	216
B.19	Main Effects for GLM-2c (Urbanicity × Model Type Interactions)	217
B.20	State Fixed Effects for GLM-2c (Reference State: NY)	218
B.21	Main Effects for GLM-2d (Urbanicity × Mobility Interactions)	219
B.22	State Fixed Effects for GLM-2d (Reference State: NY)	220
C.1	RQ2: POI and neighborhood socio-demographics by economic group, category, and MSA. Sample sizes are smaller than in RQ1 because some CBGs lack socio-demographic data (e.g., around airports). Cells shaded red denote strata with insufficient variation for estimation ($SD < 5$); these are excluded from reported regressions. Summary statistics are shown as <i>mean (sd)</i> . Abbreviations: n_cbgs = number of CBGs represented; n_pois = number of POIs; n_obs = number of observations per MSA × POI category.	231
C.2	Visit-Spend Elasticity Estimates by Economic Category, MSA, and Year (2019-2024): Table shows visit-spend elasticity estimates by economic category, MSA, and business category for 2019-2024. Confidence intervals shown in brackets below elasticity estimates. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Economic categories are color-coded: Essential Goods , Routine Services , Social & Experiential and Discretionary Retail . R^2 (Within) and R^2 (Overall) from panel regression models with POI and year fixed effects.	232
C.3	Category-Specific Elasticity Performance Metrics by Metropolitan Area. Table presents elasticity performance metrics for visit-spend relationships across business categories and metropolitan areas (2019–2024). <i>Mean Resp.</i> = average responsiveness (elasticity) across years; <i>Stability CV</i> = coefficient of variation measuring temporal consistency, Score = 1 - CV reported in main manuscript; <i>Reliability</i> = statistical significance rate; <i>COVID Impact</i> = elasticity disruption during 2020–2021 (negative values shaded); <i>Geo Consist.</i> = geographic consistency across MSAs. “—” indicates metrics could not be derived due to not meeting the reliability threshold.	233

C.4	% Changes in elasticity estimates in 2020 compared to 2019 across MSAs. The true tag in increased column indicates whether the elasticity estimate increased in 2020 compared to 2019. The Sig_2019 and Sig_2020 indicates whether the estimates were significant for 2019 and 2020 respectively.	234
C.5	Branded vs. non-branded restaurant elasticities by MSA, 2019–2024. Annual within-POI spend–visit elasticities estimated separately for branded and non-branded segments using the panel specification in Equation (1) in main manuscript with POI and year fixed effects. Columns report sample size (N_{Obs} , N_{poi}), within- and overall- R^2 , and year-specific elasticities with 95% confidence intervals in brackets. Asterisks denote conventional significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). . . .	235
C.6	Demographic Moderation of Visit-Spend Elasticities by Economic Category. Table shows demographic moderation coefficients from category-specific panel regressions. Each coefficient represents the interaction between log(visits) and the respective demographic variable. Income is measured as median household income percentile, Education as percentage with bachelor’s degrees, and racial/ethnic variables as population percentages. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. All regressions include POI fixed effects and year fixed effects with Driscoll-Kraay standard errors. Statistically significant estimates are shaded green for amplification and red for attenuation.	236
D.1	Parse success rates across evaluated LLMs. Models with parse success rate below 95% were excluded from downstream analysis.	241

List of Figures

2.1	Training-Testing approaches for LTW and STW. The LTW approach for Linear Regression is not represented because it is the standard 70% – 30% approach.	23
2.2	Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the LTW approach.	26
2.3	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with LTW approach using (a) Apple mobility data and (b) SafeGraph Inflow mobility data.	27
2.4	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ridge regression with LTW approach using (a,b) Apple Mobility data and (c,d) SafeGraph Inflow Mobility data.	28
2.5	Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the STW approach.	33
3.1	Distribution of the sensitive attributes (Race/Ethnicity and Urbanization level) across the 3,067 counties in US considered for this study.	51
3.2	Prediction fairness will be evaluated across types of models, training datasets (mobility), lookaheads and phases.	53
3.3	Forecast Hub Fairness Dashboard showing the Average Error Ratio (AER) distribution across different COVID-19 prediction models, organized by model type. Within each model type, teams are sorted in ascending order based on their median AER values. Since the user has selected “Only Race” as the variable of interest (see bottom left box) and “Hispanic” as the protected variable (see bottom center box), the AER values compare prediction errors between Hispanic and White counties, where values above 1.0 indicate higher prediction errors for Hispanic counties. Box plots show the distribution of AER values across all predictions, with the center line representing the median, boxes showing the interquartile range, and whiskers extending to the minimum and maximum values.	82
3.4	Model fairness card displaying key performance metrics including model information, prediction error differences between protected and unprotected groups, AER values, and coverage statistics.	84
4.1	Urban Census Block Groups representing all the large metropolitan areas used in our analysis.	94
4.2	Count of CBGs associated to each majority race.	94
4.3	Racial composition at selected urban locations.	95
4.4	Heatmap matrices for percentage changes in visit patterns	99
4.5	Average percentage change time series across all CBGs for each ethnicity or race. The vertical dashed lines separate the pre-pandemic phase from phases 1, 2 and 3.	100

5.1	Analytical framework for estimating visit–spend elasticities and moderation effects. For each MSA and business category, we construct a weekly POI-level panel with $\log(\text{visits})$ and $\log(\text{spend})$. RQ1 estimates yearly visit–spend elasticities using POI-week observations with POI fixed effects, yearly fixed effects, and Driscoll–Kraay standard errors to account for cross-sectional(spatial) & temporal dependence, as defined in eq. (5.1). RQ2 extends this specification by interacting visit–spend relationships with neighborhood socio-demographic characteristics to quantify moderation effects; the full specification is provided in eq. (5.8).	108
5.2	Temporal Evolution of Visit–Spend Elasticities Across Business Categories and Metropolitan Areas (2019–2024). Year-specific elasticity estimates from Eq. 5.1 for twelve business categories across six MSAs. For 2019, elasticity is given by the baseline coefficient β_1 ; for 2020–2024, elasticities are $\beta_1 + \gamma_\tau$, capturing deviations from the 2019 baseline. Each panel is a business category, with background colors indicating economic type (Essential Goods , Routine Services , Social & Experiential , Discretionary Retail). Colored lines denote MSAs; <i>black outlined circles denote non-significant estimates</i> ($p > 0.05$).	119
5.3	Comprehensive Assessment of Proxy Validity Metrics Across Business Categories and Metropolitan Areas. Each row is a business category (grouped as Essential Goods , Routine Services , Social & Experiential , Discretionary Retail); columns are MSAs (ATL, CHI, HOU, MIA, PHI, WAS). The Mean Responsiveness panel shows average elasticities (2019–2024; darker red = stronger relationships). The Temporal Stability panel shows the coefficient of variation across years (darker blue = higher stability). The Significance Rate panel shows the share of years with statistically significant elasticities (darker blue = more consistently significant). The Geo Consistency metric (single value per category, reported in Table C.3 captures cross-MSA agreement. White cells in the Elasticity and Stability panels indicate that significance rates did not meet the inclusion threshold.	120
5.4	ATL–Restaurants: co-movement and within-POI elasticities. Panel A: Weekly percent change from the January 2020 average in visits (blue, dashed) and card spending (red, solid) for restaurants in ATL; the shaded band marks initial lockdown weeks. Panel B: Within-POI relationship between residual $\log(\text{spend})$ and residual $\log(\text{visits})$ after removing POI and year fixed effects. Lines are OLS fits by year. Estimated elasticities: $\beta_{2019} = 0.125$ and $\beta_{2020} = 0.271$, indicating steeper coupling in 2020. Formal RQ1 estimates from the full specification appear in Table C.2.	123
5.5	Branded vs. non-branded restaurant elasticities, 2019–2024. Annual within-POI spend–visit elasticities estimated separately for branded and non-branded restaurants using Eq. 5.1 with POI and year fixed effects; shaded bands show 95% confidence intervals.	125

5.6	Socioeconomic Moderation of Visit-Spend Elasticities Across Business Categories and Metropolitan Areas. Each panel displays coefficients for a specific demographic variable: Income (median household income percentile) and Education (percentage with bachelor’s degrees or higher). Rows represent business categories organized by economic type: Essential Goods , Routine Services , Social & Experiential , and Discretionary Retail . Columns within each panel show the six MSAs. Color intensity indicates coefficient magnitude and direction, with blue shades representing positive moderation effects (higher demographic characteristic associated with stronger visit-spend relationship) and red/orange shades representing the converse. Asterisks denote statistical significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, while cells without asterisks indicate non-significant coefficients.	127
5.7	Racial Moderation of Visit-Spend Elasticities Across Business Categories and Metropolitan Areas. Each panel displays coefficients for a specific demographic variable: Racial/Ethnic Composition (percentage Black, Hispanic, and Asian populations, with White population as the reference category). Rows represent business categories organized by economic type: Essential Goods , Routine Services , Social & Experiential , and Discretionary Retail . Color intensity indicates coefficient magnitude and direction, with blue shades representing positive moderation effects (higher demographic characteristic associated with stronger visit-spend relationship) and red/orange shades representing the converse. Asterisks denote statistical significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, while cells without asterisks indicate non-significant coefficients.	130
6.1	Sample size distributions for CBGs across four CBSAs.	155
6.2	Mean 3-class prediction accuracy across four CBSAs for mobility outcomes inferred from socio-demographic and built-environment predictors. Outcomes are grouped into point-level, line-level, and temporal mobility features; bars show the top two LLMs and top two supervised baselines for each outcome, averaged over repeated 20% held-out test splits, with error bars denoting standard deviation across 5 repetitions. The horizontal dashed line marks chance accuracy (1/3), and the black line shows the performance gap between the best baseline and best LLM for each outcome. Across cities, supervised baselines consistently outperform zero-shot LLMs, although LLMs remain above chance on many tasks, with the strongest performance typically observed for point-level mobility outcomes.	157
6.3	Empirical and LLM-implied directional relationships between contextual predictors and mobility outcomes across the four study CBSAs. Left panels show standardized OLS coefficients for each predictor–outcome pair, where red indicates a negative empirical association and blue indicates a positive empirical association; only statistically significant coefficients are marked. Right panels show the corresponding LLM-based directional alignment results from the Jonckheere–Terpstra analysis, restricted to predictor–outcome pairs with significant OLS effects. Blue cells indicate that the LLM-implied direction is aligned with the empirical relationship, red cells indicate an opposite trend, and white cells indicate no clear monotonic trend.	161

A.1	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with LTW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).	176
A.2	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ridge regression with LTW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).	178
A.3	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with STW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).	180
A.4	Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for elastic regression with STW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).	182
B.1	Timeline Overview of COVID-19 Model Forecasts by County. This Gantt chart captures the forecast submission frequency from the 36 research teams' models analyzed in this paper, delineating the evolution of modeling efforts throughout the pandemic.	185
B.2	Distribution of the PBL metric used to evaluate COVID-19 case prediction errors. Top row shows the original (left) and square root transformed (right) distributions of the raw PBL data. Bottom row shows the same distributions after trimming the highest 1% of values, demonstrating improved visualization of the data's shape and reduction of extreme outliers.	192
B.3	Diagnostic Plots for GLM-1 (Race and Ethnicity Main Effects Model)	196
B.4	Diagnostic Plots for GLM-2 (Urbanicity Main Effects Model)	199

B.5	Regression diagnostics for the Generalized Linear Models (GLM) analyzing interactions between race/ethnicity and model characteristics: (a) Forecast lookahead interactions (GLM-1a) showing diagnostic plots for models analyzing how prediction horizons (7-28 days) affect racial disparities; (b) Pandemic phase interactions (GLM-1b) presenting diagnostic assessments for models examining how racial disparities vary across different COVID-19 phases; (c) Model type interactions (GLM-1c) showing residual patterns for analyzing racial disparities across different model architectures (compartmental, deep learning, statistical, etc.); and (d) Mobility data usage interactions (GLM-1d) displaying diagnostics for models examining how the incorporation of mobility data impacts prediction fairness across racial groups. Each subplot contains standard regression diagnostic plots: observed vs. fitted values (top left) residuals vs. fitted values (top right), normal Q-Q plot (middle left), scale-location plot (bottom left), and distribution of deviance residuals (bottom right) to validate model assumptions of normality, homoscedasticity, and identify potential influential observations.	203
B.6	Regression diagnostics for the Generalized Linear Models (GLM) analyzing interactions between urbanicity and model characteristics: (a) Forecast lookahead interactions (GLM-1a) showing diagnostic plots for models analyzing how prediction horizons (7-28 days) affect racial disparities; (b) Pandemic phase interactions (GLM-1b) presenting diagnostic assessments for models examining how racial disparities vary across different COVID-19 phases; (c) Model type interactions (GLM-1c) showing residual patterns for analyzing racial disparities across different model architectures (compartmental, deep learning, statistical, etc.); and (d) Mobility data usage interactions (GLM-1d) displaying diagnostics for models examining how the incorporation of mobility data impacts prediction fairness across racial groups. Each subplot contains standard regression diagnostic plots: observed vs. fitted values (top left) residuals vs. fitted values (top right), normal Q-Q plot (middle left), scale-location plot (bottom left), and distribution of deviance residuals (bottom right) to validate model assumptions of normality, homoscedasticity, and identify potential influential observations.	204
B.7	Forecast Hub Fairness Dashboard showing AER distribution for urbanization levels. Models are organized by type, with teams sorted within each type by median AER values. The AER values compare prediction errors between Micropolitan areas and Large Metropolitan areas (reference group), where values above 1.0 indicate higher prediction errors for Micropolitan regions.	221
B.8	Forecast Hub Fairness Dashboard displaying the interaction between race and forecast horizon (lookahead). The visualization shows how Hispanic-White AER values vary across different forecast periods (7, 14, 21, and 28 days), allowing assessment of how prediction fairness changes with forecast horizon length(lookaheads).	222
B.9	Forecast Hub Fairness Dashboard illustrating the relationship between race and pandemic phases. The visualization reveals how Hispanic-White AER values evolved across different pandemic phases (0-6), enabling analysis of how prediction fairness varied during different stages of the pandemic.	223

B.10	Forecast Hub Fairness Dashboard showing the interaction between urbanization levels and forecast horizon. The visualization demonstrates how Micropolitan-Large Metropolitan AER values change across different forecast periods (7, 14, 21, and 28 days), highlighting the relationship between prediction fairness and forecast horizon lengths (lookaheads).	224
B.11	Forecast Hub Fairness Dashboard depicting the relationship between urbanization levels and pandemic phases. The visualization shows how Micropolitan-Large Metropolitan AER values varied across different pandemic phases (0-6), revealing temporal patterns in prediction fairness across the urban-rural spectrum.	225
C.1	State-level alignment between population and SafeGraph Spend panel customers (monthly). This figure plots the monthly Pearson correlation coefficient computed across U.S. states between (i) each state’s Census Bureau annual population estimate (assigned to all months within that year) and (ii) SafeGraph Spend’s TOTAL_CUSTOMERS (the number of unique panel customers assigned to each state in that month). Higher values indicate that the geographic density of the Spend panel more closely mirrors population density at the state level. Correlations remain consistently high throughout the study period (minimum $r \geq 0.82$), suggesting that the panel’s customer distribution tracks population density closely overall.	229
C.2	Distribution of geographic sampling bias in the SafeGraph Spend panel over time. This plot summarizes the monthly distribution of absolute state bias in SafeGraph Spend panel coverage, where state bias is defined as the percentage-point difference between a state’s share of SafeGraph Spend panel unique customers (observed share) and its share of the U.S. population from ACS annual population estimates (expected share), and we plot $ bias $. For each month, the solid line shows the median $ bias $ across states, the shaded band shows the interquartile range (25th-75th percentiles), and the dashed line shows the 95th percentile, capturing tail deviations. The median and IQR remain low and stable over time, indicating modest typical state-level deviations, while the 95th percentile highlights periods when a small number of states exhibit larger deviations from proportionality.	230

Chapter 1: Introduction

1.1 Motivation & Background

High-stakes decisions are increasingly mediated by data and models [108, 151]; in domains such as public health, economic policy, and crisis response, policymakers and researchers are often required to act under conditions of urgency, uncertainty, and incomplete information. Yet the phenomena they seek to understand—disease spread, behavioral adaptation, or local economic disruption—typically evolve much faster than the traditional data systems used to measure them. Official statistics, surveys, and administrative records are often delayed, incomplete, or too spatially coarse to support timely intervention; as a result, a growing gap has emerged between the speed of social change and the pace of conventional measurement [30, 46].

The COVID-19 pandemic made this dependence especially visible; when the pandemic began, policymakers and modelers faced an unprecedented challenge of having to track disease spread and guide public response in real time [79, 132]. In the United States, this urgency gave rise to a large-scale forecasting effort across academia, industry, and public health institutions [56]. The U.S. COVID-19 Forecast Hub [33] emerged as a central repository of standardized forecasts from dozens of research groups and was used by the Centers for Disease Control and Prevention in official communications; it therefore functioned not only as a scientific coordination platform, but also as a practical component of public decision-making.

This broader turn toward modeling was not confined to epidemic forecasting; similar pressures arise whenever institutions require timely signals about fast-changing social and economic conditions [46, 160]. In the context of economic disruption and recovery, for example, researchers and policymakers increasingly need ways to monitor change as it unfolds rather than after it has already occurred. This has encouraged the use of scalable data sources and predictive models that can provide near real-time insight into business activity, recovery trajectories, and broader patterns of resilience when conventional indicators are unavailable or slow to update [46, 160].

More fundamentally, this shift reflects a fundamental measurement problem in that many high-stakes outcomes are not directly observable at the time and scale at which action must be taken. Researchers, therefore, turn to proxies, behavioral signals, and predictive models to infer underlying dynamics from partial observations [107, 176]. This logic has shaped work across multiple domains; during COVID-19, for instance, models increasingly incorporated behavioral data such as mobility to capture aspects of human movement and mixing related to transmission risk [30, 42, 101, 131, 139]. More broadly, digital trace data have been used to support decision-making in socio-economic development, public safety, disaster response, and health forecasting [27, 140, 171, 177].

Among these emerging behavioral data sources, human mobility data became especially prominent because it appeared to provide a direct lens into movement, mixing, and adaptation—processes that are central to a wide range of high-stakes outcomes [84]. Movement is a fundamental dimension of social and economic life [6, 179, 181]; it shapes how people access opportunities, interact with infrastructure, and experience places in everyday life [176]. Because mobility patterns influence access to jobs and services, pandemic spread, and broader forms of social interaction, mobility data became central to a wide range of analytical and policy questions. The increasing availability of smartphone- and mobile-phone-based location data substantially expanded this potential by providing far finer spa-

tial and temporal resolution than traditional travel surveys and related sources [18, 30].

However, mobility data are not unbiased measures of social reality. Like many digital traces, they are generated through unequal processes of device ownership, platform participation, and data collection, which can lead to systematic distortions in who and what is represented [1, 48, 77, 154]. As a result, the same data that appear valuable for real-time inference may also provide incomplete or biased views of populations and places. This concern is especially important because such distortions do not necessarily remain at the level of measurement alone; when predictive models are built on top of biased signals, those biases may be carried forward into downstream analyses, predictions, and decisions [7, 19, 151].

These concerns point to a broader challenge in data-driven inference. The central issue is not simply whether behavioral and contextual data are useful, but how they should be evaluated before being trusted in high-stakes settings. These concerns suggest that such systems should be evaluated along at least three dimensions: *utility*, *fairness*, and *validity*. This concern also extends beyond direct behavioral traces themselves. When mobility data are costly, privacy-sensitive, or difficult to access, alternative models may be used to infer behavior from context; yet these alternatives, too, require careful evaluation to determine whether they encode meaningful and empirically grounded urban knowledge.

A first gap concerns *utility*. Behavioral and contextual data are often adopted because they appear to improve inference; however, existing work on mobility-augmented prediction tends to focus on whether performance improves in aggregate, masking the more practically important questions of when it helps, by how much, and under what conditions [53, 101, 109]. In the context of COVID-19 case prediction, for instance, prior studies reported improvements from adding mobility data without systematically analyzing the counties, datasets, or modeling choices for which those improvements

did and did not materialize—leaving decision-makers without the conditional understanding needed to assess whether such data is worth its cost [1]. The first study in this dissertation addresses this gap directly, asking not whether mobility data improves prediction on average, but under what circumstances its value can be reliably demonstrated.

A second gap concerns *fairness*. Even when data-driven systems perform well on average, their errors may be distributed unevenly across populations and places. This concern is grounded in empirical observation; mobility-based analyses of small urban businesses during the pandemic have already documented that visit declines and recovery trajectories differed substantially across neighborhoods by ethnoracial composition, with majority-Asian neighborhoods experiencing both the steepest declines and the slowest recovery [3]. Given that COVID-19 burden was itself distributed unevenly across racial and ethnic groups, predictive models trained and evaluated without accounting for this heterogeneity may perform inequitably across the very populations most affected. If the behavioral signals feeding predictive models encode such disparities, there is reason to expect that model outputs may inherit and amplify them. The U.S. COVID-19 Forecast Hub [33, 51], whose predictions informed CDC communications, was extensively evaluated for accuracy; however, no systematic analysis had examined whether its models performed equitably across social determinants such as race, ethnicity, and urbanization—dimensions known to shape both COVID-19 burden and data quality [2]. Accuracy alone is insufficient as a standard when the downstream use of predictions may shape attention, resources, and intervention. The second study addresses this gap by conducting a fairness audit of the Forecast Hub’s case prediction models, treating equitable performance as a first-class evaluation criterion.

A third gap concerns *validity*. Many data-driven systems rest on the assumption that an observable signal reliably proxies a harder-to-measure outcome [140, 177, 179]—but this assumption is

rarely tested in a way that would reveal where it breaks down. In the case of mobility as an economic proxy, existing work finds generally positive co-movement between foot traffic and spending [31, 40], yet most analyses operate at regional or broad business-type aggregates that may conceal instabilities at finer levels; they provide limited evidence on whether these relationships hold across time, especially through the disruptions of the COVID-19 period. They also largely ignore how neighborhood demographic composition moderates the visit–spend relationship across business types. The third study addresses these gaps by systematically evaluating whether foot traffic reliably mirrors economic activity across categories, time periods, and demographic contexts.

Finally, these challenges point to a fourth gap concerning *alternatives*. Because direct behavioral traces are often expensive, access-restricted, or privacy-sensitive [54], there is growing interest in models that infer mobility from contextual information alone—including large language models, which may encode latent knowledge about urban structure through their training corpora. Yet existing evaluations of LLMs on urban and mobility tasks have largely assessed whether they can produce plausible outputs, without establishing whether they encode empirically meaningful priors about how mobility actually varies across neighborhoods and cities. Plausibility alone is not the same as structural validity. The fourth study addresses this gap by evaluating whether LLMs can recover the relationships between built-environment conditions, neighborhood sociodemographics, and observed mobility behavior—asking whether they constitute a genuinely informative alternative when behavioral traces are limited or not available.

Umbrella Gap		Description
Auditing	Utility	Behavioral and contextual data are often assumed to improve inference, but their benefits may be heterogeneous across tasks, locations, and modeling choices.
	Fairness	Data-driven systems may perform well on average while still producing systematically uneven errors across populations and places.
	Validity	Behavioral traces are often treated as proxies for harder-to-measure phenomena, but these proxy relationships may be unstable or context-dependent.
	Alternatives	When direct behavioral traces are costly, privacy-sensitive, or unavailable, alternative models must also be evaluated to determine whether they capture meaningful and empirically grounded structure.

Table 1.1: Broader gaps motivating this dissertation under the umbrella of auditing.

1.2 Research Objectives and Questions

Taken together, these four gaps reflect a common underlying problem: that behavioral and contextual data are routinely incorporated into high-stakes inference without the kinds of systematic evaluation needed to establish whether they can be trusted, and for whom. Addressing them requires not just improving models, but auditing the data and systems built on top of them along multiple dimensions. With this overarching goal in mind, this dissertation is organized around four recurring gaps in contemporary data-driven decision-making in high-stakes settings: *utility*, *fairness*, *validity*, and *alternatives*. Specifically, it asks when behavioral and contextual data improve inference, for whom these improvements hold, when such data serve as valid proxies for underlying phenomena, and whether alternative models can provide useful priors when direct behavioral traces are limited. Table 1.1 provides a conceptual summary of these gaps under the broader umbrella of *auditing*.

Research Objective 1 (Utility): Evaluating the Impact of Mobility Data on COVID-19 Case Prediction. This study examines the extent to which human mobility data improves county-level

COVID-19 case prediction. It evaluates whether mobility meaningfully enhances predictive performance, under what conditions those gains occur, and whether differences across counties are associated with demographic and socioeconomic characteristics.

This study addresses the following research questions:

- **RQ1.1:** Under what conditions does human mobility data provide an enhancement over county-level COVID-19 case prediction models that do not use mobility as a source of information?
- **RQ1.2:** Does mobility data bias—characterized by the demographic and socioeconomic characteristics of each county—help explain differences in the performance of COVID-19 prediction models across counties?
- **RQ1.3:** Do the improvements associated with mobility data vary across mobility datasets, predictive models, and training approaches?

Research Objective 2 (Fairness): Auditing the Fairness of Large-Scale COVID-19 Forecasting Models. This study presents a large-scale fairness audit of county-level COVID-19 case prediction models in the U.S. COVID-19 Forecast Hub. It evaluates whether forecast errors differ systematically across race, ethnicity, and urbanization levels, and whether these disparities persist across different model and data characteristics.

This study addresses the following research questions:

- **RQ2.1:** Does the prediction performance of U.S. COVID-19 Forecast Hub county case prediction models vary across race, ethnicity, and urbanization level?
- **RQ2.2:** Do these differences persist across model and data characteristics such as lookahead, pandemic phase, model type, and the inclusion of mobility data?

Research Objective 3 (Validity): Evaluating Mobility Data as a Proxy for Economic Activity. This study investigates whether foot traffic can reliably serve as a proxy for economic activity. Using point-of-interest visits and card spending data across multiple metropolitan areas and business categories, it evaluates the strength, temporal stability, and demographic moderation of visit–spend relationships.

This study addresses the following research questions:

- **RQ3.1:** Do visit–spend elasticities remain stable over time within business categories, and which categories serve as the most reliable proxies for economic activity?
- **RQ3.2:** How do neighborhood demographic characteristics moderate the visit–spend relationship across business categories?

Research Objective 4 (Alternatives): Evaluating Large Language Models as Priors for Urban Mobility. This study explores whether large language models encode useful priors about mobility behavior in the absence of direct trajectory data. Using neighborhood sociodemographic and built-environment context as input, it evaluates whether LLMs can predict neighborhood-level mobility outcomes and whether those predictions align with empirically observed mobility structure.

This study addresses the following research questions:

- **RQ4.1:** To what extent can LLMs predict neighborhood-level mobility outcomes given built-environment and sociodemographic context?
- **RQ4.2:** To what extent do LLM predictions align with empirically observed relationships between neighborhood characteristics and mobility behavior?
- **RQ4.3:** Do the mobility-relevant priors encoded by LLMs transfer consistently across different

cities and mobility outcomes?

1.3 Dissertation Organization

The remainder of this dissertation is organized as follows.

Chapter 2 evaluates the impact of human mobility data on county-level COVID-19 case prediction. It compares mobility-based models with non-mobility baselines, quantifies when mobility improves predictive performance, and examines whether differences in those improvements are associated with demographic and socioeconomic characteristics of counties.

Chapter 3 presents a large-scale fairness audit of county-level COVID-19 case prediction models from the U.S. COVID-19 Forecast Hub. It examines whether forecast errors vary systematically across race, ethnicity, and urbanization level, and whether these disparities persist across different models and data characteristics such as lookahead, pandemic phase, model type, and the use of mobility data.

Chapter 4 presents an exploratory analysis of ethnoracial disparities in COVID-19’s impact on small urban restaurant businesses, illustrating how mobility-derived behavioral signals can also reveal unequal economic disruption across communities.

Chapter 5 investigates whether foot traffic reliably reflects economic activity. Using visit and spending data across multiple metropolitan areas and business categories, it evaluates the temporal stability of visit–spend relationships and examines how these relationships vary across neighborhood demographic contexts.

Chapter 6 evaluates large language models as priors for urban mobility. It examines whether LLMs can infer neighborhood-level mobility outcomes from built-environment and sociodemographic

context, whether those predictions align with empirically observed mobility patterns, and whether such priors transfer across cities and mobility outcomes.

Chapter 7 concludes the dissertation by summarizing the main findings, discussing their broader implications for the use of data and models in high-stakes settings.

Chapter 2: Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models

2.1 Introduction

The COVID-19 pandemic has mainstreamed human mobility data into the public domain, and beyond academic networks. During the early stages of the pandemic, the importance of limiting mobility to contain the epidemic became evident, with cities, states and countries taking various non-pharmacological interventions (NPIs) focused on mobility such as national lockdowns or work-from-home approaches [65, 139]. To evaluate the effect of these interventions, public health experts, the CDC, city departments and journalists explored the use of mobility data which, at the time, was made open and freely available. Companies like Apple, Google, SafeGraph or Descartes shared different types of aggregated mobility datasets to characterize behaviors such as the volume of visits to specific places (e.g., schools, workplaces or restaurants), the volume of trips between regions (e.g., trips between two counties), or the volume of trips by type of transportation (e.g., driving vs. public transit).

Beyond understanding the impact of mobility reduction policies, the increased access to mobility data sources has also supported research on regional COVID-19 case prediction models, with the assumption that how people moved within a region in the past could potentially provide additional information about how people get infected in the future. COVID-19 case prediction models focus on

providing regional-level estimates for future number of cases in the short- and long-term via lookahead analysis performance *i.e.*, measuring region-level prediction performance for various temporal windows such as daily, weekly or monthly [101]. For example, researchers have shown that SafeGraph data can help predict weekly COVID-19 cases at the county level in the US, providing higher accuracy when compared to non-mobility baselines [131]. There exist a wide variety of models to predict regional COVID-19 cases including epidemiological [43, 98, 174], machine learning [78, 109, 166] and statistical models [78, 101]. In this paper, we focus on statistical models (linear regression and ARIMA) because we are interested in the deployment of models that are interpretable by decision makers, rather than implementing black-box predictive approaches that are harder to explain [149].

Nevertheless, there are several gaps in the current state of the art in regional COVID-19 case prediction using mobility data. First, performance results - measured as RMSE or correlation between actual and predicted regional COVID-19 cases - are reported as averages across regions, masking individual region-level performance, which is critical to inform local interventions and policies [101]. For example, past research has shown that mobility data enhances COVID-19 case predictions, on average, across counties in the US; however, that average might be masking counties for which it did not work [53, 109]. Second, performance results are often times not compared against non-mobility baselines, making it hard to measure the effectiveness of adding mobility data to the prediction model [53, 109, 166]. Third, prior work has shown that mobility data might suffer from sampling bias whereby certain demographic groups *e.g.*, Black, elder and low-income individuals can be under-represented in the data due to lower smartphone and cell phone ownership rates [48, 154]. Nevertheless, prior work focused on building COVID-19 case prediction models tends to ignore the bias present in the mobility data, which in turn, might affect the performance of regional COVID-19 case prediction models depending on the population of that region [101, 109, 126]. Fourth, current

approaches tend to provide narrow evaluations, focused on a few models, or on one or a few mobility datasets, with little research broadly looking into the impact of different prediction models, mobility datasets, and training approaches that use more or less data, on model performance. Given (1) the high cost of acquiring human mobility data for COVID-19 prediction purposes, now that it is no longer freely accessible, and (2) that COVID-19 case predictions are going to be used to assess NPIs such as mobility reduction, or vaccine distribution at the local level, we posit that it is critical to understand the conditions under which mobility data helps (or not) at the individual regional level so that it can adequately inform local decision-making.

In this paper, we aim to analyze the conditions under which human mobility data provides an enhancement over individual regional COVID-19 case prediction models that do not use mobility as a source of information. Our main objective is to inform regional decision makers about the potential of region-level COVID-19 case prediction models that use mobility data, which we posit should be well understood given the high cost of human mobility data. The main contributions of this paper are:

- Focusing on US counties, we evaluate the number of counties that benefit from adding mobility data, and quantify the improvements. Our analyses show that, at most, 60% of counties improve their performance over non-mobility baselines; and that those improvements are modest, happening mostly for longer-term predictions. Looking into the counties that benefit from adding mobility data, 50% of those counties show modest correlation improvements of at most 0.1 and 25% show correlation improvements of at most 0.3.
- We present and discuss an approach to assess whether mobility data bias - characterized by demographic and socio-economic characteristics of each county - might explain the differences in the performance of COVID-19 prediction models across counties. We show that correlation

improvements were lower for counties with higher Black, Hispanic, and other non-White populations as well as low-income and rural populations, pointing to potential bias in the mobility data negatively impacting predictive performance.

- We analyze whether the differences in the performance of mobility-based models over non-mobility baselines vary depending on the mobility datasets; the predictive model; or the training approach. Our results reveal that the improvements brought about by mobility data are similar across mobility datasets, albeit with slightly better values for Apple and SafeGraph; that linear regressions are associated with larger improvements; and that the training approach might also affect the scale of the improvements.

2.2 Related Work

2.2.1 Human mobility data and COVID-19 case predictions

Human mobility data has been used in the past to model and characterize human behaviors in the built environment [70, 93, 148, 164, 170], to support decision making for socio-economic development [71, 72, 73, 75, 95], for public safety [171, 172], as well as during epidemics and disasters [23, 81, 94, 96, 102, 169]. During the COVID-19 pandemic, human mobility has also played a central role in driving decision making, for example, with social distancing policies significantly reducing the spread of the virus [14].

Related work has shown that COVID-19 case prediction models can be enhanced using human mobility data, when compared to non-mobility baselines [41, 101, 131]. For example, Ilin et al. [101] analyzed the use of mobility data to forecast COVID-19 cases using interpretable statistical models

like regressions [101]. Working at various spatial scales (from county to state to country), the authors revealed that adding mobility data significantly helps decrease the mean percentage prediction error; and that the improvements were higher for longer forecasting lengths. Nevertheless, a significant amount of papers focused on COVID-19 case prediction using mobility data fail to compare model performance against non-mobility baselines [53, 109]. More importantly, several papers have revealed settings in which mobility data did not help. For example, Curtis et al. [52] and Venter et al. [163] found only a small correlation between COVID-19 cases and mobility in parks and natural areas (blue-green spaces); and Mehrab et al. [126] showed that the performance of mobility-based prediction differed considerably across 50 US counties that correspond to land-grant universities.

Human mobility data is generally collected from smartphones and cell phones; however, due to the differences in access to that technology, not all individuals are equally represented in mobility datasets. In fact, prior work has shown that, for example, Black and elder individuals were under-represented in SafeGraph’s dataset for the state of North Carolina [48]; that wealthier individuals tend to be over-represented in cell phone data from several countries including Sierra Leone or Iraq [154]; and that the relationship between COVID-19 cases and mobility is stronger in urban than rural areas [107]. Given this evidence, we posit that the sampling bias found in mobility data might also affect the performance of regional COVID-19 case prediction models.

In this paper, we aim to provide a much needed systematic analysis to evaluate the conditions under which mobility data can enhance county COVID-19 case prediction models, and to quantify by how much when compared to non-mobility baselines. Thus, we will extensively evaluate performance across types of prediction models, temporal prediction windows (lookaheads), mobility datasets, training-testing approaches as well as county-level demographic and socio-economic characteristics associated with potential mobility data bias that could affect performance.

2.2.2 COVID-19 prediction models

There exist a wide variety of models to predict regional COVID-19 cases including epidemiological, machine learning and statistical models. Epidemiological models, such as SEIR and SIR models, have been used to predict infection rates, and in certain cases, related work has shown that the models can be improved with human mobility data characterizing how people (agents) travel and might infect others [43, 98, 174]. However, SEIR/SIR models have a number of pitfalls, such as the large number of parameters that need to be adjusted [183]; or the complexities of adding mobility responses of the population as a function of time and space [146]. For example, Roda et al. [146] found simpler SIR models to be more effective in predicting COVID-19 cases than the more complex SEIR models.

A number of machine learning techniques have also been widely applied for regional COVID-19 case prediction with mobility data, including tree-based and K-nearest neighbors models [109]. Furthermore, several articles [78, 166] make use of a range of deep learning architectures, including sequential models like Long Short-Term Memory (LSTM) [131], Gated Recurrent Unit (GRU), Recurrent Neural Networks (RNN) [88] and spatio-temporal models like Graph Neural Networks (GNNs) [106]. Some researchers have also incorporated static and dynamic mobility flows - characterizing average and daily mobility patterns between regions - as well as friendship networks to understand the spatio-temporal dependencies between regions that might affect infection rates and which could inform the prediction of regional COVID-19 cases [74, 166].

Statistical models have also been popular for predicting COVID-19 cases since they are transparent and simple to interpret - which is highly important for decision makers; these models can also easily incorporate mobility data along with other features. Statistical models such as autoregressive time series and linear regression have been used for COVID-19 case prediction showing that - early

on during the pandemic - mobility data improved non-mobility baselines across countries, states and other administrative levels [78, 101]. In this paper, we focus on statistical models (linear regression and ARIMAX) because we are interested in the deployment of models that are interpretable by decision makers, rather than implementing black-box predictive approaches that are harder to explain [149].

2.3 Datasets

To assess the effectiveness of mobility data on county-level COVID-19 case prediction models we evaluate model performance across nine mobility datasets from four different companies: SafeGraph, Google, Descartes and Apple. We describe each of the COVID-19 and mobility datasets in detail. All the datasets used in this paper, were freely available during the onset of the pandemic. We focus on the period from 2020-03-18 to 2020-11-30 (258 days): the former marks the start of the consistent case data availability in the United States and the latter the date when vaccines were introduced (the Pfizer-BioNTech COVID-19 vaccine was made available on Dec.11, 2020). We focus on the pre-vaccine period to prevent immunity levels from acting as a confounder, since the relationship between mobility and infections post-vaccines is less clear in the literature [80].

COVID-19¹. We use the COVID-19 case data compiled by the New York Times at the county level. To account for peaks in daily COVID-19 case counts due to delayed reporting, we use 7-day daily rolling average of COVID-19 cases (computed as the average of its current value and the prior 6 days) instead of raw counts. We acknowledge that especially during the early stages of the pandemic, case numbers might not be reflective of the actual spread of COVID-19, in large part due to the lack of testing resources [145].

¹<https://github.com/nytimes/COVID-19-data>

SafeGraph.² Curated by tracking the movements of millions of *anonymized* users via mobile app SDKs, SafeGraph open-sourced the mobility patterns of app users at the onset of the pandemic. Based on the data available, we use two types of features from SafeGraph datasets: daily O-D flows [105] and daily visits to points of interest (POIs). O-D (county-to-county) flows represent daily volumes of trips between pairs of counties across the US, while visit to POIs represent the daily volume of visits to grocery stores, restaurants, religious organizations and schools in a given county. For O-D flows, we retrieve from SafeGraph both Inflows *i.e.*, incoming flows to county D from county O, and outflows *i.e.*, outgoing flows from county O to county D. All mobility features are measured as changes in volumes with respect to a baseline of *normal* behavior computed by SafeGraph using mobility data from February 17, 2020 till March 7, 2020.

Google.³ Google collects mobility data from users who have the location data collection option selected. During the pandemic, Google provided daily county level mobility scores across different POI categories including parks, residential, and transit stations. Mobility scores are calculated as the ratio between the volume of visits on a given day during the pandemic and the volume of visits during a pre-pandemic baseline; with the baseline computed by Google as the median value for each day of week in the 5-week period from January 3rd to February 6th, 2020. Among all POI available, we selected workplace, the category with the greatest number of counties with daily data availability in our chosen time period.

Descartes Labs.⁴ This mobility data is calculated using geolocation data from mobile devices, and captures the median of the maximum distances traveled by individuals in each county each day. As with Google, this median is converted by Descartes to a ratio of pandemic mobility to baseline

²<https://www.safegraph.com/>

³<https://www.google.com/covid19/mobility/>

⁴<https://github.com/descarteslabs/DL-COVID-19>

pre-pandemic mobility.

Apple⁵. Apple’s mobility data is collected from Apple Maps, and divides its categories by transportation method: driving, walking, and transit. We selected the *driving* category to measure the volume of individuals driving on a daily basis at the county level as, similar to Google, this category had more consistent daily data availability than other categories.

Across all datasets, we only consider counties that have COVID-19 case data and mobility data available daily throughout the time period of study. Table 2.1 shows the number of counties that fit this criteria for each dataset; notably, Google has by far the fewest counties with 990 out of the total 3,143 US counties being represented.

Mobility Dataset	County Counts
Apple	2064
Descartes	2551
Google	990
SafeGraph Inflows	3116
SafeGraph Outflows	3116
SafeGraph POI (Grocery Stores)	3065
SafeGraph POI (Religious Organizations)	3076
SafeGraph POI (Restaurants)	3091
SafeGraph POI (Schools)	3084

Table 2.1: Number of counties considered for each mobility dataset.

2.4 Methodology

In this paper, we aim to provide a much needed systematic analysis to evaluate the conditions under which mobility data can enhance county COVID-19 case prediction models, and to quantify by how much when compared to non-mobility baselines. Our main objective is to inform regional decision makers about the potential and pitfalls of COVID-19 case prediction models that use mobility

⁵<https://covid19.apple.com/mobility>

data, given the high cost of acquiring such data. Next, we describe the prediction models we evaluate and their parameter adjustment, the training-testing approaches, and the overall evaluation approach.

2.4.1 Prediction Models

As described in section 2.2, we consider two types of predictive models commonly used by decision makers due to their interpretability: linear regressions and time series. In contrast to more complex epidemiological models that are hard to tune due to their parametric nature, and to deep learning models with black box architectures, linear models and time series are easier to interpret providing decision makers with the ability to clearly explain their policies [146, 149, 183].

To evaluate the effectiveness of mobility data in the prediction of county-level COVID-19 cases, we analyze the performance of linear regressions and time series using (1) only past county-level COVID-19 case data as an independent variable to predict future COVID-19 cases (these are the non-mobility baselines), and (2) both past county-level COVID-19 case data as well as county-level mobility data as independent variables; with the assumption that how people moved in the past could potentially provide additional information about how people get infected in the future. Both non-mobility baselines and mobility-based models will be evaluated across five different temporal prediction windows (a.k.a. lookaheads): 1-day, 7-days, 14-days, 21-days and 28-days. Next, we provide further details for each predictive model; and in section 4.4, we discuss details about the parameter tuning.

2.4.1.1 Linear Regression.

We train one linear regression model per county and lookahead. For the non-mobility baselines, the number of county-level COVID-19 cases for a given lookahead is predicted using the COVID-19

cases from the previous lookahead value *e.g.*, for lookahead 1, the COVID-19 cases in day x are used to predict cases for day $x+1$, while for lookahead 14, the COVID-19 cases in day x are used to predict cases in day $x+14$. For mobility-based models, the number of county-level COVID-19 cases for a given lookahead is predicted using (i) the COVID-19 cases from the previous lookahead value as well as (ii) the 10-day lagged mobility features *i.e.*, the mobility features from day x are used to predict COVID-19 cases for day $x+10$, since the infection gap - incubation period from exposed to being able to spread the virus - has been associated with that lag by many reports in the literature [78, 101, 182]. Lagged mobility data is only included in the mobility-based models that will be compared against the non-mobility baselines.

2.4.1.2 ARIMAX

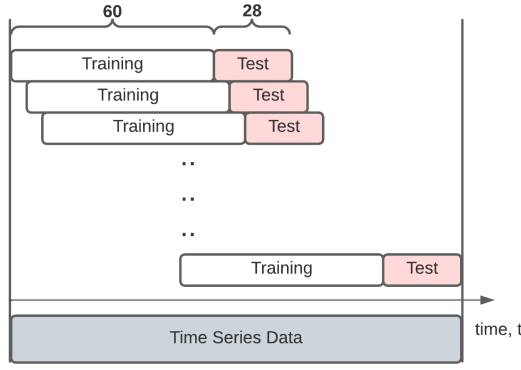
We train one ARIMA time series forecasting model per county, and use it to predict county-level COVID-19 cases for all five lookaheads. Similarly to linear regression, we train non-mobility baselines exclusively with COVID-19 case time series, while mobility-based models incorporate 10-day lagged mobility data series. Mobility data is incorporated using ARIMAX models that allow to add exogenous covariates into time series forecasting models. Arima(x) forecasting models have three components that function as parameters: (1) p (autoregressive order, AR) indicates the number of lag observations that the dependent variable regresses with in the model; (2) d (integrated order, I) represents the number of times the time series needs to be differenced to achieve stationarity; and (3) q (moving average, MA) represents the size of the moving average window that models the relationship between the error terms of the moving average model and the lagged observations. We discuss the tuning of these parameters in section 4.4.

2.4.2 Training-Testing Approaches

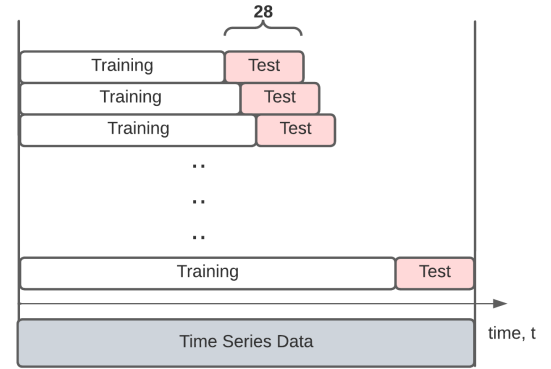
In order to explore the conditions under which mobility helps in improving predictive performance, we aim to test two different training-testing approaches that have been used in the literature: the long training window (LTW) approach and the short training window (STW). The STW approach focuses on the use of small training datasets to test prediction accuracy for the next 28 days (lookaheads 1, 7, 14, 21 and 28); while the LTW approach uses much larger training datasets to predict values for the five lookaheads. Given the high cost of acquiring human mobility data - in fact, none of the companies described earlier offer their data for free anymore - the objective of evaluating these two training approaches is to understand the impact of cost on the performance of COVID-19 prediction models *i.e.*, buying more (LTW) versus less mobility data (STW). Next, we explain the implementation of LTW and STW for linear regression and ARIMAX models (also depicted in Figure 2.1).

2.4.2.1 Linear Regression

The LTW approach creates one regression model per lookahead per county, trained on the first 70% of days in our dataset and tested on the remaining 30%. This splitting approach is consistent across lookaheads, easing interpretability, but also meaning that the split is not exactly 70/30 for each lookahead *i.e.*, at higher lookaheads, the training set becomes smaller while the testing set remains the same size. The STW approach uses a 60-day sliding window for the training set and a sliding 28-day window for the testing set (smaller training sizes produced extremely low performing results and were not considered). Unlike the LTW approach, the testing set consists of 1 date for each lookahead (1 day after the end of the training set, 7 days after, etc.). Results are reported by averaging performance metrics across all testing datasets.



(a) Short training window (STW) for ARIMAX and Linear



(b) Long training window (LTW) for ARIMAX

Figure 2.1: Training-Testing approaches for LTW and STW. The LTW approach for Linear Regression is not represented because it is the standard 70% – 30% approach.

2.4.2.2 ARIMAX

The ARIMAX LTW approach uses an expanding window protocol [143] [100] that first trains the ARIMAX with a data window that expands over 70% of the entire time series, and is tested for the next 28 days to evaluate accuracy per lookahead. After that, the training data window is expanded by one day at a time, without dropping older data points, and tested for subsequent 28-day windows to assess accuracy at each lookahead. On the other hand, the STW adopts a sliding window approach [143] whereby the training window length remains fixed at each train-test step with a length of 60 days. Each training window is computed by shifting by one with respect to the prior window, effectively discarding older data points. Results are reported by averaging performance metrics across all testing datasets.

2.4.3 Model performance

We measure individual county prediction performance as the correlation between the predicted volume of COVID-19 cases for that county and its actual case numbers retrieved from the COVID-19 official report dataset. To assess the effectiveness of enhancing COVID-19 case prediction models with mobility data, we will report the correlation improvement (ci) of the mobility-based models over non-mobility baselines. Given that the only difference between mobility-based models and non-mobility baselines is the mobility data, we posit that ci allows us to measure the effectiveness of using mobility data in COVID-19 predictive settings. Specifically, we compute ci as $ci = pcorr_Mobility - pcorr_Baseline$ with $pcorr_Mobility$ representing model performance (correlation) when mobility data is used and $pcorr_Baseline$ measuring the correlation when no mobility data is used in the COVID-19 case prediction model. Given that correlation values go between -1 and +1, ci will be within the $(-2, 2)$ range.

2.4.4 Adjusting Predictive Models

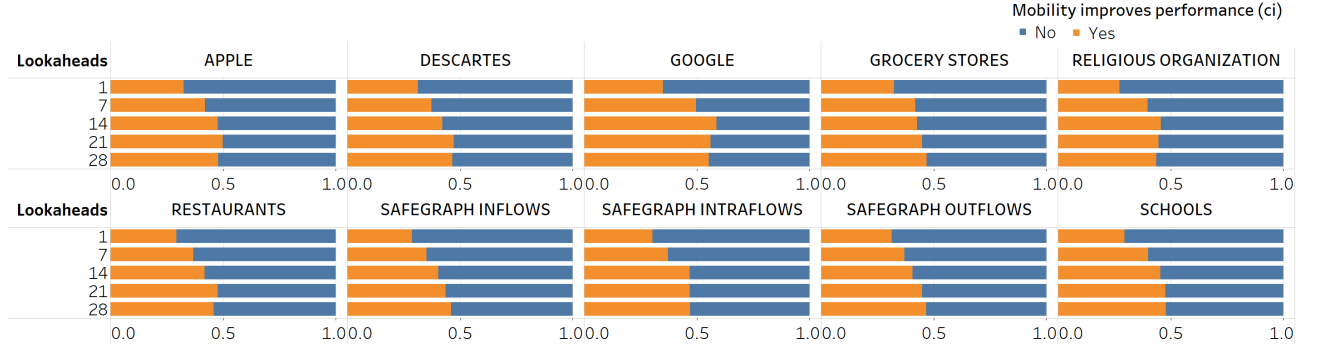
To identify the best linear regression implementation, we computed the average performance of ridge, lasso, elastic net, and OLS trained with and without mobility data for each lookahead and training approach (LTW and STW), and across all mobility datasets. We then selected the implementation with the majority of best performance values: Ridge for LTW and ElasticNet for STW. Table A.1 in the Appendix shows the detailed numbers. On the other hand, the optimal p , d and q values for the LTW and STW ARIMAX models were chosen based on a grid search and the minimum Akaike Information Criteria (AIC) value. The p , d and q values were selected across lookaheads, and the same values were used for mobility-based models and non-mobility baseline models for comparison analyses. Table A.8

in the Appendix shows a summary of the p , d and q values identified across all counties.

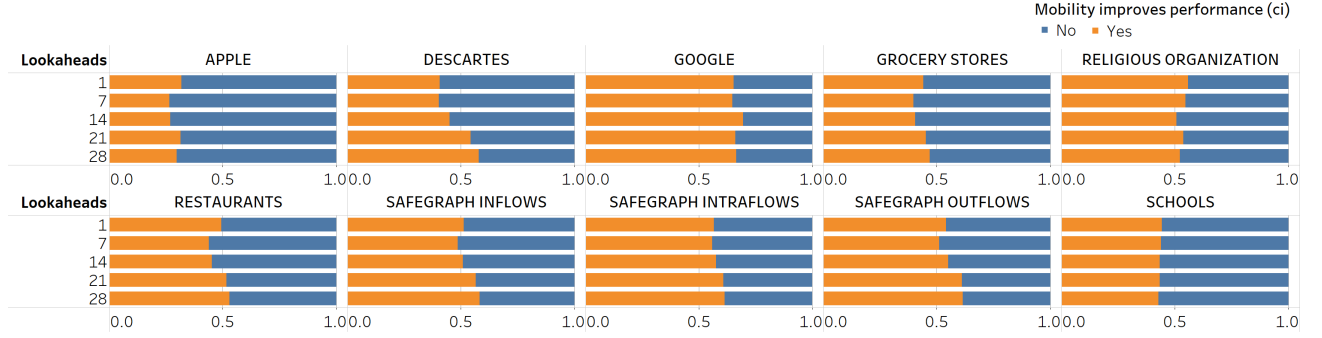
2.4.5 Evaluation approach

To analyze the conditions under which mobility data provides (or not) an enhancement over county-level COVID-19 case prediction models that do not use mobility as a source of information, we propose to carry out the following analyses (results are presented in section 2.5).

In section 5.1, we analyze the number of counties for which mobility data improved the individual prediction performance of non-mobility baselines, and quantify the improvements, with a focus on models trained with the long training window approach (LTW). In section 5.2, we delve into the demographic and socio-economic characteristics of these counties to assess if mobility data bias - whereby certain demographic groups are over- or under-represented in the data - might explain the differences in the performance of COVID-19 prediction models across counties. We also analyze whether the differences in the performance of mobility-based models over non-mobility baselines across counties vary depending on (1) the training approach: STW versus LTW (in section 5.3); (2) the mobility datasets used: ten different data sources (in section 5.4); and (3) the predictive model: linear regression versus ARIMAX (in section 5.5). As stated earlier, our main objective is to bring light into the use of mobility data in county-level COVID-19 case prediction models, to understand when it works (or not), and why, so as to inform decision makers assessing effectiveness-cost trade offs given that mobility data is not freely accessible.



(a) Results for ARIMAX COVID-19 case prediction model.



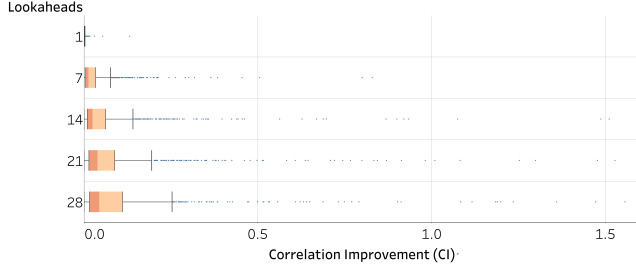
(b) Results for Ridge Regression COVID-19 case prediction model.

Figure 2.2: Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the LTW approach.

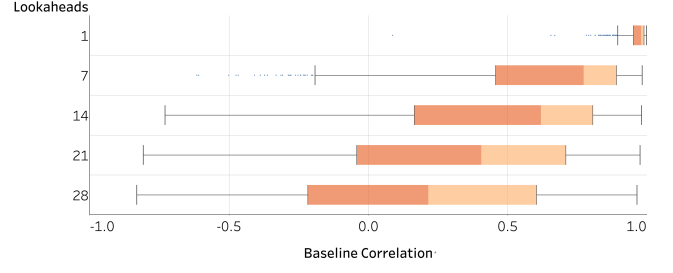
2.5 Results and Analysis

2.5.1 Does mobility data help, and by how much?

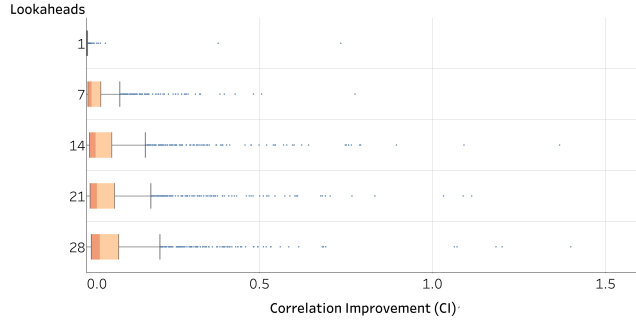
Figures 2.2a and 2.2b show the percentage of counties for which the prediction performance of mobility-based models improves over its corresponding non-mobility baselines for ARIMAX and ridge regression models, respectively, using the LTW approach (a comparison with STW is discussed in section 2.5.3). In other words, these plots show the percentage of counties for which the correlation improvement $ci > 0$. These plots show several important insights: (1) incorporating mobility data into county-based COVID-19 case prediction models helps in, at most, 60% of the counties analyzed,



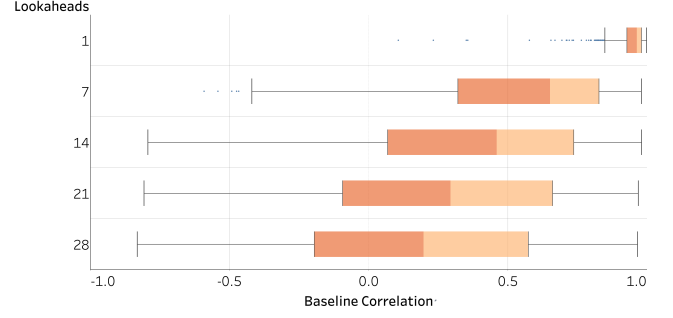
(a) CI for Apple Mobility



(b) Baseline Correlation for Apple Mobility



(c) CI for SafeGraph Inflows Mobility

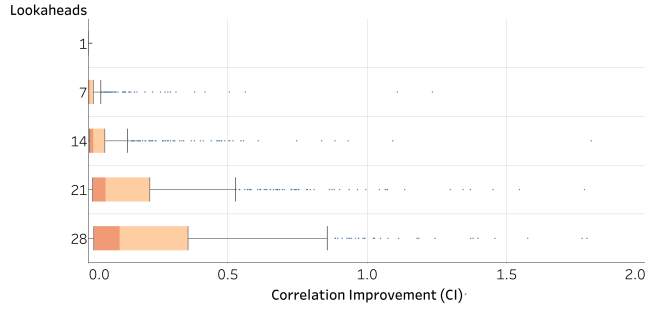


(d) Baseline Correlation for SafeGraph Inflows Mobility

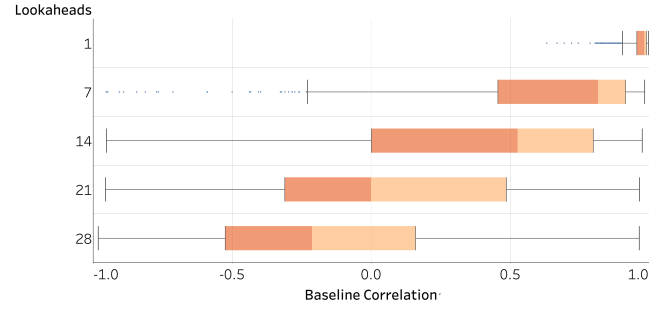
Figure 2.3: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with LTW approach using (a) Apple mobility data and (b) SafeGraph Inflow mobility data.

leaving at least 40% or more of the other counties with prediction performances lower than their non-mobility counterparts *i.e.*, mobility data is frequently hurting prediction performance; and (2) mobility data appears to help more in longer-term predictions (lookaheads 14, 21 and 28) than in shorter term predictions *i.e.*, in the short-term, COVID-19 statistics are generally informative enough and provide predictions that are more accurate than those when mobility data is added to the model; while for longer lookahead predictions, adding mobility data to predictive COVID-19 case models provides additional information that frequently improves the predictive accuracy of the non-mobility baseline models. This trend appears clear in the ARIMAX model; while for ridge, it is more apparent for the Descartes dataset and for SafeGraph in- and out-flow datasets.

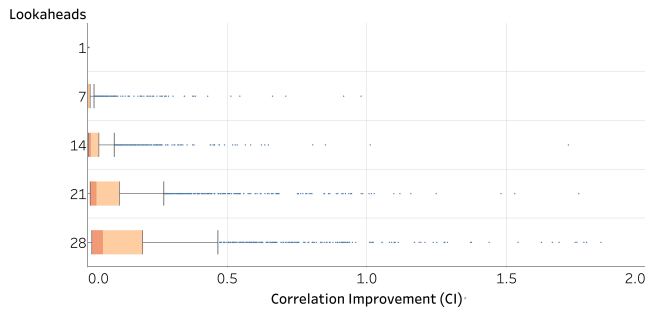
We have shown that adding mobility data to COVID-19 case prediction models improves their performance for - at most - 60% of the counties, across datasets, models (arimax and ridge) and



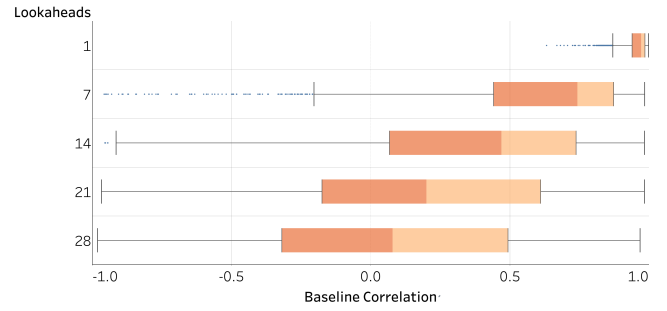
(a) CI for Apple Mobility



(b) Baseline Correlation for Apple Mobility



(c) CI for SafeGraph Inflows Mobility



(d) Baseline Correlation for SafeGraph Inflows Mobility

Figure 2.4: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ridge regression with LTW approach using (a,b) Apple Mobility data and (c,d) SafeGraph Inflow Mobility data.

lookaheads for the long training window approach (LTW). Next, we aim to quantify the performance improvement *i.e.*, does mobility data produce small or large prediction improvements when compared to non-mobility baselines?

Quantifying improvements. For this analysis, we look into two metrics per dataset and model: (1) the distribution of the county correlation coefficients ci for each lookahead, to explore the correlation improvements brought about by adding mobility data to the county prediction models - since we are assessing the improvement over non-mobility baselines, only counties with correlation improvements are considered; and (2) to contextualize these numeric improvements, we also look into the distribution of the correlations (between predicted and actual cases) for the non-mobility baseline models, and only for the counties that showed a correlation improvement so as to match the distributions in the first metric. By analyzing both the improvements brought about by adding mobility data, and by comparing these improvements with the actual baseline correlations, we are able to provide nuanced insights about when mobility data aids COVID-19 case prediction models. In this section, we focus on the LTW approach. Discussions about differences across training approaches, datasets and models will be covered in sections 5.3, 5.4 and 5.5. Figures 2.3 and 2.4 show an example of the distribution of the county correlation improvement (ci) and the distribution of the non-mobility baseline correlation values for ridge and ARIMAX, respectively, across lookaheads and using the LTW approach. For clarity purposes, both figures only show distributions for Apple (a) and SafeGraph Inflow (b) datasets. Plots for the remaining datasets can be found in the Appendix (see Figures A.1 and A.2).

The correlation improvements (ci) across datasets show that median ci values are between 0.0 – 0.1 *i.e.*, for counties where adding mobility data improves the prediction accuracy, it does so by a maximum of 0.1 for 50% of the counties, across models (ridge/arimax), datasets and lookaheads

(see Figures 2.3 and 2.4 as a sample trend, and Appendix Figures A.1 and A.2 for the full spectrum trend); with the largest correlation improvements associated with higher lookaheads. These are modest correlation improvements that might not change the strength of the non-mobility baseline correlation *e.g.*, a baseline moderate correlation of 0.5 will still be moderate after a 0.1 improvement [5]. Looking at upper quartile values (Q3), we observe slightly better correlation improvements in the 0.0 – 0.4 range, with the majority of Q3 values under 0.3, across models, datasets and lookaheads; revealing *ci* values that could improve correlation strengths. Maximum values (Q4) are in the 0 – 0.9 range across models, datasets and lookaheads, with the majority of maximum values under 0.5; and outliers can reach correlation improvement values of up to 1.9. These less frequent, and extreme, correlation improvements point to situations where negative baseline correlations - from the non-mobility models - are being changed to positive correlations, revealing counties where adding mobility helps reverse bad performing models to good ones, although these large improvements happen only for a handful of counties.

The correlation improvement plots also show that for higher lookaheads, the median value and the right skewness of the correlation improvement also increases across datasets and models. This finding shows that adding mobility data to COVID-19 prediction models produces better performing models over non-mobility baselines for longer term predictions, while short-term prediction models (next day) barely benefit from adding mobility data, with correlation improvement values close to zero. We posit that this might be due to the fact that non-mobility baselines achieve high correlations for lower lookaheads making it very hard to improve the baselines when adding mobility data. The baseline correlations plots in Figures 2.3 and 2.4 (as well as Appendix Figures A.1 and A.2) show that correlations for non-mobility baselines and lookahead 1 (next day prediction) have median values of at least 0.9; while, for higher lookaheads (21 and 28), we observe that non-mobility baselines have much

lower correlations, with median values in the range $(-0.21, 0.21)$.

In summary, these findings show that mobility data helps at most in 60% of the counties in the datasets analyzed, and that correlation improvements range from minimal for next day predictions, to small for higher lookaheads, with 50% of the counties showing modest correlation improvements of at most 0.1 and with 25% of the counties showing correlation improvements of at most 0.3. Larger outlier improvements of up to 1.9 in higher lookaheads are found for a handful of counties.

2.5.2 What types of counties benefit from mobility data? A mobility data bias analysis

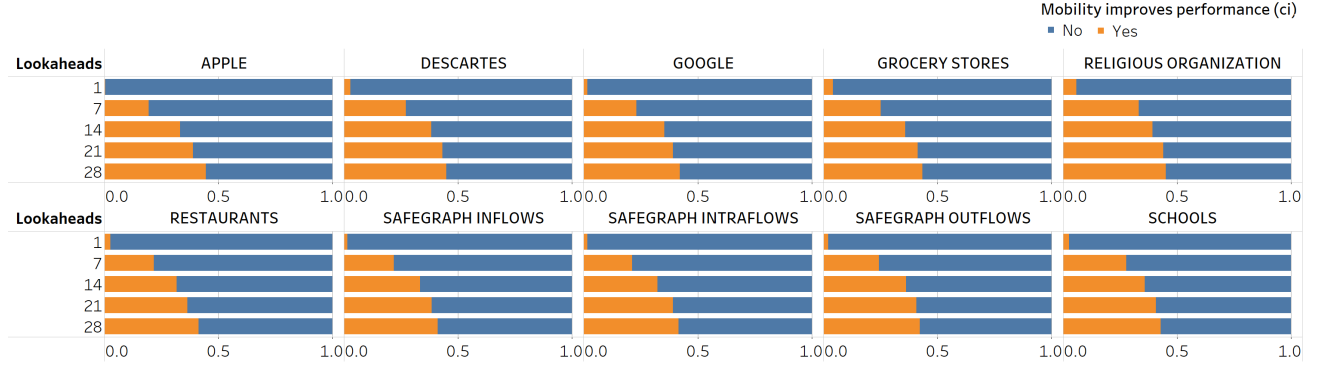
As stated in the Introduction, we hypothesize that a major reason why adding mobility data is not helping improve the accuracy of COVID-19 case prediction models in many more counties might be the bias present in the mobility data *i.e.*, counties with large percentages of racial minorities, elder or low-income population might be under-represented in the mobility datasets, thus negatively impacting the prediction performance [48, 154]. To assess that hypothesis, we carry out regression coefficient analysis for each mobility dataset and predictive model using demographic and socio-economic county characteristics as independent variables and the positive correlation improvements across lookaheads as the dependent variable. Our analysis will reveal the statistically significant associations of demographic and socio-economic features with correlation improvements, and the *direction* of that association (positive or negative).

Based on prior work on mobility data bias, we consider the following demographic and socio-economic county variables from the 2019 Census data: **age 65+**: percentage of county residents age 65 or greater; **income**: median household income for that county; **Black**: percentage of county residents

that identify as Black; **Hispanic**: percentage of county residents that identify as Hispanic; **race-Other**: percentage of county population that identifies as not White, Black, or Hispanic; and **urban-rural**: National Center for Health Statistics (NCHS) Urban-Rural Classification from 1 to 6, where 1 is *Large Central Metro* and 6 is *Non-core* (extreme rural). We also consider - as independent variables in the regression model - all possible paired interactions between the variables described by multiplying the values of each pair of variables, and by standardizing them. For interaction interpretability, we also ensure that higher values for each feature match with our hypothesis for lower performance. Thus, we change the directionality of the income by negating its values to match our hypothesis that lower income might be associated with worse coverage of mobility data and thus worse performance (all other features stay the same).

Finally, we would like to clarify that the COVID-19 case data itself might also suffer from different types of bias due to inaccurate data collection processes [154]. Nevertheless, since that bias affects both mobility-based and non-mobility baseline models, and since we are looking at the correlation improvement differences between the two, we can claim that any performance differences observed can be attributed to the mobility data, which is the only difference between the two models. Next, we discuss the demographic and socio-economic variables that were found to be statistically significantly related to correlation improvements for the ridge and ARIMAX models with the LTW training approach (results for the STW approach will be discussed in section 2.5.3). A detailed presentation of the coefficients and their significance can be checked in the Appendix, in Tables A.2 and A.4.

Findings. For the ridge model, we observe that the percentage of Black and Hispanic population, rurality as well as age above 65 are significantly and negatively related to correlation improvements across most datasets *i.e.*, counties with higher percentages of minority race and ethnicity, rural or elder populations benefit less from the addition of mobility data, with increases in these populations related



(a) Results for ARIMAX COVID-19 case prediction model.



(b) Results for Elastic Regression COVID-19 case prediction model.

Figure 2.5: Percentage of counties for which adding mobility data to the COVID-19 case prediction model improves the prediction performance ($ci \geq 0$). Results are for the STW approach.

to lower performance improvements over the non-mobility baselines. We posit that this is probably due to potential sampling bias in the mobility datasets, with under-representation of race, ethnicity, old age and rurality in mobility datasets pointing to worse overall predictive performance. Lower income, on the other hand, was significantly, positively associated with higher correlation improvements when mobility was added to individual county-level COVID-19 case prediction models. In principle, this result was counter-intuitive since we were expecting that lower incomes would be associated with lower access to smart phones. Nevertheless, looking at the interaction terms between low income and age 65+ as well as low income and race/ethnicity (Black and Hispanic) we observe significant negative coefficients, pointing to the fact that counties with higher percentages of low income black and hispanic groups as well as low income elder population are associated with lower correlation

improvements possibly due to these groups not being well represented in the datasets.

For the ARIMAX model, the trends were not pervasive across mobility datasets, with specific coefficients observed for different datasets. Age 65+ and rurality were negatively related to correlation improvements for the Apple and some SafeGraph datasets, pointing to the fact that elder and rural populations might not be fairly represented in the datasets, and confirming prior work on SafeGraph bias analysis [48]. Apple and SafeGraph also show a significant negative relation between race-*Other* (non Black, White or Hispanic) and correlation improvements potentially pointing to the fact that other minority races/ethnicities might not be as represented in the mobility datasets as Whites, Blacks and Hispanics. There were two instances with counter-intuitive results. First, in the Descartes dataset, race-*Other* was - unexpectedly - significantly positively associated with correlation improvements. However, when considering the interaction between race-*Other* and rurality the significant coefficients were negative *i.e.*, counties with high percentages of other minority races in rural settings are negatively related to correlation improvements and thus potentially not fairly represented in the Descartes dataset. Second, lower income had a significant positive coefficient for Apple *i.e.*, counties with lower incomes were associated with higher correlation improvements. Nevertheless, the interaction term between low income and race-*Other* for the Apple dataset, has a significant negative coefficient, revealing worse correlation improvements for counties with high percentages of minority races and low income, potentially due to sampling bias. It is important to clarify that, despite the significance of many coefficients, the R-square values for the regressions were low, pointing to the fact that there exist other behavioral or pandemic features that could also explain correlation improvements such as masking mandates, masking behaviors or transmission rates, among others. Nevertheless, statistics for these features at the county level and for our period of study were not accessible.

lookaheads	p-value	avg. % benefit STW	avg. % benefit LTW
1	0.00002	0.16473	0.40074
7	0.00004	0.31053	0.43324
14	0.08103	0.41427	0.46833
21	0.20845	0.4722	0.49895
28	0.3104	0.48988	0.50206

Table 2.2: Mann-Whitney U test between the STW and LTW distributions of the percentage of counties that benefit from adding mobility data across the ten datasets and two models for each training approach.

2.5.3 Do correlation improvements change across training approaches?

In this section, we explore whether the improvements brought about by adding mobility data to county-based COVID-19 case prediction models are different depending on the type of training approach. Specifically, we discuss the differences in percentages of counties that benefit from adding mobility data, quantify correlation improvements and discuss the impact of mobility data bias on county-level COVID-19 case prediction models that have been trained with STW as opposed to LTW approaches. Our main objective is to understand if using short-term training windows - which requires considerably less data and hence reduces data costs - has an impact on how COVID-19 case prediction models benefit from mobility data.

Percentage of Counties. Figure 2.5 shows the percentage of counties whose STW-trained COVID-19 case prediction performance improves when adding mobility data. Compared to Figure 2.2, we can observe that - for lower lookaheads (1, 7 and 14) - the percentage of counties that benefit from adding mobility data to STW-trained models is smaller than the percentage for LTW-trained models, while the percentage of counties remains similar for higher lookaheads (21 and 28) across both linear and ARIMAX models. To assess the statistical significance of these observations, we run a Mann-Whitney U test between the STW percentages and the LTW percentages for each lookahead,

lookaheads	p-value	avg. median ci STW	avg. median ci LTW
1	0.00334	0.00183	0.00029
7	0.00001	0.01693	0.00738
14	0	0.0321	0.01674
21	0.00009	0.04836	0.02939
28	0.01143	0.06145	0.04425

Table 2.3: Mann-Whitney U test per lookahead between the STW and LTW distributions of the correlation improvements across the ten datasets and two models for each training approach.

and across the ten datasets and two predictive models (see Table 2.2 for details). The distributions were found to be statistically significantly different ($p - value < 0.05$) for lookaheads 1 and 7, with STW-trained models having significantly smaller median percentages of counties benefiting from adding mobility data (from 16% – 31%) than LTW-trained models (from 40% – 43%).

Correlation Improvements. To quantify correlation improvement differences between STW- and LTW-trained models, we run a Mann-Whitney U test for each lookahead between the correlation improvement (ci) distributions for STW and LTW across the ten mobility datasets and two predictive models *i.e.*, we measure whether the improvements brought about by adding mobility data to COVID-19 case prediction models are statistically significantly different across training approaches, and by how much (distribution plots can be checked in the Appendix, Figures A.1, A.2, A.3 and A.4). The tests - shown in Table 2.3 - reveal that the two distributions are statistically significantly different ($p - value < 0.05$) across all lookaheads, with lower median correlation improvement values for LTW-trained models than their STW-trained counterpart (0.0002 – 0.0442 vs. 0.0018 – 0.0614); and with a slightly lower maximum median ci value of 0.1, as opposed to STW-trained maximum $ci = 0.13$. In other words, although a higher percentage of LTW-trained counties improved their performance when adding mobility data, the median improvement range is slightly lower than their STW counterpart.

Mobility data bias. We repeat the regression coefficient analysis discussed in section 5.2 for

the STW training approach; and a comparison with the associations revealed for the LTW training in that section revealed similar findings. For regression models, similarly to LTW age 65+, income, race Black, race-*Other* and rurality are all negatively related to correlation improvement *i.e.*, counties with higher percentages of elder, rural, low income or minority populations do not benefit as much from the use of mobility data, possibly due to lack of representativity of these groups in the mobility datasets. Nevertheless, the number of significant coefficients across datasets for STW is lower than for LTW - although all datasets except for Google show one or another type of bias. This result highlights that mobility data bias might more constrained when shorter amounts of training data are used in the model.

For ARIMAX models, and similarly to LTW, we found that age 65+, ethnicity Hispanic and race-*Other* are negatively associated with correlation improvements *i.e.*, adding mobility data in counties with elder people and other minority race/ethnicity worsens the predictive performance when compared to non-mobility baselines, potentially revealing bias in the data collected for these groups. On the other hand, although low income and Black race are positively related to correlation improvements; interaction terms between these two factors and age 65+ showed that counties with high percentage of elder black population or elder low-income population were significantly negatively related to correlation improvement *i.e.*, the mobility data associated with these groups might not be representative thus affecting the quality of the predictions. Finally, the interaction between rurality and age 65+ partially showed that counties with higher percentages of elder rural communities did not always benefit from using SafeGraph mobility data. For further details, we encourage the reader to compare Table A.4 with Table A.5 in the Appendix (for regression models) and Tables A.2 and A.3 for ARIMAX models.

These analyses show that the training-testing approach impacts the number of counties that benefit from adding mobility data and creates a trade-off: LTW improves performance for more counties

Dataset	Avg. Percentage of counties improved	Avg. Median ci improvement
Apple Mobility	0.38107	0.0337
Descartes Mobility	0.42670	0.02363
Google Mobility	0.53613	0.01293
(SafeGraph) Grocery Store Mobility	0.42302	0.01816
(SafeGraph) Religious Org Mobility	0.44899	0.02298
(SafeGraph) Restaurants Mobility	0.42864	0.02375
(SafeGraph) Schools Mobility	0.43075	0.02123
SafeGraph Inflow Mobility	0.42365	0.02751
SafeGraph Intraflow Mobility	0.46907	0.02720
SafeGraph Outflow Mobility	0.43595	0.02938

Table 2.4: Improvement statistics per mobility dataset across training approaches and predictive models. The percentage of counties column represents the average percentage of counties whose COVID-19 case prediction models benefit from adding mobility data; while the correlation improvement (ci) column quantifies the average median *ci* improvement over the non-mobility baseline.

albeit with smaller correlation improvements with respect to their baselines and with more bias across socio-economic and demographic variables when compared to STW, which requires less data and constitutes a more affordable approach.

2.5.4 Do correlation improvements change across mobility datasets?

In this section, we explore whether the improvements brought about by adding mobility data to county-based COVID-19 case prediction models are different depending on the type of dataset. As in previous sections, we will analyze differences across percentage of counties that benefit from adding mobility data to their predictive models; differences across correlation improvements; and evaluate the role of mobility data bias on those differences. Table 2.4 summarizes average improvements per mobility dataset across lookaheads, training approaches and predictive models for (1) percentage of counties and (2) correlation improvements. To evaluate statistically significant differences across datasets, we compute the Friedman test, a non-parametric test that evaluates whether median values across datasets are statistically significantly different.

Percentage of Counties. We run the Friedman test with the distribution of the percentage of

counties that benefit from adding mobility data across the five lookaheads, two models and two training approaches for each mobility dataset. The test was rejected ($\chi^2 = 28$ and $p - value = 0.0009$) thus pointing to statistically significant differences between the average percentages across datasets. The Friedman test does not identify which distribution(s) are different; however, Table 2.4 shows that the percentage of counties with improvements is considerably higher for the Google dataset when compared to others, possibly due to the smaller number of counties used in the analysis. Thus, we removed the Google dataset from our set of distributions and repeated the Friedman test, which did not find any statistically significant difference between datasets *i.e.*, the percentage of counties that benefit from adding mobility data to COVID-19 prediction models does not significantly change depending on the mobility dataset except for Google (see Table A.6 in the Appendix for further test details).

Correlation Improvements. We run the Friedman test to evaluate whether the correlation improvement brought about by different mobility datasets was statistically significantly different across datasets. The Friedman test between the median correlation improvements across models, training approaches and lookaheads for each dataset rejected the null hypothesis ($\chi^2 = 182.7$, $p - value = 0$) thus pointing to statistically significant differences across datasets. Removal of individual datasets did not change the result of the tests (null hypotheses rejected) revealing that correlation improvements are significantly different across mobility datasets with Apple and SafeGraph - which includes O-D flows - having the highest improvements across training approaches and predictive models. Nevertheless, it is important to highlight that despite significant, the median correlation improvements across datasets were very modest with a maximum value of 0.033 (see Table A.7 in the Appendix for further test details).

Mobility Data Bias. Finally, in terms of bias, all mobility datasets were associated with bias in race, age 65+ and income and rurality either as independent features or via feature interactions; with

lookaheads	p-value	average % ARIMAX	average % linear regression
1	0.0002	0.16672	0.39875
7	0.00512	0.32085	0.42292
14	0.00069	0.40169	0.48091
21	0.00004	0.4376	0.53354
28	0.00004	0.45166	0.54029

Table 2.5: Mann-Whitney U test between the ARIMAX and linear distributions of the percentage of counties that benefit from adding mobility data across the ten datasets and two training approaches for each predictive model.

the exception of the Google dataset and SafeGraph Grocery Stores that were associated only to race, age and income bias, but rurality did not appear to play a role in the correlation improvements (see Tables [A.2](#), [A.3](#), [A.4](#) and [A.5](#) in Appendix for further details).

To summarize the analysis presented in this section, Apple and SafeGraph datasets appear to bring about very modest statistically significantly higher correlation improvements (maximum value of 0.033), but the percentage of counties and the bias identified are similar across mobility datasets.

2.5.5 Do correlation improvements change across predictive models?

In this section, we explore whether the improvements brought about adding mobility data to county-based COVID-19 case prediction models are different depending on the type of predictive model. Revisiting Figures [2.2](#) and [2.5](#), we can visually observe that the percentage of counties that benefit from adding mobility data to county-level COVID-19 prediction models is larger when the models are linear regressions than ARIMAX. To assess the statistical significance of this difference, we run a Mann-Whitney U test for each lookahead, between the percentage of counties that benefited from adding mobility data to linear models and the percentage of counties that benefited from adding mobility data to ARIMAX models across the ten datasets and two training approaches. The test showed

lookaheads	p-value	average median ci ARIMAX	average median ci regression
1	0.00007	0.00196	0.00029
7	0.94738	0.01544	0.0147
14	0.00298	0.02564	0.03478
21	0.00003	0.03196	0.0612
28	0.00001	0.0381	0.08195

Table 2.6: Mann-Whitney U test per lookahead between the ARIMAX and linear distributions of the correlation improvements across the ten datasets and two training approaches for each model type.

that the differences between the two types of models are statistically significantly different across all five lookaheads with the percentage of counties that benefit from adding mobility data being smaller for ARIMAX models (with values between 0.16% and 0.45%) than for linear models (0.39%–0.54%).

Table 2.5 shows further details of the statistical test.

On the other hand, to assess the statistical significance of the differences in the correlation improvements (*ci*) between linear and ARIMAX models, we run a Mann-Whitney U test for each lookahead, between the ci values associated with adding mobility data to linear models and the ci values associated with adding mobility data to ARIMAX models across the ten datasets and two training approaches. Except for lookahead 7, all other lookaheads show significantly different ci distributions, with larger improvements associated with linear regression models at higher lookaheads (0.00029 – 0.08195 for linear vs. 0.00196 – 0.0381 for ARIMAX). Full test details are available in Table 2.6.

Finally, in terms of bias differences across models, we can observe that both models suffer from similar types of bias across datasets and training approaches *i.e.*, income, race, age 65+ and (a little bit less frequently) rurality. Nevertheless, there is one main difference between linear regression and ARIMAX models. Linear models trained with Google mobility data (both with STW and LTW training approaches) did not identify any significant socio-economic or demographic features in the bias anal-

ysis, meaning that these features do not play a role in the performance improvement of county-level COVID-19 case linear predictions. It is important to clarify that although some bias was identified for ARIMAX trained with Google mobility data, the number of features identified was also considerably lower than for any other dataset. We posit that this could be potentially related to the smaller number of counties available (990, see Table 2.1). Detailed bias numbers can be found in the Appendix, in Tables A.2 and A.3 (arimax) and Tables A.4 and A.5 (linear regressions). Overall, linear regressions appear to be a better choice, with a larger number of counties benefiting from adding mobility data, with larger correlation improvements at higher lookaheads and with similar bias.

2.6 Discussion

Our analysis on the use of mobility data for COVID-19 case prediction has shown the heterogeneity and limitations of the benefits of mobility data inclusion. As discussed in Section 2.5, at most 60% of the counties improve their performance when adding mobility data to the prediction model (top value for linear regressions with the LTW approach). The median correlation improvements across lookaheads, datasets and models are minimal for next day predictions, and modest for higher lookaheads, with 50% of the counties showing correlation improvements of at most 0.1 and 25% of the counties showing correlation improvements of at most 0.3.

Our results have shown that, even in the best case, 40% of counties would face no improvement when mobility data is added to their COVID-19 case prediction models, making the question of whether or not mobility data would be helpful to county-level decision makers essentially a coin flip. While companies made mobility data freely available at the initial height of the COVID-19 pandemic, companies have already begun to discontinue these practices (Apple’s reports are no longer available,

and Google has stopped releasing data as of October 15th, 2022). As decision makers face the continued spread of COVID-19 and potential future diseases, they must consider whether it will be worth purchasing mobility data. Our analysis shows purchasing county-level mobility data will not benefit many counties, and decision makers should proceed with caution accordingly.

It is also concerning that the extent to which mobility data improves predictions is in part a function of the composition of the population in the county. In fact, across most of the mobility datasets, correlation improvements were lower for counties with higher Black, Hispanic, and other non-White population as well as low-income and rural populations. As older and minority-race patients have been disproportionately affected by the COVID-19 pandemic, we would hope to provide more and better resources to these groups to ameliorate the disparities. Instead, we see that mobility data could serve to entrench these disparities, providing decision makers in counties with more vulnerable populations with worse-performing models, leading to worse-informed policy decisions.

Our analyses have also demonstrated that linear regression models outperform ARIMAX models in both: (1) the percentage of counties that benefit from adding mobility data, and (2) the correlation improvement when using the same mobility data. These differences are statistically significant. Based on this, we recommend decision makers favor linear regression models, when interested in using interpretable models. We have also shown that the performance across datasets is very similar, with Apple and some SafeGraph datasets having slightly superior correlation improvements, albeit still so small that in many cases will not lead to a change in the strength of the correlation. This result highlights that decision makers looking to use mobility data as a source of behavioral information should not worry about the dataset they gain access to, since all seem to similarly improve the correlation over their non-mobility baselines. Finally, we have discussed how the training-testing approach presents a trade off for decision makers, with LTW approaches increasing the number of counties that benefit

from adding mobility data; while STW approaches - that require considerably less data and reduce costs - increase the correlation improvements (albeit with small values) and have lower bias.

Chapter 3: Auditing the Fairness of COVID-19 Forecast Hub’s Case Prediction Models

3.1 Introduction

3.1.1 The US COVID-19 Forecast Hub: Development and Context

The US COVID-19 Forecast Hub was founded in 2020 and serves as a “*central repository of COVID-19 forecasts from over 50 independent research groups*” [34]. Participant research groups submit county, state and national US COVID-19 forecasts with a standardized format; and the Forecast Hub provides an interactive visualization tool to help decision makers and the general public analyze weekly predictions for COVID-19 hospitalizations, cases and deaths. The standardized predictions collected from all research groups, as well as the predictions for an ensemble model that brings all individual predictions together, are also shared with the Centers for Disease Control and Prevention (CDC) who uses these results for their official COVID-19 communications [33].

Over the past four years, numerous research groups from both academia and industry have focused on developing models to forecast COVID-19 cases, hospitalizations and deaths in the United States. The COVID-19 Forecast Hub [51] has been instrumental in collating these efforts. These models vary in approach, ranging from deep learning methods [10, 111, 119, 185] to compartmental models [138, 187], statistical models [47, 76], or combinations of these via ensemble models [4, 51].

COVID-19 forecast models are usually trained with historical data (e.g., past cases or hospitalizations) together with other contextual information such as human mobility data. Human mobility data has been used in the past to model and characterize human behaviors in the built environment [70, 93, 148, 164, 170], to support decision making for socio-economic development [71, 72, 73, 75, 95], for public safety [171, 172], as well as during epidemics and disasters [23, 81, 94, 96, 102, 169]. During the COVID-19 pandemic, human mobility has also played a central role in driving decision making, and more than 50% of the Forecast Hub models have incorporated mobility data into their prediction models, acknowledging the impact of human movement on virus propagation [1, 10, 13, 60, 111, 119].

3.1.2 Fairness in COVID-19 Prediction Models: A Critical Gap

The US COVID-19 Forecast Hub has been, and continues to be, a critical centralized resource to promote transparent decision making. While the Forecast Hub has made significant contributions through its accuracy-focused predictions at different spatial granularities (*e.g.*, county or state), there is an opportunity to expand its evaluation framework to examine how prediction performance varies across social determinants like race, ethnicity and urbanization levels that have been shown to play an important role in COVID-19, including race, ethnicity and rurality [87, 158].

The pandemic has highlighted existing disparities in healthcare, with significant differences in COVID-19 infection rates, hospital admissions, and deaths among different racial and ethnic groups as well as across the urban-rural spectrum [158, 162]. These disparities risk being perpetuated in model predictions if not adequately addressed. Diverse prediction performance across social determinants - for example, higher prediction errors for a given minority race or ethnicity - could negatively impact resource allocation and intervention decisions *e.g.*, hospital beds or stay-at-home orders, given that

the CDC appears to be using the Forecast Hub predictions for official communications that subsequently inform policy decisions [33]. Given the urgent need for rapid pandemic response modeling, initial Forecast Hub efforts necessarily focused on developing accurate predictions. As these models continue to inform CDC communications and policy decisions, incorporating fairness analyses could further enhance their utility for equitable resource allocation and intervention planning across diverse communities [142].

There are many reasons why the COVID-19 prediction performance can be different across social determinants such as race, ethnicity or urbanization levels. The Forecast Hub's COVID-19 prediction models are trained on datasets containing COVID-19 statistics for hospitalizations, cases or deaths. Given the unprecedented scale and urgency of the pandemic, data collection faced several challenges [57, 87]. For example, a lack of consistency in reporting race and ethnicity across jurisdictions, has generated a lot of missing racial data. That data is often excluded due to its incompleteness, potentially affecting the actual total hospitalizations, cases or deaths for minority race and ethnicity groups which might be less reported. In addition, there are occasions where the race is reported by the medical staff instead of being self-reported, which is the most accurate source and prevents errors [25]. For example, the CDC reports that the latest research on race and Hispanic origin misclassification on COVID-19 death certificates shows that deaths are underreported by 33% for non-Hispanic American Indian or Alaska Natives, by 3% for non-Hispanic Asian or Pacific Islanders, and by 3% for Hispanic decedents [36]. Testing availability and access varied across communities, with some minority groups experiencing more limited access, such as Latino communities [55], thus affecting the accuracy of the overall COVID-19 statistics, with under-reporting bias perpetuating the invisibility of racial and ethnic minorities in general COVID-19 statistics. A similar effect has been observed in rural counties and states, with rural areas associated to lower testing rates, thus disproportionately detecting fewer cases

of COVID-19 in these regions [158].

To exacerbate this situation even more, COVID-19 prediction performance across social determinants can also be affected by additional datasets used in the training of some of the COVID-19 prediction models. Specifically, around 50% of the Forecast Hub’s models use human mobility data from Safegraph [150], Apple [8] or Google [85] among others, to complement COVID-19 predictions (see Fig 3.2a). Human mobility data can characterize origin-destination trips, visits to specific points of interest (POI), or the volumes of different types of trips (*e.g.*, car vs. public transit). Research has shown that mobility data can improve the prediction accuracy of COVID-19 cases, deaths and hospitalizations [1, 78, 101]. Nevertheless, researchers have also identified that mobility data suffers from sampling bias across race and age groups [60] with, for example, elder and Black communities being less represented [48]. Similarly to the COVID-19 case under-reporting bias, mobility data sampling bias could also affect the fairness of COVID-19 predictions across social groups.

In this paper, we propose - to the best of our knowledge - the first thorough fairness analysis of the COVID-19 prediction models in the Forecast Hub. Specifically, we focus on COVID-19 case prediction models at the county level, since these are closer to local realities and allow for more actionable decision making than state-level predictions. We use error parity as a measure of group fairness [89] *i.e.*, lack of fairness in our context is associated with significantly different error distributions across two social determinants: race or ethnicity and urbanization level. Accurately revealing differences across racial and ethnic groups would require access to county-level COVID-19 case data stratified by race or ethnicity, which would allow us to compare predicted versus actual case county statistics for each racial and ethnic group. Due to the complexity and scope of pandemic data collection efforts, many counties in the US faced significant challenges in collecting comprehensive demographic data [104]. Hence, to be able to carry out a fairness analysis of the Forecast Hub’s COVID-19 predic-

tion models, we propose a regression analysis to evaluate the associations between prediction errors in a given county and the race and ethnicity distributions for that county, while controlling for underlying health conditions and age groups. A similar regression analysis is proposed for the urbanization levels.

Additionally, to support researchers in the Forecast Hub, we also investigate how group fairness metrics for race, ethnicity and urbanicity levels change across model characteristics such as model type (*e.g.*, deep learning versus statistical), training data (*e.g.*, *with or without mobility data*), lookaheads (*e.g.*, predicting cases for next week versus in four weeks) or pandemic phases. Finally, we also describe a dashboard that we have designed to allow decision makers and researchers explore *fairness nutritional cards* for each Forecast Hub model [161]. To sum up, the main contributions of this paper are:

- We present a thorough fairness analysis of the CDC Forecast Hub’s COVID-19 county case prediction models across race, ethnicity and urbanization levels. Our research shows statistically significant differences in predictive errors with some minority racial and ethnic groups as well as less urbanized areas associated with significantly higher errors than the majority White race, while controlling for underlying health conditions, age groups and state.
- We carry out interaction analyses identifying differences in performance across racial/ethnic groups and urbanicity levels with respect to COVID-19 prediction type models, COVID-19 datasets, prediction lookaheads and COVID-19 phases, while controlling for underlying health conditions and age groups. Our results show significant prediction performance differences for certain minority groups and less urbanized areas, when compartmental or statistical models are used. On the other hand, short-term forecasting and certain pandemic phases with higher case volumes are also associated with higher prediction performance differences for certain minority

racial and ethnic groups as well as for less urbanized areas.

- We present a dashboard where researchers and decision makers at the CDC and beyond will be able to explore fairness nutritional cards per individual model across race, ethnicity and urbanization level, and how fairness might vary across model and data characteristics.

3.2 Materials and Methods

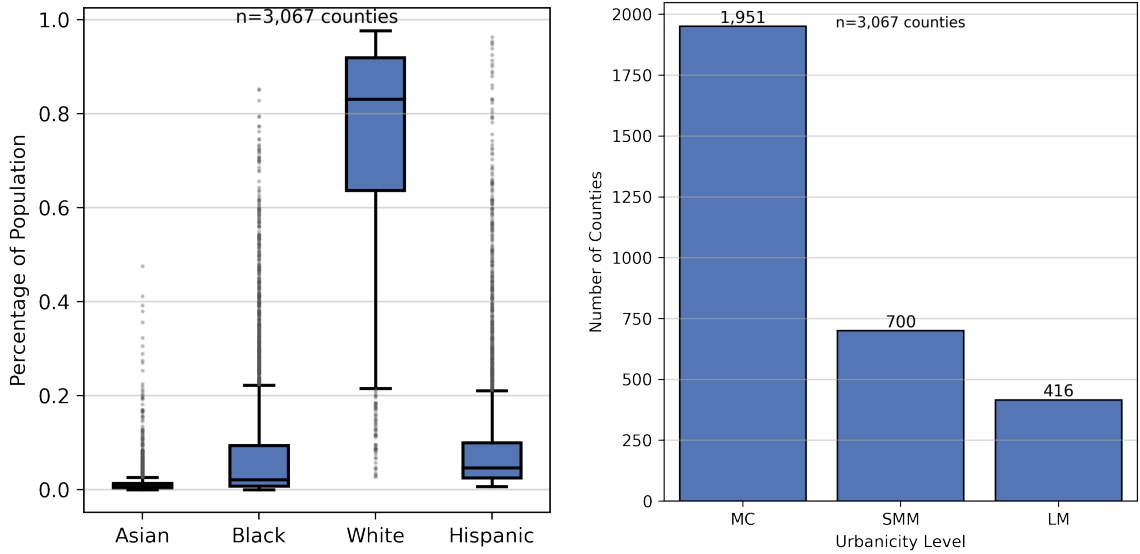
3.2.1 Data Sources and Variables

3.2.1.1 COVID-19 Forecast Data

For the purpose of our study, we focus exclusively on the weekly, county-level COVID-19 case predictions publicly available from the COVID-19 Forecast Hub across all US counties [34]. We focus on county-level forecasts because these are closer to local realities and allow for more actionable decision-making than state-level predictions. On the other hand, since hospitalizations and deaths are only available at the state level [51], we focus on COVID-19 case predictions. The weekly incidence predictions in the Forecast Hub are uploaded by participating teams and defined as the newly anticipated COVID-19 cases per county within the following *epidemiological week*, extending from Sunday to Saturday. We use the weekly forecasts during the period from July 2020 to October 2022.

The hub’s data repository offers both point forecasts and quantile-based probabilistic forecasts. Our study employs the latter, leveraging the seven provided quantiles ([0.025, 0.100, 0.250, 0.500, 0.750, 0.900, 0.975]) to gain insights into the uncertainty ranges and confidence intervals posited by the forecasting models. From the entire cohort of models and teams contributing to the Forecast Hub, we selected 36 teams that met our inclusion criteria: they provided comprehensive quantile forecasts

throughout our period of analysis and they submitted predictions at the county-level. A Gantt chart depicting the specific quantile forecasts used to evaluate each model is shown in Figure B.1.



(a) Distribution of Racial Demographics Across US Counties (b) Distribution of Counties by Urbanicity Level

Figure 3.1: Distribution of the sensitive attributes (Race/Ethnicity and Urbanization level) across the 3,067 counties in US considered for this study.

3.2.1.2 Sensitive Attributes

To empirically evaluate fairness in COVID-19 forecasts, we measure error parity across two sensitive demographic attributes: (1) RACE AND ETHNICITY population composition, and (2) county-level URBANIZATION LEVEL classification. For this analysis, we use two primary data sources: the American Community Survey (ACS) [39] for demographic information and the CDC urbanization classification for county-level urban-rural designations [37]. In our analysis, we focus on Asian, Black, Hispanic, and White racial/ethnic groups while using White as the reference category (see Fig 3.1a for the racial and ethnic distribution across the 3,067 counties considered for this study). Other racial categories, including American Indian/Alaska Native (AIAN), Native Hawaiian/Pacific Islander (NHPI),

and multi-racial groups, were excluded from our primary analysis due to limited variable distribution. As Table 3.1 shows, 97.69% of counties have NHPI populations under 1%, and 79.03% have AIAN populations under 1%, providing insufficient variation for meaningful regression analysis. In contrast, between 1.21%-66.22% of the counties have less than 1% of Asian, Black, or Hispanic population, providing adequate statistical power for detecting potential effects in our regression analysis. Furthermore, Asian populations represent a key demographic group in metropolitan and suburban areas where COVID-19 impacts were particularly pronounced, making their inclusion crucial for understanding prediction accuracy in these important contexts. Further details about this selection, and exclusion of some counties can be found in the Section B.1.1

Table 3.1: County-Level Demographic Distribution Statistics

Racial/Ethnic Group	Median (%)	Mean (%)	Counties > 1%	% Counties > 1%
White	83.09	75.64	3,067/3,067	100.00
Hispanic	4.64	9.88	3,030/3,067	98.79
Black	2.14	8.88	2,019/3,067	65.83
Asian	0.70	1.53	1,036/3,067	33.78
American Indian/Alaska Native	0.40	2.00	643/3,067	20.97
Native Hawaiian/Pacific Islander	0.12	0.24	71/3,067	2.31

Notes: Statistics are computed across all 3,067 U.S. counties included in our analysis. "Counties > 1%" shows the number of counties where the group comprises more than 1% of the population.

On the other hand, the CDC urban-rural classification scheme classifies counties into six urbanization levels[37], from highly urban (1) to rural (6). For this paper, we group them into three labels: Large Metropolitan areas (**LM**, which correspond to codes 1 and 2), Small and Medium Metropolitan (**SMM**, codes 3 and 4) and Micropolitan and Non-core areas (**MC**, codes 5 and 6). This grouping ensures sufficient sample sizes and variations within each category for robust statistical analysis, given the uneven distribution of US counties across urbanization levels (see Fig 3.1b). At the same time the three-level classification provides a clearer narrative about urban-rural disparities while maintaining meaningful distinctions in population density and healthcare infrastructure.

3.2.1.3 Model-Data Characteristics

Given our interest in understanding how group fairness metrics for race, ethnicity and urbanization might change across model and data characteristics, we break down the prediction performance and fairness analyses across four aspects:

- **Model Type:** Based on information reported in the papers associated to each of the 36 predictive models, we have manually classified them into five categories, namely: Statistical, Compartmental, Deep Learning, Baseline, and Ensemble (see Fig 3.2a for model statistics). This classification aims to discern the potential influence of model typologies on forecast performance and to identify any systematic biases inherent to specific modeling approaches (Section B.1.2 and Table B.1 provide more details about the taxonomy).

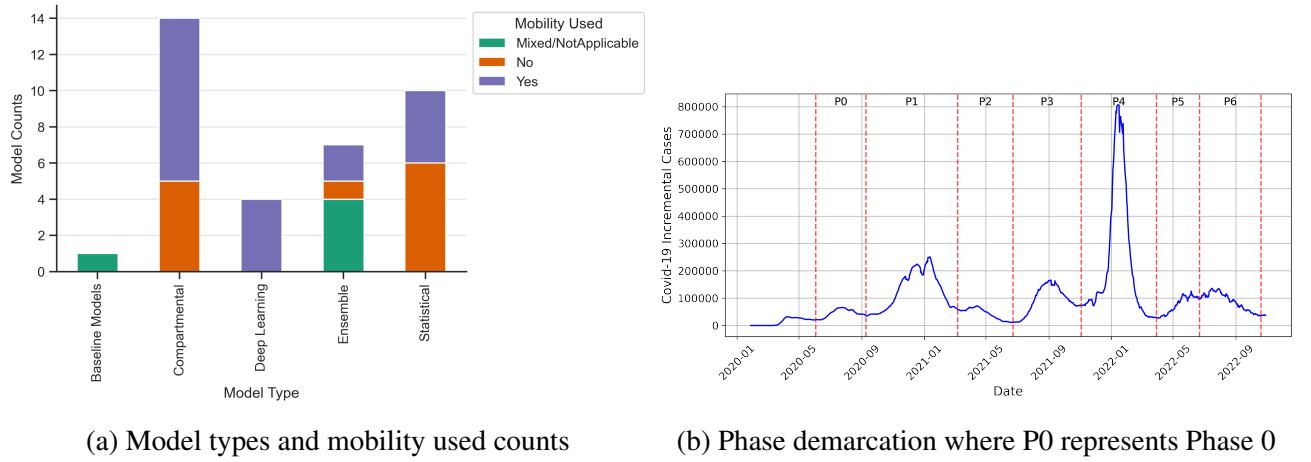


Figure 3.2: Prediction fairness will be evaluated across types of models, training datasets (mobility), lookaheads and phases.

- **Mobility Used:** Based on information reported in the papers for each predictive model analyzed, we distinguish between models that integrate mobility data and those that do not. The main objective of this feature is to allow us to explore the effect of mobility data on forecast accuracy

and the possible introduction of biases resulting from this additional data dimension. Fig 3.2a displays a categorization of the forecasting models, differentiated by both the type of model and the incorporation of mobility data, as quantified by their respective counts. *Mixed models* are COVIDHub Ensemble models (like COVIDhub CDC-ensemble) that combine models that use (or not) mobility data in their training.

- **Lookahead:** We use forecasts ranging from 1 week to 4 weeks (a.k.a lookaheads), allowing us to evaluate differences in the predictive accuracy of the hub’s models across race, ethnicity and the rural-urban spectrum for short and medium-term horizons: at 7, 14, 21 and 28 days.
- **Phase:** COVID-19 case numbers differ a lot across pandemic stages. To understand whether phases play a role in model fairness across race, ethnicity and urbanization level, we divide the period under study into seven phases, identified based on the presence of valleys and peaks in the volume of COVID-19 cases (see Fig 3.2b). A more detailed explanation of the phase identification process can be found in Section B.1.3.

3.2.1.4 Control Variables

Our analysis controls for potential confounding factors by incorporating county-level demographic and health variables in our analyses. Specifically, we include the proportion of residents aged 65 and above (sourced from ACS [39]) to account for age-related COVID-19 risk variations. To control for underlying health conditions known to affect COVID-19 susceptibility and severity we also utilize age-adjusted prevalence of nine key comorbidities [38]: high blood pressure, cancer, diabetes, obesity, stroke, chronic obstructive pulmonary disease (COPD), Chronic kidney disease (CKD), current asthma, and coronary heart disease (CHD) extracted from the CDC PLACES [35] dataset. These

specific health conditions were selected based on extensive epidemiological evidence linking them to increased COVID-19 severity and mortality [180].

3.2.2 Analytical Approach

To systematically evaluate fairness across race, ethnicity, and urbanization levels in COVID-19 case predictions from the Forecast Hub, we developed a three-step analytical approach that combines error calculation, fairness assessment across sensitive attributes, and interaction analysis with model-data characteristics, while controlling for relevant demographic and health factors.

Step One: Calculating Pinball Loss. We focus on county error parity as a measure of fairness. Hence, we first need to compute the weekly forecast error at the county level for all the counties in the US. To evaluate the accuracy of COVID-19 case forecasts, and given that forecasts in the Forecast Hub are uploaded by teams as quantiles, we employ the pinball loss metric (PBL). This metric quantifies the error of a probabilistic forecast by measuring the distance between observed values and the predicted quantiles, penalizing over- or under-estimation asymmetrically to reflect the actual cost of errors in the prediction. The pinball loss $L_\tau(y, f)$ is represented as: $(\tau - 1) \cdot (y - f)$, if $y < f$ and as $\tau \cdot (f - y)$, if $y \geq f$, where $L_\tau(y, f)$ denotes a county's pinball loss for a given quantile τ , y is the observed value i.e., reported number of cases or ground truth extracted from the JHU CSSE COVID-19 case counts dataset [56], and f is the forecasted value at quantile τ . For our analysis, we use the average county PBL, computed across the set of 7 quantiles: $\overline{PBL}_\tau = \frac{1}{7} \sum_{i=1}^7 L_\tau(y_i, f_i)$, and normalized by the county population.

Step Two: Fairness across Sensitive Attributes Once weekly average PBLs have been computed per county, we aim to evaluate associations between county prediction errors and the two pro-

tected attributes, RACE AND ETHNICITY and URBANIZATION LEVEL , using regressions. Coefficient analyses will enable us to identify cases in which error parity is violated, pointing to a lack of prediction fairness i.e., significantly different errors across racial or ethnic groups or across urbanization levels.

Associating counties, and their PBL errors, to URBANIZATION LEVEL is straight forward using the CDC urban-rural classification scheme [37] that associates a county with a given urbanization level. On the other hand, associating counties, and their PBL prediction errors, with race and ethnicity would require access to race-stratified predictions. However, due to systemic data collection challenges during the pandemic, the race-stratified COVID-19 case data necessary to compute race-stratified prediction errors were not collected, hence limiting the predictions provided by the Forecast Hub to county level predictions across all racial and ethnic groups [104]. Next, we describe how we proceed in the evaluation of the relationship between forecast errors and sensitive attributes race, ethnicity and urbanization level.

Race and Ethnicity Analysis. To evaluate whether there exist differences between PBL error distributions across racial and ethnic groups, we conduct a regression analysis where county PBL errors are regressed against the racial and ethnic percentage distributions within each county (with White race as the reference group). In the regression, we control for underlying health conditions, percentage of the population aged 65 and above, the state and data-model characteristics (i.e., type of predictive model, mobility data, lookahead and phase). An analysis of the resulting regression coefficients can provide insights into how different racial compositions might be positively or negatively associated with forecast accuracy, potentially identifying unfairly higher errors for certain racial or ethnic groups. The main effects regression model, **Model-1**, can be formally defined as:

$$PBL_{c,t} = \alpha + \sum_i \beta_i \cdot Race_{i,c} + \sum_{j,k} \gamma_{j,k} \cdot DM_{j,k,c,t} + \sum_l \sigma_l \cdot HO_{l,c} + \theta \cdot age65_c + \sum_s \lambda_s \cdot State_{s,c} + \epsilon_{c,t} \quad (3.1)$$

where:

- $PBL_{c,t}$: The dependent variable representing prediction error for county c at time t .
- $\alpha, \epsilon_{c,t}$: The intercept term and the error terms for county c at time t , respectively.
- $\beta_{Asian}, \beta_{Black}, \beta_{Hispanic}$: Relationship between each race category (Asian, Black, and Hispanic) and PBL , with White as the reference category.
- γ_{jk} : Relationship between each model-data characteristic j with corresponding category k and the PBL , with $j \in \{\text{Lookahead, Phase, Model Type, Mobility Used}\}$. For example, for the characteristic Lookahead, $k \in (14, 21, 28)$ (7-days reference group). Reference groups for each model-data characteristic are: 7-days, phase 0, compartmental model and no mobility data used.
- σ_l : Relationship between age-adjusted health outcomes (HO) and the PBL , with $l \in \{\text{BPHIGH, CANCER, DIABETES, OBESITY, STROKE, COPD, KIDNEY, CASTHMA, CHD}\}$. For example, σ_{Asthma} represents the effect of asthma prevalence in a given county on the PBL .
- θ : Relationship between the percentage of population aged 65 and older in a given county and the PBL .
- λ_s : Individual state s fixed effect on the error distribution, with $s \in \{\text{List of States in US}\}$.

Urbanization Analysis. For the URBANIZATION LEVEL we replace the Race percentage variables in Equation 3.1 with the three urbanization levels and we refer to this as **Model-2**.

$$PBL_{c,t} = \alpha + \sum_i \beta_i \cdot Urb_{i,c} + \sum_{jk} \gamma_{jk} \cdot DM_{j,k,c,t} + \sum_l \sigma_l \cdot HO_{l,c} + \theta \cdot age65_c + \sum_s \lambda_s \cdot State_{s,c} + \epsilon_{c,t} \quad (3.2)$$

where, $Urb_{i,c}$ where, $Urb_{i,c}$ represents the urbanization level i for county c , with $i \in SMM, MC$, with LM being the reference category.

Step Three: Interaction with Model-Data Characteristics Building upon the main effects models, we are also interested in looking into whether the fairness metrics across sensitive attributes change when model and/or data characteristics are considered. In other words, we evaluate if the relationship between county forecast errors and their racial/ethnic and urbanization groups changes when model and data characteristics are taken into account. Next, we describe the methodological approach for each sensitive attribute.

Race and Ethnicity Analysis. To evaluate whether the relationship between PBL error distributions and race/ethnicity changes across model or data characteristics, we regress the county PBL errors against racial and ethnic percentages for that county (with the White group as a baseline) while adding interaction terms between race/ethnicity and each model-data characteristic, and while controlling for age-adjusted Health Outcome prevalence, the percentage of individuals in the county who are 65+, and the county state. Since we are interested in four model-data characteristics, namely model type, use of mobility data, lookahead and phase, we create for separate regressions (**Model-1a, -1b, -1c and -1d**, respectively) to account for both main and interaction effects for each model-data characteristic:

$$\begin{aligned}
PBL_{c,t} = & \alpha + \sum_i \beta_i \cdot Race_{i,c} + \sum_{jk} \gamma_{jk} \cdot DM_{j,k,c,t} + \sum_{i,k} \delta_{ijk} \cdot (Race_{i,c} \times DM_{j,k,c,t}) \\
& + \sum_l \sigma_l \cdot HO_{l,c} + \theta \cdot age_65_c + \sum_s \lambda_s \cdot State_{s,c} + \epsilon_{c,t}
\end{aligned} \tag{3}$$

where δ_{ijk} represents the relationship between the PBL and changes in a given race percentage i and model-data characteristic category j, k with respect to their reference groups (e.g., $\delta_{Hispanic, Lookahead, 14-days}$ helps us evaluate how the PBL changes for a 1% increase in $i = Hispanic$ county population with respect to the White reference group, with $j = Lookahead$ and $k = 14 - days$ with respect to reference category 7-days). All other terms remain the same as Equation 3.1. To be able to evaluate significant changes in the error distribution (PBL) between a minority race and the White reference group for a given model-data characteristic category j, k , we will examine the relative effect computed as $\beta_i + \delta_{ijk}$.

Urbanization Analysis. Similarly to the race and ethnicity analysis, we construct four regression models (**Model-2a, -2b, -2c and -2d**) that mirror the race/ethnicity interaction models, but replace racial percentages with urbanization categories. Using Large Metropolitan (LM) areas as the reference group, we examine interactions between urbanization levels (SMM and MC) and each model-data characteristic. The regression equation takes the form:

$$\begin{aligned}
PBL_{c,t} = & \alpha + \sum_i \beta_i \cdot Urb_{i,c} + \sum_{jk} \gamma_{jk} \cdot DM_{j,k,c,t} + \sum_l \sigma_l \cdot HO_{l,c} \\
& + \sum_{i,k} \delta_{ijk} \cdot (Urb_{i,c} \times DM_{j,k,c,t}) + \theta \cdot age_65_c + \sum_s \lambda_s \cdot State_{s,c} + \epsilon_{c,t}
\end{aligned} \tag{4}$$

where $Urb_{i,c}$ represents the urbanization level i for county c , and the interaction term δ_{ijk} captures the relationship between a given urbanization level and model-data characteristic category j, k with respect to their reference groups, on the PBL. We will examine the relative effect of a given urbanization level i on the PBL (with respect to LM) and for a given model-data characteristic category j, k computing the relative effect as $\beta_i + \delta_{ijk}$.

Table 3.2 provides a comprehensive overview of our regression framework, detailing the dependent and independent variables for each model. The table systematically presents our main effects models (Model-1 and Model-2) and their corresponding interaction models, along with the control variables and model specifications used throughout our analysis.

Table 3.2: Summary of Generalized Linear Models and Their Specifications

Model	Sensitive Attributes	Interaction Terms	Control Variables
<i>Main Effects Models</i>			
GLM-1	Race/Ethnicity (% Black, Hispanic, Asian) % White (ref.)	None	Health Outcomes Data-Model Characteristics State Fixed Effects
GLM-2	Urbanicity (SMM, MC), LM (ref.)	None	Health Outcomes Data-Model Characteristics State Fixed Effects
<i>Main + Interaction Effects: Race Interaction Models</i>			
GLM-1a	Race/Ethnicity	Race $\times$ Lookahead	Same as GLM-1
GLM-1b	Race/Ethnicity	Race $\times$ Phase	Same as GLM-1
GLM-1c	Race/Ethnicity	Race $\times$ Model Type	Same as GLM-1
GLM-1d	Race/Ethnicity	Race $\times$ Mobility Used	Same as GLM-1
<i>Main + Interaction Effects: Urbanicity Interaction Models</i>			
GLM-2a	Urbanicity	Urbanicity $\times$ Lookahead	Same as GLM-2
GLM-2b	Urbanicity	Urbanicity $\times$ Phase	Same as GLM-2
GLM-2c	Urbanicity	Urbanicity $\times$ Model Type	Same as GLM-2
GLM-2d	Urbanicity	Urbanicity $\times$ Mobility Used	Same as GLM-2

Distribution Family: Gaussian with log link function for all models

Health Outcomes: Asthma, Obesity, COPD, CHD, CKD, Diabetes, Obesity, Cancer, Stroke, Age 65+

Data-Model Characteristics: Lookaheads (7(reference) , 14, 21, 28), Phase (0 (reference) 1-6), Model Type (Compartmental (reference), Baseline, Deep Learning, Ensemble, Statistical), Mobility Usage (No (reference), Yes, Mixed)

Dependent Variable: Square root of PBL (sqrt_pbl) with 1% trimming

Model Assumptions: Independent observations, exponential family distribution (Gaussian), linear predictor through log link function, constant variance of residuals on link scale, no perfect multicollinearity (adjusted GVIF < 2)

3.2.3 Model Selection and Evaluation

Initial exploratory analysis of the Pinball Loss norm (PBL) distribution revealed several characteristics that influenced our model selection. The original distribution exhibited extreme right-skewness with a heavy tail, through quantile analysis (see Section B.2.1 and Figure B.2. These extreme values, while valid measurements, can exert excessive leverage in regression models and potentially obscure patterns in the majority of the data. Therefore, we made the methodological choice to trim the top 1% values, preserving 99% of our observations while substantially improving the stability of the model. We also consider squared root transformation during model evaluation. Given these distributional characteristics - which violate basic linear regression assumptions- and the strictly positive nature of our dependent variable, we tested both Gamma and Gaussian GLM families with log-link and identity-link functions.

Initial analysis revealed substantial multicollinearity among independent variables in the GLM models, particularly between health variables and racial demographics. To address this, we employed an iterative variable selection process using the Generalized Variance Inflation Factor (GVIF¹) [68]. Variables were retained only if their adjusted GVIF values (calculated as $(GVIF^{\frac{1}{2 \times Df}})$, where Df represents degrees of freedom) fell below a threshold of 2. Importantly, we proceeded with our analysis only after confirming that our variables of interest - race, ethnicity, urbanization level and data-model characteristics - all demonstrated adjusted GVIF values below this threshold, ensuring the reliability of our primary coefficient estimates.

Using diagnostic plots, residual patterns, and model performance metrics (pseudo- R^2 values) for model evaluation, we identified the Gaussian GLM with log-link function applied to square root

¹<https://rdrr.io/cran/glmttoolbox/man/gvif.html>

transformed data (with 1% trimming) as the best fit. Multicollinearity analyses and model diagnostics for these GLMs are discussed in depth in Section B.2.2, Section B.2.3.

In the next section, we evaluate the relationship of race, ethnicity and urbanization levels with the forecast error (PBL) distribution via Model-1 and Model-2, which we rename to GLM-1 and GLM-2. We also evaluate how that relationship is modulated when interaction terms between race, ethnicity, urbanization levels and the four model and data characteristics are incorporated into the model i.e., Model-1a through Model-1d and Model-2a through Model-2d that we rename as GLM-1a through GLM-1d and GLM-2a through GLM-2d (see Table 3.2 for model summary). Following the Results section, we will present hypotheses for the reasons behind these findings in the Discussion section.

3.3 Results

3.3.1 Fairness of COVID-19 case predictions across race and ethnicity (GLM-1)

Based on Equation 3.1, the generalized linear model (GLM-1) characteristics are illustrated in Table 3.3. The table provides insights into the relationship between race, ethnicity and the PBL while controlling for the impact of various data-model features, health outcomes and state fixed effects on prediction accuracy. As a note, some of the health outcome covariates were left out due to inflated GVIF.

For the race and ethnicity variables, the regression results indicate that for every 1% increase in a county's Hispanic population (with respect to the reference group White) the prediction errors increased by approximately $(1.216 - 1) * 100\% = 21.6\%$ ($\beta = 0.196$, $\exp(\beta) = 1.216$, $p < 0.001$). In contrast, regions with larger Asian populations showed markedly lower prediction errors than the White population baseline ($\beta = -0.663$, $\exp(\beta) = 0.515$, $p < 0.001$). Our results indicate that for

Table 3.3: Summary of Regression Results: GLM-1

Variable	e Coef. (Std. Error)	95% CI	z-value
Intercept	0.009*** (0.009)	[0.009, 0.009]	-510.099
<i>Sensitive Attributes</i>			
% Black	0.976*** (0.003)	[0.970, 0.983]	-6.864
% Hispanic	1.216*** (0.003)	[1.209, 1.223]	63.844
% Asian	0.515*** (0.015)	[0.500, 0.530]	-44.647
<i>Lookahead Period</i>			
14 days	1.119*** (0.001)	[1.118, 1.121]	143.586
21 days	1.211*** (0.001)	[1.209, 1.213]	249.396
28 days	1.300*** (0.001)	[1.298, 1.302]	352.668
<i>Phase Effects</i>			
Phase 1	1.445*** (0.001)	[1.443, 1.448]	377.204
Phase 2	0.890*** (0.001)	[0.888, 0.892]	-99.535
Phase 3	1.285*** (0.001)	[1.283, 1.288]	236.425
Phase 4	1.553*** (0.001)	[1.549, 1.556]	416.701
Phase 5	0.969*** (0.001)	[0.967, 0.972]	-22.748
Phase 6	1.152*** (0.001)	[1.149, 1.154]	112.373
<i>Model Type</i>			
Baseline	1.121*** (0.002)	[1.118, 1.124]	72.937
Deep Learning	1.037*** (0.001)	[1.035, 1.039]	40.193
Ensemble	1.007*** (0.001)	[1.005, 1.009]	7.959
Statistical	1.076*** (0.001)	[1.074, 1.078]	89.646
<i>Mobility</i>			
Mixed	0.864*** (0.001)	[0.862, 0.866]	-124.255
Yes	0.996*** (0.001)	[0.994, 0.997]	-5.637
<i>Health Controls</i>			
Asthma	0.959*** (0.001)	[0.957, 0.961]	-43.776
Obesity	1.003*** (0.000)	[1.003, 1.004]	25.942
COPD	1.046*** (0.000)	[1.045, 1.047]	93.094
% Age 65+	1.214*** (0.007)	[1.198, 1.230]	28.712

Model Statistics: Pseudo R^2 (Cox-Snell) = 0.460; Log-Likelihood = 5,981,789; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* sqrt_pbl. *Link Function:* log. *Regression Family:* Gaussian. State fixed effects included but are reported in Table 5 in [Auditing the Fairness of COVID-19 Forecast Hub's Case Prediction Models](#). $\exp(\text{Coef})$ represents the multiplicative effect on the outcome. CI: Confidence Interval.

every 1% increase in the Asian population (with respect to White population), the prediction accuracy is approximately 48.5% better when compared to the White group. For predominantly Black communities, we observed a slight but statistically significant improvement in prediction accuracy ($\beta = -0.024$, $\exp(\beta) = 0.976$, $p < 0.001$), with 1% increase in Black population (with respect to White) being associated with predictions approximately 2.4% more accurate than the baseline.

Model diagnostics. Model assessments indicate that Gaussian-GLM demonstrates adequate fit and reliability. The pseudo- R^2 is 0.46, and the residuals show an approximately normal distribution with mild heteroskedasticity at extreme values. Cook’s distance analysis identified no influential points that would substantially affect our findings. Although there are some deviations from normality in the tails of the residual distribution, these do not materially affect our main conclusions regarding demographic disparities in prediction accuracy (detailed diagnostics are provided in Section 2.3.1, Fig 3 in [Auditing the Fairness of COVID-19 Forecast Hub’s Case Prediction Models](#)).

Summary: Race

COVID-19 forecast errors show substantial racial disparities, particularly for **Hispanic** Race:

⇒ 21.6% *higher* for Hispanic populations

⇒ 48.5% *lower* for Asian populations

⇒ 2.4% *lower* for Black population

3.3.2 Fairness of COVID-19 case predictions across urbanization levels (GLM-2)

Table 3.4 illustrates the estimated coefficients of the generalized linear model for the fairness analysis of COVID-19 prediction accuracy across urbanization levels (GLM-2, Equation 3.2). The table reveals insights into the relationship between urbanization levels and the PBL while controlling

for the impact of various data-model features, health outcomes and state fixed effects on prediction accuracy.

Table 3.4: Summary of Regression Results: GLM-2

Variable	e ^{Coef.} (Std. Error)	95% CI	z-value
Intercept	0.009*** (0.018)	[0.009, 0.009]	-263.017
<i>Sensitive Attributes</i>			
Micropolitan	1.065*** (0.001)	[1.063, 1.067]	60.326
Small and Medium Metro	1.027*** (0.001)	[1.025, 1.029]	26.509
<i>Lookahead Period</i>			
14 days	1.119*** (0.001)	[1.117, 1.121]	143.691
21 days	1.211*** (0.001)	[1.209, 1.213]	249.533
28 days	1.299*** (0.001)	[1.298, 1.301]	352.841
<i>Phase Effects</i>			
Phase 1	1.445*** (0.001)	[1.442, 1.448]	377.540
Phase 2	0.890*** (0.001)	[0.888, 0.892]	-100.050
Phase 3	1.285*** (0.001)	[1.282, 1.287]	236.167
Phase 4	1.552*** (0.001)	[1.549, 1.555]	416.754
Phase 5	0.969*** (0.001)	[0.967, 0.972]	-22.827
Phase 6	1.151*** (0.001)	[1.148, 1.154]	112.311
<i>Model Type</i>			
Baseline	1.121*** (0.002)	[1.118, 1.124]	73.043
Deep Learning	1.037*** (0.001)	[1.035, 1.039]	40.217
Ensemble	1.007*** (0.001)	[1.005, 1.009]	7.952
Statistical	1.076*** (0.001)	[1.074, 1.078]	89.919
<i>Mobility</i>			
Mixed	0.864*** (0.001)	[0.862, 0.866]	-124.509
Yes	0.996*** (0.001)	[0.994, 0.997]	-5.631
<i>Health Controls</i>			
Asthma	0.956*** (0.001)	[0.954, 0.958]	-34.853
Obesity	1.004*** (0.000)	[1.004, 1.005]	32.018
BPHigh	0.998*** (0.000)	[0.997, 0.998]	-7.667
COPD	1.000 (0.001)	[0.998, 1.002]	-0.089
Stroke	0.981*** (0.003)	[0.974, 0.988]	-5.605
Cancer	0.984*** (0.003)	[0.979, 0.990]	-5.650
CHD	1.089*** (0.003)	[1.084, 1.095]	33.418
Diabetes	0.995*** (0.001)	[0.993, 0.997]	-5.442
CKD	1.043*** (0.006)	[1.030, 1.056]	6.559
% Age 65+	1.111*** (0.007)	[1.096, 1.126]	15.328
Model Statistics: Pseudo R^2 (CS) = 0.462; Log-Likelihood = 5,983,292; N = 1,526,869			
Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. <i>Dependent Variable:</i> sqrt_pbl. <i>Link Function:</i> log. <i>Regression Family:</i> Gaussian. State fixed effects included but are reported in Table B.6. <i>exp(Coef)</i> represents the multiplicative effect on the outcome. CI: Confidence Interval.			

Relative to LM areas, prediction errors varied significantly across urbanization levels. MC areas, which are urban clusters with populations up to 50,000, showed notably higher prediction errors ($\beta = 0.063$, $\exp(\beta) = 1.065$, $p < 0.001$), indicating approximately 6.5% higher errors in these regions. This suggests particular challenges in forecasting COVID-19 cases in smaller urban areas that may have different healthcare infrastructure and reporting systems compared to LM areas. SMM areas also demonstrated increased prediction errors ($\beta = 0.027$, $\exp(\beta) = 1.027$, $p < 0.001$), though the effect

was more modest with 2.7% higher errors than large metropolitan regions.

Model diagnostics. Our model assessments reveal that GLM-2 achieves similar statistical properties to GLM-1, with a pseudo- R^2 of 0.46. While the residual distribution shows some deviation from normality at the tails, these departures do not substantially impact our key findings regarding urbanicity-based disparities in prediction accuracy. The complete characteristics of the model and the GVIF analysis can be found in Section [B.2.3.2](#), Figure [B.4](#).

Summary: Urbanicity

⇒ Prediction disparities worsen for more rural areas.

3.3.3 Fairness of COVID-19 case predictions across race, ethnicity and model-data characteristics

3.3.3.1 GLM-1a: Race, Ethnicity and Forecast Lookahead

Table [3.5](#) presents the interaction effects between race, ethnicity and lookahead periods, revealing notable variations in forecast accuracy across different time horizons, and pointing to unequal (unfair) error distributions (see Equation [3](#)). The trends described for the main effects (GLM-1) persist when adding the interaction effects *i.e.*, increases in Black or Asian population in a county (with respect to White) and for a given lookahead, are associated with lower PBLs; while increases in Hispanic population are associated with higher PBLs.

Our analysis reveals that increases in Hispanic population (with respect to White) are related to persistent disparities in forecast accuracy across all prediction horizons, though these disparities show a gradual decrease for predictions over longer timeframes. When examining the relative effects, we find that areas with higher Hispanic populations experience prediction errors that are 24.6%

Table 3.5: **GLM-1a**: Race $\times$ Lookahead Effects Relative to White Reference Group

Variable	Coefficient Estimates		Relative Effect	
	e ^{Coef.} (SE)	95% CI	e ^{Coef.}	% Diff from White
% Asian (7-day ref.)	0.228*** (0.030)	[0.215, 0.243]	0.228***	-77.2%
× 14-day ahead	1.652*** (0.037)	[1.536, 1.777]	0.377***	-62.3%
× 21-day ahead	2.689*** (0.035)	[2.509, 2.882]	0.613***	-38.7%
× 28-day ahead	3.505*** (0.034)	[3.279, 3.746]	0.799***	-20.1%
% Black (7-day ref.)	0.937*** (0.005)	[0.928, 0.947]	0.937***	-6.3%
× 14-day ahead	1.034*** (0.006)	[1.023, 1.046]	0.969***	-3.1%
× 21-day ahead	1.053*** (0.005)	[1.042, 1.065]	0.987***	-1.3%
× 28-day ahead	1.067*** (0.005)	[1.056, 1.078]	1.000	0.0%
% Hispanic (7-day ref.)	1.298*** (0.005)	[1.287, 1.310]	1.298***	+29.8%
× 14-day ahead	0.960*** (0.005)	[0.951, 0.970]	1.246***	+24.6%
× 21-day ahead	0.913*** (0.005)	[0.904, 0.923]	1.185***	+18.5%
× 28-day ahead	0.900*** (0.005)	[0.891, 0.909]	1.168***	+16.8%

Model Statistics: Pseudo R^2 (CS) = 0.461; Log-Likelihood = 5,982,892; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-1a. We only discuss race, ethnicity and lookahead interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Tables 7, 8 in [Auditing the Fairness of COVID-19 Forecast Hub's Case Prediction Models](#). Model Diagnostics are provided in Fig 5a in [Auditing the Fairness of COVID-19 Forecast Hub's Case Prediction Models](#). For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular race or ethnicity compared to the White population within each lookahead. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 3. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to White population for each lookahead variable (*% Diff from White*). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given race/ethnicity and lookahead value, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to the White population (e.g., +24.6% means that the PBL error for Hispanic counties at 14-day lookahead is 24.6% higher PBL when compared to White). All effects should be interpreted as the relative difference compared to White population within each specific lookahead value. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

higher than predominantly White areas at 14-day forecasts, declining to 18.5% at 21-day forecasts, and further reducing to 16.8% at 28-day forecasts. This pattern indicates that although the disparity in prediction accuracy moderates somewhat over longer forecast horizons, significant inequities in model performance persist throughout all prediction timeframes for Hispanic communities.

For Black counties, the relative effect shows progressively decreasing significant differences in forecast errors between Black and the White reference group, with PBL errors being 3.1% higher for

White counties for the 14-day forecast ($\exp(\beta) = 0.969$, $p < 0.001$), decreasing to 1.3% higher for 21 days ($\exp(\beta) = 0.987$, $p < 0.001$). By the 28-day forecast horizon, the prediction accuracy for Black counties shows no significant difference from White areas (relative effect: 1.000, 0.0% difference).

Asian counties reveal the most pronounced variation across forecast horizons. The relative effect analysis shows that while predictions remain more accurate than for White counties across all horizons, this advantage decreases substantially over longer forecast periods. The relative effect shows that areas with higher Asian populations have prediction errors 62.3% lower than White areas at 14-day forecasts, improving to 38.7% lower at 21-day forecasts, and 20.1% lower at 28-day forecasts.

3.3.3.2 GLM-1b: Race, Ethnicity and COVID-19 Phase

When examining how racial and ethnic disparities vary across different pandemic phases, we observe substantial heterogeneity in PBL, with patterns varying markedly by both race and phase (see Table 3.6). The results reveal complex temporal dynamics in prediction fairness across different demographic groups.

For Hispanic populations, the analysis reveals significant variation in prediction errors across pandemic phases relative to White populations. At initial stages of the pandemic (Phase 0), Hispanic counties show substantially higher errors (+101.8%) compared to White counties. While this disparity persists across most phases, its magnitude fluctuates notably. Phase 1 also exhibits a higher disparity (+36.6%), followed by Phase 3 (+15.6%) and Phase 6 (+12.0%). However, in Phases 4 and 5, this pattern reverses, with Hispanic counties showing slightly lower errors than White counties (-8.2% and -0.3% respectively), suggesting that prediction fairness for Hispanic communities varied significantly with pandemic phase characteristics.

Table 3.6: **GLM-1b**: Race $\times$ Phase Effects Relative to White Reference Group

Variable	Coefficient Estimates		Relative Effect	
	e ^{Coef.}	(SE) 95% CI	e ^{Coef.}	% Diff from White
% Asian (Phase 0 ref.)	0.137***	(0.042) [0.126, 0.149]	0.137***	-86.3%
× Phase 1	2.092***	(0.046) [1.912, 2.290]	0.287***	-71.3%
× Phase 2	8.767***	(0.049) [7.957, 9.659]	1.201***	+20.1%
× Phase 3	0.812***	(0.053) [0.732, 0.900]	0.111***	-88.9%
× Phase 4	6.310***	(0.047) [5.756, 6.918]	0.864***	-13.6%
× Phase 5	35.577***	(0.049) [32.323, 39.159]	4.871***	+387.1%
× Phase 6	3.260***	(0.057) [2.916, 3.645]	0.447***	-55.3%
% Black (Phase 0 ref.)	2.130***	(0.005) [2.108, 2.153]	2.130***	113.0%
× Phase 1	0.438***	(0.006) [0.434, 0.443]	0.933***	-6.7%
× Phase 2	0.393***	(0.007) [0.388, 0.399]	0.837***	-16.3%
× Phase 3	0.541***	(0.006) [0.535, 0.548]	1.152***	+15.2%
× Phase 4	0.270***	(0.007) [0.267, 0.274]	0.575***	-42.5%
× Phase 5	0.375***	(0.009) [0.369, 0.382]	0.799***	-20.1%
× Phase 6	0.529***	(0.008) [0.521, 0.536]	1.127***	+12.7%
% Hispanic (Phase 0 ref.)	2.018***	(0.005) [1.998, 2.039]	2.018***	101.8%
× Phase 1	0.677***	(0.006) [0.670, 0.685]	1.366***	+36.6%
× Phase 2	0.549***	(0.007) [0.541, 0.557]	1.108***	+10.8%
× Phase 3	0.573***	(0.006) [0.566, 0.580]	1.156***	+15.6%
× Phase 4	0.455***	(0.006) [0.449, 0.460]	0.918***	-8.2%
× Phase 5	0.494***	(0.009) [0.485, 0.503]	0.997	-0.3%
× Phase 6	0.555***	(0.008) [0.547, 0.564]	1.120***	+12.0%

Model Statistics: Pseudo R^2 (CS) = 0.495; Log-Likelihood = 6,012,390; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-1b. We only discuss race, ethnicity and phase interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Tables B.9 and B.10. Model Diagnostics are provided in Figure B.5. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular race or ethnicity compared to the White population within each phase. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 3. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to White population for each phase variable (% Diff from White). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given race/ethnicity and phase value, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to the White population (e.g., +101.8% means that the PBL error for Hispanic counties in Phase 0 is 101.8% higher PBL when compared to White). All effects should be interpreted as the relative difference compared to White population within each specific phase value. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

For Black populations, the results show complex phase-dependent patterns. Starting with notably higher errors in Phase 0 (+113.0% compared to White areas), the disparities shift dramatically across phases. Phase 4 shows the most favorable performance for Black communities, with errors

42.5% lower than White areas. However, Phases 3 and 6 show notably higher prediction errors (+15.2% and +12.7% respectively) compared to White areas. Phases 1, 2, and 5 show better performance for Black communities (-6.7%, -16.3%, and -20.1% respectively).

Asian populations exhibit the most extreme phase-dependent variations in prediction accuracy. While starting with substantially lower errors than White areas in Phase 0 (-86.3%), the disparities show dramatic swings across phases. Phase 5 stands out with strikingly higher errors (+387.1% compared to White areas), while Phase 3 shows the best relative performance (-88.9%). Phase 2 also shows notably higher errors (+20.1%), while Phases 1, 4, and 6 maintain lower errors compared to White areas (-71.3%, -13.6%, and -55.3% respectively). These extreme variations suggest that predictions for Asian communities were particularly sensitive to phase-specific characteristics of the pandemic.

All these phase-dependent variations are statistically significant ($p < 0.001$), and they reveal that prediction fairness across racial and ethnic groups is not consistent throughout the pandemic, with certain phases (particularly Phase 0) associated with the largest disparities relative to White populations.

3.3.3.3 GLM-1c: Race, Ethnicity and Model Type

Our analysis of how prediction disparities vary across different minority race/ethnic groups and model types with respect to the White reference group reveals distinct patterns and demonstrates that model architecture choices have significant implications for prediction fairness (see Table 3.7).

Increases in Hispanic population with respect to the White reference group are associated with large performance disparities across all model types. Baseline models show the largest disparity (28.1% higher PBL compared to White areas), while Compartmental models show somewhat reduced,

Table 3.7: **GLM-1c**: Race $\times$ Model Type Effects Relative to White Reference Group

Variable	Coefficient Estimates		Relative Effect	
	e ^{Coef.} (SE)	95% CI	e ^{Coef.}	% Diff from White
% Asian (Compartmental ref.)	0.463*** (0.019)	[0.446, 0.481]	0.463***	-53.7%
× Baseline Models	1.635*** (0.046)	[1.494, 1.789]	0.757***	-24.3%
× Deep Learning	1.549*** (0.042)	[1.427, 1.681]	0.717***	-28.3%
× Ensemble	1.232*** (0.027)	[1.168, 1.300]	0.571***	-42.9%
× Statistical	0.934* (0.033)	[0.876, 0.995]	0.432***	-56.8%
% Black (Compartmental ref.)	1.009* (0.004)	[1.001, 1.017]	1.009*	+0.9%
× Baseline Models	1.012 (0.008)	[0.995, 1.028]	1.021*	+2.1%
× Deep Learning	0.902*** (0.006)	[0.890, 0.913]	0.910***	-9.0%
× Ensemble	0.942*** (0.005)	[0.933, 0.950]	0.950***	-5.0%
× Statistical	0.951*** (0.005)	[0.941, 0.960]	0.959***	-4.1%
% Hispanic (Compartmental ref.)	1.235*** (0.004)	[1.226, 1.243]	1.235***	+23.5%
× Baseline Models	1.037*** (0.008)	[1.021, 1.053]	1.281***	+28.1%
× Deep Learning	0.966*** (0.006)	[0.955, 0.978]	1.193***	+19.3%
× Ensemble	0.958*** (0.004)	[0.950, 0.967]	1.183***	+18.3%
× Statistical	0.982*** (0.005)	[0.972, 0.991]	1.213***	+21.3%

Model Statistics: Pseudo R^2 (CS) = 0.461; Log-Likelihood = 5,982,160; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-1c. We only discuss race, ethnicity and model type interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Tables B.11 and B.12. Model Diagnostics are provided in Figure B.5. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular race or ethnicity compared to the White population within each model type. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 3. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to White population for each model type (% Diff from White). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$, for a given race/ethnicity and model type, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to the White population (e.g., +23.5% means that the PBL error for Hispanic counties for Compartmental models is 23.5% higher PBL when compared to White). All effects should be interpreted as the relative difference compared to White population within each specific model type. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

though still substantial, disparities (+30.7%). Ensemble models register the lowest disparity (+18.3%).

Increases in Black population are associated with relatively modest variations in PBL across model types when compared to the White reference group. Compartmental and Baseline Models show marginally higher prediction errors (0.9% & +2.1% difference with White areas), while Deep Learning & Ensemble models demonstrate better relative performance (-9.0%, -5%). Statistical models perform better for Black communities (-4.1%).

The Asian subgroup, on the other hand, shows improved relative effect with respect to White populations, and these disparities remain relatively consistent across model types, with PBL errors being at least 24.3% lower than errors for the White group across all model types.

These findings suggest that model architecture choices significantly impact prediction fairness, with Deep Learning and Ensemble models providing the most balanced performance across racial and ethnic groups, particularly for Hispanic and Black populations

3.3.3.4 GLM-1d: Race, Ethnicity and Mobility Data

Analysis of how prediction disparities vary with mobility data usage also reveal interesting patterns in the relationship between data inputs and prediction fairness (Table 3.8).

The relative effect analysis for Hispanic population indicates significant disparities in forecast performance when compared to the White reference group across mobility data uses in prediction models. For forecast models that use mobility data, increases in Hispanic population with respect to the White reference group are associated with a PBL 18.5% higher than that of the White population. That number increases to 22% for mixed models (ensembles of CDC ForecastHub models trained with and without mobility data). On the other hand, when mobility data is not used, PBL errors are 31.2% higher for each 1% increase in Hispanic population with respect to the White reference group, suggesting that mobility data usage can improve prediction fairness for Hispanic communities. However, considerable disparities remain (higher PBL errors) when compared to the White population.

For Black population, and when compared to the White reference group, not using mobility data in the forecast models was associated with PBL errors being 6.8% higher. However, using mobility data reversed that trend, with PBL errors being 5.1% lower when the Black population increases 1%

Table 3.8: **GLM-1d**: Race $\times$ Mobility Data Inclusion Effects Relative to White

Variable	Coefficient Estimates			Relative Effect	
	e ^{Coef.}	(SE)	95% CI	e ^{Coef.}	% Diff from White
% Asian (No Mobility ref.)	0.351***	(0.027)	[0.332, 0.370]	0.351***	-64.9%
× Mixed	1.585***	(0.035)	[1.480, 1.697]	0.556***	-44.4%
× Mobility Used	1.626***	(0.029)	[1.536, 1.722]	0.571***	-42.9%
% Black (No Mobility ref.)	1.068***	(0.005)	[1.058, 1.079]	1.068***	+6.8%
× Mixed	0.909***	(0.006)	[0.899, 0.919]	0.971***	-2.9%
× Mobility Used	0.889***	(0.005)	[0.881, 0.897]	0.949***	-5.1%
% Hispanic (No Mobility ref.)	1.312***	(0.005)	[1.301, 1.324]	1.312***	+31.2%
× Mixed	0.930***	(0.005)	[0.920, 0.940]	1.220***	+22.0%
× Mobility Used	0.903***	(0.004)	[0.895, 0.911]	1.185***	+18.5%

Model Statistics: Pseudo R^2 (CS) = 0.461; Log-Likelihood = 5,982,399; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-1d. We only discuss race, ethnicity and mobility usage interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed Tables B.13 and B.14. Model Diagnostics are provided in Figure B.5. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular race or ethnicity compared to the White population within each mobility usage category. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 3. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to White population for each mobility usage category (*% Diff from White*). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$, for a given race/ethnicity and mobility usage category, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to the White population (e.g., +31.2% means that the PBL error for Hispanic counties for the given mobility usage category is 31.2% higher PBL when compared to White). All effects should be interpreted as the relative difference compared to White population within each specific mobility usage category. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

with respect to the White group.

Prediction performance for Asian populations when considering the use of mobility data was significantly better than the performance for White counties, independently of whether mobility data was used or not in the models, with PBL errors being between 42.9% and 64.9% significantly lower than errors for the White reference group. Interestingly, we also observe that use of mobility data helps reduce the disparity in performance errors (42.9% and 44.4% vs. 64.9%).

These results reveal significant prediction performance differences between minority racial and ethnic groups with respect to White across mobility data use settings; while suggesting that the incorporation of mobility data might help reduce prediction disparities for minority racial groups.

Summary: Race X Model-Data Characteristics

- ⇒ Hispanic communities consistently experience higher prediction errors compared to White areas across most configurations.
- ⇒ Disparities decrease with longer forecast horizons.
- ⇒ Initial phases of the pandemic saw widest disparity for both Hispanic and Black subgroups
- ⇒ Asian communities maintained lower error rates overall & showed substantial phase-dependent variations (-88.9% to +387.1%)
- ⇒ Deep Learning and Ensemble models demonstrated most balanced performance across all racial and ethnic groups
- ⇒ Mobility data usage generally helps reduce prediction disparities

3.3.4 Fairness of COVID-19 case predictions across urbanicity and model-data characteristics

3.3.4.1 GLM-2a: Urbanicity and Forecast Lookahead

Table 3.9 exhibits significant disparities in COVID-19 case prediction errors between urbanization levels across different forecast horizons. Our findings demonstrate that both MC and SMM areas consistently experience higher prediction errors when compared to LM areas (reference group) across lookaheads; and that the magnitude of these disparities decreases as the prediction horizon extends.

MC areas show the most pronounced disparity, with baseline (7-day) prediction errors 13.3% higher than LM areas. This disparity, while persistent, diminishes at longer prediction horizons. For 14-day forecasts, the relative effect shows that MC areas still experience 9.1% higher errors than LM

Table 3.9: **GLM-2a**: Urbanicity $\times$ Lookahead Effects Relative to LM Reference Group

Variable	Coefficient Estimates			Relative Effect	
	e ^{Coef.}	(SE)	95% CI	e ^{Coef.}	% Diff from LM
Micropolitan (MC) (7-day ref.)	1.133***	(0.002)	[1.128, 1.137]	1.133***	+13.3%
× 14-day ahead	0.963***	(0.003)	[0.958, 0.968]	1.091***	+9.1%
× 21-day ahead	0.926***	(0.003)	[0.921, 0.930]	1.049***	+4.9%
× 28-day ahead	0.900***	(0.002)	[0.896, 0.905]	1.020***	+2.0%
Small/Medium Metro (SMM) (7-day ref.)	1.048***	(0.002)	[1.044, 1.053]	1.048***	+4.8%
× 14-day ahead	0.984***	(0.003)	[0.979, 0.990]	1.031***	+3.1%
× 21-day ahead	0.975***	(0.003)	[0.970, 0.981]	1.022***	+2.2%
× 28-day ahead	0.969***	(0.003)	[0.964, 0.975]	1.015***	+1.5%

Model Statistics: Pseudo R^2 (CS) = 0.464; Log-Likelihood = 5,985,040; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-2a. We only discuss urbanicity and lookahead interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Section B.2.3.2 and table B.16. Model Diagnostics are provided in Figure B.6. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular urbanicity level compared to the Large Metropolitan (LM) areas within each lookahead. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 4. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to LM areas for each lookahead variable (*% Diff from LM*). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given urbanicity level and lookahead value, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to LM areas. All effects should be interpreted as the relative difference compared to LM areas within each specific lookahead value. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

areas. This gap continues to narrow, decreasing to 4.9% for 21-day forecasts and further reducing to 2.0% for 28-day predictions.

SMM areas exhibit a similar pattern but with smaller magnitudes of disparity compared to MC. These areas show prediction errors 4.8% higher than LM areas for 7-day forecasts. Again the pattern of the disparity decreasing is also observed for SMMs, with relative effects showing 3.1% higher errors for 14-day forecasts, 2.2% for 21-day forecasts, and 1.5% for 28-day forecasts compared to LM areas. These differences, while smaller, remain statistically significant across all forecast horizons ($p < 0.001$).

Table 3.10: **GLM-2b**: Urbanicity $\times$ Phase Effects Relative to LM Areas

Variable	Coefficient Estimates		Relative Effect	
	$e^{\text{Coef.}}$ (SE)	95% CI	$e^{\text{Coef.}}$	% Diff from LM
Micropolitan (MC) (Phase 0 ref.)	1.116*** (0.003)	[1.110, 1.122]	1.116***	+11.6%
× Phase 1	0.997 (0.003)	[0.991, 1.003]	1.113***	+11.3%
× Phase 2	0.858*** (0.004)	[0.852, 0.864]	0.958***	-4.2%
× Phase 3	1.007* (0.003)	[1.001, 1.014]	1.124***	+12.4%
× Phase 4	0.915*** (0.003)	[0.909, 0.921]	1.021***	+2.1%
× Phase 5	0.820*** (0.004)	[0.814, 0.827]	0.915***	-8.5%
× Phase 6	1.004 (0.004)	[0.996, 1.013]	1.120***	+12.0%
Small/Medium Metro (SMM) (Phase 0 ref.)	1.091*** (0.003)	[1.085, 1.098]	1.091***	+9.1%
× Phase 1	0.960*** (0.004)	[0.953, 0.966]	1.047***	+4.7%
× Phase 2	0.876*** (0.004)	[0.869, 0.883]	0.956***	-4.4%
× Phase 3	0.998 (0.004)	[0.991, 1.006]	1.089***	+8.9%
× Phase 4	0.898*** (0.004)	[0.891, 0.904]	0.980***	-2.0%
× Phase 5	0.862*** (0.005)	[0.854, 0.870]	0.940***	-6.0%
× Phase 6	0.971*** (0.005)	[0.963, 0.980]	1.059***	+5.9%

Model Statistics: Pseudo R^2 (CS) = 0.466; Log-Likelihood = 5,987,062; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-2b. We only discuss urbanicity and phase interaction coefficients. For clarity purposes, all other main effects and control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Tables B.17 and B.18. Model Diagnostics are provided in Figure B.6. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular urbanicity level compared to the Large Metropolitan (LM) areas within each phase. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 4. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to LM areas for each phase (% Diff from LM). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given urbanicity level and phase, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to LM areas. All effects should be interpreted as the relative difference compared to LM areas within each specific phase. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

3.3.4.2 GLM-2b: Urbanicity and COVID-19 Phase

Both MC and SMM areas show significantly different prediction performance compared to LM areas, with these disparities varying substantially across different pandemic phases (See Table 3.10).

MC areas demonstrate the highest baseline disparity (Phase 0), with prediction errors 11.6% higher than LM areas. The magnitude of this disparity fluctuates notably across different pandemic phases. During Phase 3, MC areas experienced their largest disparity, with errors 12.4% higher than LM areas. Conversely, Phases 2 and 5 show a reversal of this pattern, with MC areas actually perform-

ing better than LM areas, showing 4.2% and 8.5% lower errors respectively ($p < 0.001$). Phases 1 and 6 maintained similar disparities to the baseline (11.3% and 12.0% higher errors)

SMM areas show a similar but less pronounced pattern of disparities. The baseline prediction errors for SMM areas are 9.1% higher than LM areas. Like MC areas, SMM areas show varying performance across phases, but with generally smaller magnitudes of disparity. The pattern of better performance in Phases 2 and 5 is repeated, with SMM areas showing 4.4% and 6.0% lower errors than LM areas respectively ($p < 0.001$). The highest disparities for SMM areas occur in Phase 3 (8.9% higher errors) and Phase 6 (5.9% higher errors).

3.3.4.3 GLM-2c Urbanicity and Model Type

Table 3.11: **GLM-2c**: Urbanicity $\times$ Model Type Effects Relative to LM Areas

Variable	Coefficient Estimates			Relative Effect	
	e ^{Coef.}	(SE)	95% CI	e ^{Coef.}	% Diff from LM
Micropolitan (MC) (Compartmental ref.)	1.066***	(0.001)	[1.063, 1.069]	1.066***	+6.6%
× Baseline Models	0.984***	(0.004)	[0.977, 0.991]	1.049***	+4.9%
× Deep Learning	0.979***	(0.003)	[0.974, 0.985]	1.044***	+4.4%
× Ensemble	0.988***	(0.002)	[0.984, 0.992]	1.053***	+5.3%
× Statistical	1.029***	(0.002)	[1.024, 1.034]	1.097***	+9.7%
Small/Medium Metro (SMM) (Compartmental ref.)	1.028***	(0.001)	[1.025, 1.031]	1.028***	+2.8%
× Baseline Models	1.000	(0.004)	[0.992, 1.009]	1.028***	+2.8%
× Deep Learning	0.983***	(0.003)	[0.976, 0.989]	1.011***	+1.1%
× Ensemble	0.990***	(0.002)	[0.986, 0.995]	1.018***	+1.8%
× Statistical	1.019***	(0.003)	[1.014, 1.025]	1.048***	+4.8%

Model Statistics: Pseudo R^2 (CS) = 0.462; Log-Likelihood = 5,983,479; N = 1,526,869

Notes: $p < 0.001$, $p < 0.01$, $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-2c. We only discuss urbanicity and model type interaction coefficients. For clarity purposes, all other control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Tables B.19 and B.20. Model Diagnostics are provided in Figure B.6. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular urbanicity level compared to the Large Metropolitan (LM) areas within each model type. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 4. To evaluate better the relative effect, we also discuss the percentage change in forecast error when compared to LM areas for each model type (% Diff from LM). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given urbanicity level and model type, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to LM areas. All effects should be interpreted as the relative difference compared to LM areas within each specific model type. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

Table 3.11 shows the interaction effect analysis for urbanicity levels and types of forecast models. Overall, we observe that MC and SMM suffer from higher PBL errors than LM areas across all types of forecasting models.

The highest performance disparities across model types are observed between MC and LM counties. Specifically, statistical models are the ones with the most pronounced differences, with MC areas experiencing 9.7% higher errors than LM areas; followed by Compartmental models (reference group) at 6.6%, Ensemble models at 5.3%, baseline models at 4.9% and deep learning models at 4.4%. All these differences remain statistically significant ($p < 0.001$).

As seen before for prediction lookaheads and some phases, SMM areas demonstrate a similar pattern but with smaller magnitudes of disparity. Statistical models show the largest disparity between SMM and LM counties, with PBL errors being 4.8% higher in SMM areas. Compartmental (reference group) and baseline model prediction errors for SMM counties are 2.8% higher than LM areas. The disparities are notably smaller for ensemble and deep learning models. Ensemble models have PBL errors 1.8% higher in SMM counties when compared to LM counties, while deep learning models forecasting cases in SMM counties have PBL errors 1.1% higher than LMs.

These findings suggest that deep learning models may be most effective at minimizing urbanization-related disparities in prediction accuracy. Conversely, statistical models appear to amplify these disparities across both MC and SMM areas. The consistent pattern of higher disparities in MC areas compared to SMM areas, regardless of model type, indicates that predictive challenges in less urbanized areas persist across modeling approaches, though their magnitude can be influenced by model selection.

3.3.4.4 GLM-2d: Urbanicity and Mobility Data

Table 3.12: **GLM-2d**: Urbanicity $\times$ Mobility Usage Effects Relative to LM

Variable	Coefficient Estimates			Relative Effect	
	e ^{Coef.}	(SE)	95 % CI	e ^{Coef.}	% Diff from Rural
Micropolitan (MC) (No Mobility ref.)	1.088 ^{***}	(0.002)	[1.084, 1.092]	1.088 ^{***}	+8.8%
$\times$ Mixed	0.979 ^{***}	(0.003)	[0.974, 0.984]	1.065 ^{***}	+6.5%
$\times$ Mobility Used	0.972 ^{***}	(0.002)	[0.968, 0.976]	1.058 ^{***}	+5.8%
Small/Medium Metro(SMM) (No Mobility ref.)	1.056 ^{***}	(0.002)	[1.052, 1.061]	1.056 ^{***}	+5.6%
$\times$ Mixed	0.976 ^{***}	(0.003)	[0.970, 0.981]	1.031 ^{***}	+3.1%
$\times$ Mobility Used	0.963 ^{***}	(0.002)	[0.958, 0.967]	1.017 ^{***}	+1.7%

Model Statistics: Pseudo R^2 (CS) = 0.462; Log-Likelihood = 5,983,425; N = 1,526,869

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Dependent Variable:* Square root PBL. *Link Function:* Log. *Regression Family:* Gaussian. The table shows the GLM coefficients and their significance for model GLM-2d. We only discuss urbanicity and mobility usage interaction coefficients. For clarity purposes, all other control variables: Health outcomes, age 65+ and state fixed effects are only shown and discussed in Section B.2.3.2 and table B.22. Model Diagnostics are provided in Figure B.6. For the Coefficient Estimates, $\exp(\text{Coef})$ represents the multiplicative effect on the outcome, SE: Standard Error and CI: Confidence Interval. The Relative Effect represents the multiplicative effect on the forecast error (PBL) of a particular urbanicity level compared to the Large Metropolitan (LM) areas within each mobility usage category. The relative effect is represented by $\exp(\text{Coef})$ and computed as $e^{\beta_i + \delta_{ij}}$ with coefficients from Equation 4. To better evaluate the relative effect, we also discuss the percentage change in forecast error when compared to LM areas for each mobility usage category (% Diff from LM). This change is computed as $(1 - e^{\beta_i + \delta_{ij}}) * 100\%$ for a given urbanicity level and mobility usage category, and it represents the percentage increase or decrease in the forecast error (PBL) with respect to LM areas. All effects should be interpreted as the relative difference compared to LM areas within each specific mobility usage category. The relative coefficient significance is evaluated using `linearHypothesis` in R (*car* package [69])

Table 3.12 represents the interaction analysis between urbanization levels and the use of mobility data in forecasting models. The relative effects analyses show a similar pattern to the other model-data characteristics we have discussed *i.e.*, when compared against LM counties, MC and SMM counties are associated with significantly higher prediction errors across mobility data use approaches (no mobility data, mobility data or mixed model).

MC areas exhibit the largest disparity, with prediction errors 8.8% higher than LM areas when no mobility data is used to train the COVID-19 case forecasting models. This disparity is modestly reduced when mobility data is used, with MC areas showing 5.8% higher errors compared to LM areas, and 6.5% higher errors for mixed models (ensemble of models trained with and without mobility data). Both reductions in disparity are statistically significant ($p < 0.001$).

On the other hand, SMM areas demonstrate a similar pattern but with smaller magnitudes of disparity. Not using mobility causes prediction errors for SMM areas to be 5.6% higher than LM areas. The incorporation of mobility data appears to be more effective in reducing disparities for SMM areas as well, with models using mobility data showing only 1.7% higher errors compared to LM areas. Mixed models, based on CDC ForecastHub ensembles of both models trained with and without mobility data show an intermediate improvement, with 3.1% higher errors compared to LM areas.

Summary: Urbanicity X Model-Data Characteristics

- ⇒ SMM and MC are consistently associated to higher prediction errors compared to LM across most configurations.
- ⇒ Disparities decrease with longer forecast horizons.
- ⇒ Phases 2 and 5 saw a reversal in the pattern with SMM and MC performing better than LM counties.
- ⇒ Deep learning models demonstrated most balanced performance across urbanization levels.
- ⇒ Mobility data usage generally helps reduce prediction disparities

3.4 Dashboard

This paper has revealed significant disparities in COVID-19 case prediction accuracy across race, ethnicity and urbanization level. The regression analyses we have presented, evaluate the relationship between COVID-19 case prediction errors and racial and ethnic groups, urbanization level, model

type, the use of mobility data, lookahead, and phase. Our findings are based on global trends across all Forecast Hub models, and provide general recommendations for researchers working in COVID-19 prediction models and for decision makers using case predictions to inform pandemic policies. For example, we have shown that mobility data helps reduce the forecast error disparities between racial groups and urbanization levels. This fact can be used modelers and decision makers to support the use of mobility data to enhance forecast models.

Nevertheless, it is important to acknowledge that researchers and decision makers might also want to assess the specific performance of each COVID-19 county case prediction model individually, exploring PBL error differences between racial and ethnic groups or urbanization levels for a given model, their statistical significance, or whether these differences persist when considering specific lookaheads or phases. To enable individual model evaluation, we have created an interactive dashboard (see Figure 3.3), that will be made publicly available upon the publication of this paper. The dashboard displays a model’s performance error (PBL) for a given protected attribute - race and ethnicity or urbanization level - that can be selected by the user from the user interface.

To allow for meaningful explorations, the individual model errors are displayed using the Accuracy Equality Ratio (AER) [32], which measures the difference in error distributions between protected and unprotected groups for a given protected attribute. The AER is computed as a quotient between the model’s performance error (PBL) for a given protected group g across all counties and the model’s performance error for the unprotected group across all counties: $AER_g = \frac{PBL(protected_group_g)}{PBL(unprotected_group)}$ where $PBL(protected_group_g)$ and $PBL(unprotected_group)$ are the pinball ball loss metric for protected and unprotected groups respectively. For the race and ethnicity protected attribute we define three protected groups with respect to the White unprotected group: Asian, Black and Hispanic, and associate the plurality race to each county, i.e., the race or ethnicity that makes up the largest percent-

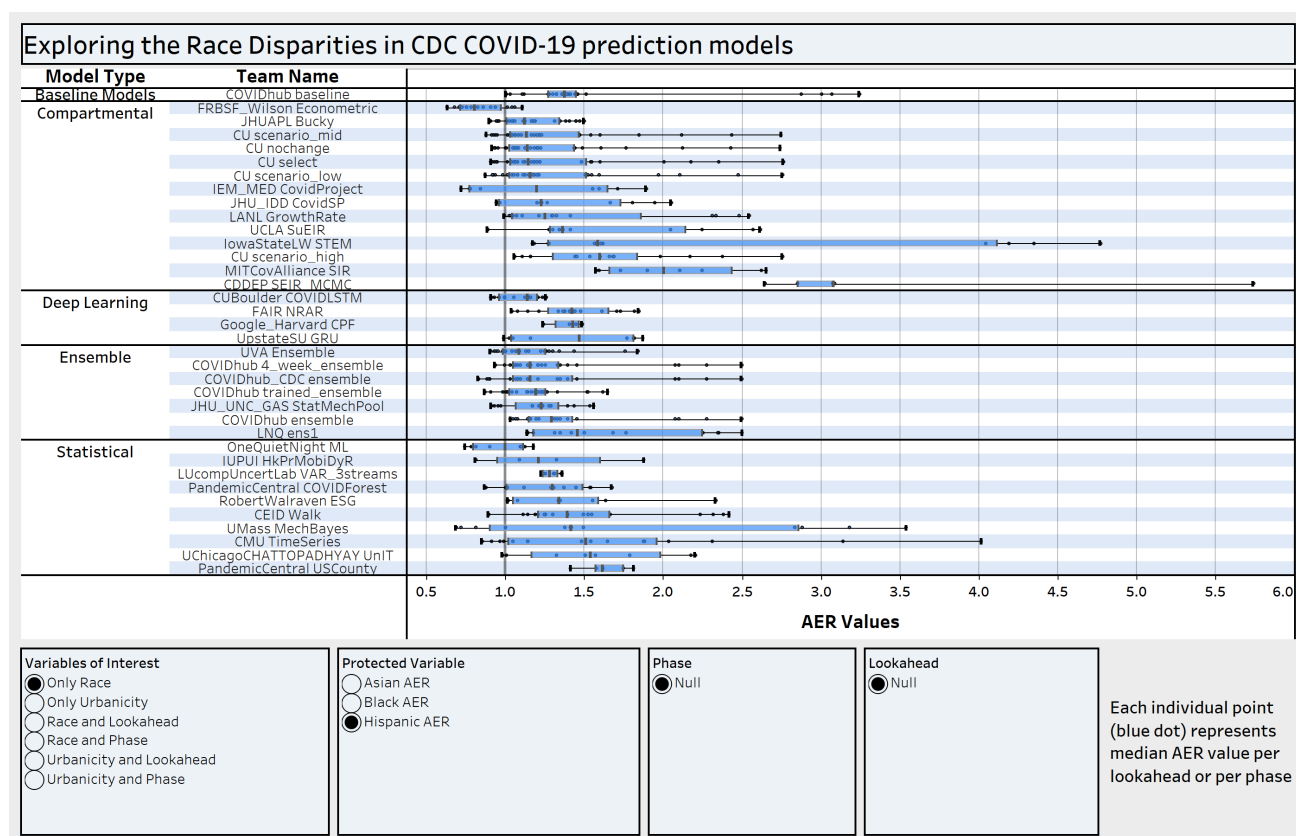


Figure 3.3: **Forecast Hub Fairness Dashboard** showing the Average Error Ratio (AER) distribution across different COVID-19 prediction models, organized by model type. Within each model type, teams are sorted in ascending order based on their median AER values. Since the user has selected “Only Race” as the variable of interest (see bottom left box) and “Hispanic” as the protected variable (see bottom center box), the AER values compare prediction errors between Hispanic and White counties, where values above 1.0 indicate higher prediction errors for Hispanic counties. Box plots show the distribution of AER values across all predictions, with the center line representing the median, boxes showing the interquartile range, and whiskers extending to the minimum and maximum values.

age for that county. For the urbanization code, we use the protected groups described in the paper: Micropolitan (MC) and Small and Medium Metro Areas (SMM) with Large Metropolitan Areas being the unprotected group. Similar error distributions between the protected and the unprotected groups will produce AER values close to one. AER values larger than one point to higher errors for the protected group, and AER values smaller than one point to higher errors for White or large metropolitan counties (baseline groups).

Interactive Features. The dashboard allows users to carry out different types of racial/ethnic and urbanization fairness analyses for each individual model. As shown in the four boxes at the bottom of Figure 3.3, the dashboard includes the following key interactive features:

- **Variables of Interest Selection:** Users can choose from six different analytical perspectives to analyze the fairness of a given COVID-19 forecast model:
 - Race or ethnicity analysis, that allows to analyze the fairness of the predictions for a given minority race with respect to White (see example in Figure 3.3)
 - Urbanicity analysis, that allow to evaluate the fairness of the predictions for a given urbanization level with respect to large metropolitan areas (see Figure B.7)
 - Analysis at the intersection of race/ethnicity and lookahead periods or pandemic phases, that allows to evaluate differences in COVID-19 forecast fairness for a given minority race (with respect to White) and for a given lookahead or pandemic phase (see Figure B.8, Figure B.9 for a couple of examples)
 - Urbanicity and lookahead periods or pandemic phases analysis, allowing users to explore differences in COVID-19 forecast fairness for a given urbanization level (with respect to large metropolitan areas) and a given lookahead or phase (see Figure B.10, Figure B.11 for a couple of examples)
- **Protected Variable Selection:** The dashboard allows users to focus on specific demographic groups:
 - For racial/ethnic analysis: Black, Hispanic, or Asian AER (White is the baseline group)
 - For urbanicity analysis: Micropolitan or Small/Medium Metro AER (Large Metropolitan

MODEL FAIRNESS CARD		MODEL FAIRNESS CARD	
Model Information		Model Information	
Team Name:	LUcompUncertLab VAR_3streams	Team Name:	IowaStateLW STEM
Variables of Interest:	Only Race	Variables of Interest:	Only Race
Mean Difference with Unprotected Reference Group		Mean Difference with Unprotected Reference Group	
Mean Differences in PBL with unprotected group: 0.0000748		Mean Differences in PBL with unprotected group: 0.0002463	
Upper Difference:	0.000274	Upper Difference:	0.000446
Lower Difference:	-0.0001243	Lower Difference:	0.0000463
Team AER Values		Team AER Values	
Median AER Value:	1.280	Median AER Value:	1.588
AER Value:	1.239	AER Value:	1.615
Coverage Info		Coverage Info	
Total counties coverage	3143	Total counties coverage	3104
Total Number of Predictions	53284	Total Number of Predictions	80704
(a) LUcompUncertLabVAR_3streams		(b) IowaStateLW STEM	

Figure 3.4: Model fairness card displaying key performance metrics including model information, prediction error differences between protected and unprotected groups, AER values, and coverage statistics.

is the baseline group)

- **Temporal Analysis Options:**

- Phase selection (0-6) for analyzing performance across different pandemic periods
- Lookahead periods (7, 14, 21, or 28 days) for examining how prediction fairness varies with forecast horizon

Sample Use Case. Figure 3.3 shows an example of the dashboard for the exploration of individual model performance by race and type of model, with a focus on the relationship between the prediction errors for Hispanic and White counties ($AER_{Hispanic}$). The box plots for each model represent its AER distribution across all counties; and a user can explore the median AER as well as its quantiles for each predictive model. In this example, most of the AERs are above one, pointing to unfair forecasts (higher errors) for the Hispanic group when compared to White counties.

Hovering over the model points displays all the information in the format of a ‘fairness nutritional card’ as shown in Figure 3.4. ‘Nutritional labels’ were proposed by Stoyanovich and Howe

to assess model fairness [161] drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. We have adapted these cards to the COVID-19 fairness context, as a way to provide detailed COVID-19 forecast model fairness information. The fairness nutritional cards in our dashboard (see Figure 3.4 for a sample) provide detailed information organized into four key sections: (1) *Model Information*, which identifies the team name and the variables being analyzed; (2) *Mean Difference with Unprotected Reference Group*, which quantifies the prediction error differences between protected and unprotected groups in terms of PBL values, including upper and lower bounds; (3) *Team AER Values*, showing both the median and specific AER values that indicate the relative performance between groups; and (4) *Coverage Info*, which provides context about the number of counties and total predictions covered by each team.

Looking into Figure 3.3 and Figure 3.4, we can observe that the *IowaStateLW STEM* model exhibits a wide range of median AER values (min: 1.172, max: 4.772, median: 1.588) across its predictions, indicating highly variable fairness performance when assessed across different phases and lookaheads. On the other hand, the *LUcompUncertLab VAR_3streams* model shows consistent median AER values within a small range from 1.228 to 1.334 (median: 1.280). Comparing both models, we observe that although both of them are systematically producing predictions that are less fair for Hispanic counties (AER values are larger than 1), the *LUcompUncertLab VAR_3streams* model has lower max AER values, suggesting that model might be a better choice.

Overall, we posit that the dashboard facilitates dynamic exploration of a diverse set of metrics across different dimensions, allowing users to examine how fairness measures vary with changes in pandemic phases and lookahead through the user interface. When exploring these temporal variables, the nutritional card automatically updates to display the relevant phase or lookahead information for the selected view.

3.5 Discussion

This study highlights the critical need to audit COVID-19 prediction models due to significant disparities in prediction accuracy. Our findings reveal that certain minority groups, especially Hispanic communities, and less urbanized areas consistently experience higher prediction errors. The race and ethnicity analysis revealed that increases in Hispanic population (when compared to White) exhibit significantly higher PBL errors; while increases in Asian and Black population are associated with lower PBL errors when compared to the White population (reference group), and while controlling for health outcomes and older population at the county level. The analysis on urbanization levels, on the other hand, revealed an inverse relationship between the level of urbanization and the magnitude of prediction errors underscoring the unique challenges encountered by rural areas. Rural counties consistently face higher prediction errors than their urban counterparts, a pattern that persists across various model types and forecast windows.

Potential Reasons. These findings could be related to data quality (COVID-19 cases, mobility data) or structural problems. Hispanic groups or less urbanized areas being associated with significantly higher errors (when compared to their baselines: White population and large metropolitan area), could point to lack of quality COVID-19 case or mobility data for the Hispanic population and less urbanized areas. It could also be due to more complex spreading patterns that make COVID-19 cases harder to predict for the Hispanic population or rural areas, or two structural differences such as reduced access to medical facilities or testing sites. On the other hand, these findings are also pointing to better COVID-19 case or mobility data for Asian and Black population when compared to White; or to simpler spreading patterns that are easier to forecast, hence producing lower errors.

The implications of these findings are significant, since systematic disparities in model perfor-

mance could lead to unfair distribution of public health resources or to less effective pandemic response efforts in Hispanic counties and in less densely populated regions, when compared to White and urban regions. Our definition of prediction fairness is focused on achieving similar prediction errors across racial, ethnic and urban-rural groups because COVID-19 cases have been used to make resource allocation and intervention decisions e.g., hospital beds or stay-at-home orders. Hence, higher prediction errors for minority racial groups or rural regions could in turn translate into unfair resource allocation for communities that have borne the brunt of the pandemic. Ultimately, we want to ensure our findings serve as a critical call to action for researchers and decision makers to analyze model performance disaggregated by racial/ethnic and urban-rural variables.

Interaction Analysis. Our interaction analysis provides a more multifaceted understanding of fairness in COVID-19 modeling. We find that deep learning models tend to produce the lowest disparities in errors across racial, ethnic and urban-rural groups, while compartmental and statistical models tend to be associated with the highest disparities. Our results have also shown that the use of mobility data helps reduce prediction error disparities for racial and ethnic groups as well as across urbanization levels. Short-term lookaheads and certain pandemic phases (case-peak phases 1, 3 and 4) are also associated with higher prediction error disparities for minority racial groups and rural areas. These findings highlight the complex interplay between model characteristics, data inputs, and social determinants in shaping prediction fairness.

Potential Reasons. Decreasing disparities in prediction errors for higher lookaheads could be due to a reduction in the effect of COVID-19 case data bias (positive or negative) for minority groups and less urbanized areas. In fact, as the lookahead increases, prior work has shown that case prediction becomes more difficult [1], and this forecasting complexity appears to have a stronger effect than data bias on the prediction errors, thus making all errors more similar (more fair) across racial/ethnic groups

and urbanization levels. When looking into prediction error disparities across phases and race/ethnicity or urbanization levels, we argue that the higher prediction differences tend to take place during initial phases of the pandemic, which could point to factors like initial data collection issues, testing accessibility, or reporting practices that may have varied across racial and ethnic groups during the early stages of the pandemic, especially for Hispanic groups and less urbanized areas. It also revealing of the fact that there were very limited historical COVID-19 case data to learn from in the early phases.

Looking into types of models, deep learning (DL) models appear to be the best choice to reduce performance error differences between LMs and less urbanized areas; while DL models also appear to be a good compromise across minority groups. This could be pointing to DL modeling being able to better capture spatio-temporal dependencies without the behavioral assumptions of compartmental or statistical models, that show higher performance differences for some racial groups and across less urbanized levels.

Finally, mobility data appears to be providing additional information (insights into behavioral patterns) that helps reduce biases in model performance across the aforementioned sensitive groups.

These interaction results highlight the need for researchers and modelers to carefully examine their data sources, model assumptions, and potential biases that could lead to unfair predictions for certain population groups. Incorporating fairness considerations into the model development, validation, and deployment processes is essential to ensure equitable outcomes. Public health officials and policymakers should be aware of the potential disparities in the accuracy of COVID-19 prediction models and work closely with modelers to mitigate these disparities. Failure to address these issues could lead to the perpetuation of health inequities and could eventually undermine the effectiveness of pandemic response efforts.

In addition to our findings, this study has several limitations that should be acknowledged.

Firstly, we had to exclude some U.S. counties from our analysis due to insufficient data availability for the data sources we used. Second, while our work primarily focused on urbanicity and race/ethnicity as fairness-related variables, other important attributes, such as socioeconomic status or access to health-care, etc, were not considered and could be explored in future research. Additionally, we were unable to incorporate all minority racial groups like AIAN and NHPI due to inadequate population sizes, which constrained our ability to assess fairness comprehensively across all demographic groups.

Moving Forward. In our study, we have focused exclusively on county-level predictions because these are closer to local realities and allow for more actionable decision-making than state-level predictions. However, county-level statistics were collected only for COVID-19 cases, with hospitalizations or deaths only accounted for at the state level. Since prior work has shown that case counts might be more biased than hospitalization or death statistics [51], the results reflected in this paper could potentially change if hospitalization or death data were available at the county level and this study was replicated.

We posit that future research in COVID-19 case prediction models should focus on developing and validating bias mitigation strategies that account for performance disparities across race, ethnicity and urbanization levels. This may involve exploring alternative data sources, refining model architectures, and incorporating techniques to ensure fairness across different population groups. Additionally, more comprehensive and standardized race and ethnicity data collection in public health surveillance systems is crucial to enable accurate assessments of model fairness and to guide equitable decision-making.

Chapter 4: Assessing Ethnoracial Disparities in COVID-19's Impact on Small Businesses in US Metro Areas Using Mobility Data

4.1 Introduction

The outbreak of the COVID-19 pandemic brought unprecedented challenges to communities worldwide, prompting the implementation of various non-pharmaceutical interventions (NPIs) aimed at limiting mobility and curbing the spread of the virus [15, 110]. These interventions, such as nationwide lockdowns and remote work strategies, underscored the crucial role of restricting movement in containing the epidemic [134, 135]. While these measures were effective in mitigating the health crisis, they also had significant repercussions for businesses, particularly small establishments that form the backbone of local economies [50, 59].

The pandemic has also exposed and magnified pre-existing socioeconomic disparities and systemic inequities within our society. In fact, vulnerable populations, including racial and ethnic minorities, have borne a disproportionate burden of the pandemic's impact including anti-Asian racism [99] as well the condition of being frontline workers [133]. Understanding the impact of COVID-19 on small businesses and its intersection with racial composition is also vital for policymakers, community leaders, and stakeholders. By identifying disparities and informing targeted interventions, tangible work can be done towards fostering a more equitable recovery and supporting the long-term sustain-

ability of small businesses.

To better understand the implications of COVID-19 on small businesses and to shed light on the intersection of ethnicity and business resilience, this study leverages aggregated geolocated cell phone data from SafeGraph, characterizing visits to small urban businesses in the US, with demographic data related to the social composition of the areas where these businesses are located. Analyzing business visitation patterns at the level of Census Block Groups in the US, we aim to provide insights into the changes in small urban business activity and their association with the race and ethnicity of their tracts. Although many types of businesses were affected by the pandemic, we focus on non-chain urban restaurants because these small businesses are major contributors to job creation while also being more vulnerable to falling demand [50].

The objective of this research is twofold. Firstly, we examine the overall changes of restaurant visitation patterns following the onset of the pandemic, assessing the extent to which small urban restaurant businesses in the US have been affected and are recovering from the initial shock. Secondly, we investigate whether disparities in visitation patterns exist across different racial and ethnic groups, seeking to uncover potential inequities in the recovery process across metropolitan areas. By delving into the visitation patterns and changes in visitation volumes across restaurants, we can gain valuable insights into the specific challenges faced by small urban restaurant businesses in the US during the COVID-19 pandemic. This information will contribute to a more nuanced understanding of the restaurant industry's dynamics and inform targeted strategies and support mechanisms for these vulnerable businesses.

The rest of the paper is organized as follows. In Section 2, we present a comprehensive description of the mobility and census data we will use in the analysis. Section 3 describes the statistical and visualization methods we use to explore changes in restaurant visitation patterns during the pan-

demic, and to examine the relationship between ethnorace and visitation changes as well as recovery patterns. Finally, Section 4 will present our main findings, followed by Section 5 with a description of the proposed future work.

4.2 Data

4.2.1 Mobility Data

The mobility patterns of millions of anonymized users in the US were tracked using mobile app software development kits (SDKs) by SafeGraph, a data provider that made this information publicly available at the beginning of the pandemic. To analyze the impact of COVID-19 on business visitation patterns, we utilized the points of interest (POIs) dataset from SafeGraph. We focused on two components of the POI data:

- **Core POI:** It includes the name and brand (indicating whether the POI is a part of a chain)
- **Patterns:** It includes the volume of daily visits per identified POI in the Safegraph database.

To focus specifically on small business restaurants, our analysis involves extracting points of interest (POIs) that meet two criteria: they are not associated with any chain (brand), and they fall within the category of “*Restaurants and Other Places*”. Around 80% of POIs in SafeGraph have no brand associated as they are single commercial locations. The visit patterns were measured as changes in volumes compared to the traffic before the pandemic. We collected SafeGraph data from January 1st, 2020 to January 31st, 2021. The data collected between January 1st, 2020, and March 7, 2020, was defined as the baseline to assess changes in visitation patterns during the pandemic, which we mark with a starting date of March 15, 2020. To reveal long-term trends that may be concealed by occasional

fluctuations we utilized a 7-day rolling average to smooth the time series data. This approach enables us to identify significant changes and discern overall changes in the data, providing a clearer picture of the long-term effects of the pandemic on small business restaurant visitation patterns.

4.2.2 ACS Data

To evaluate disparities in the impact of COVID-19 on small urban restaurant businesses with respect to race and ethnicity, we utilized three datasets obtained from the American Census Survey datahub¹. These datasets provided valuable information for our research:

- **CBG ShapeFiles:** We utilized CBG ShapeFiles to gather the boundaries of the Census Block Groups. We use CBG as the geographical unit of analysis, assigning to each CBG the changes in volumes of visits to small restaurants located within the CBG.
- **Urban-rural classification:** This dataset helped us identify urban CBGs by categorizing regions based on their urban level. We specifically focused on CBGs classified as level 1, which corresponds to large metropolitan counties.
- **Race Distribution:** This dataset provided detailed information on the racial composition of the population at the Census Block Group (CBG) level. It allowed us to examine the distribution of different racial groups within specific geographic areas. To carry out the proposed analysis, each CGB was associated to the majority group race i.e., the race with the largest population volume in that CBG.

Figure 4.1 shows all the urban Census Block Groups representing large metropolitan areas, providing an overview of the locations used in our analysis. Figure 4.2 represents the count of CBGs

¹<https://data.census.gov/>

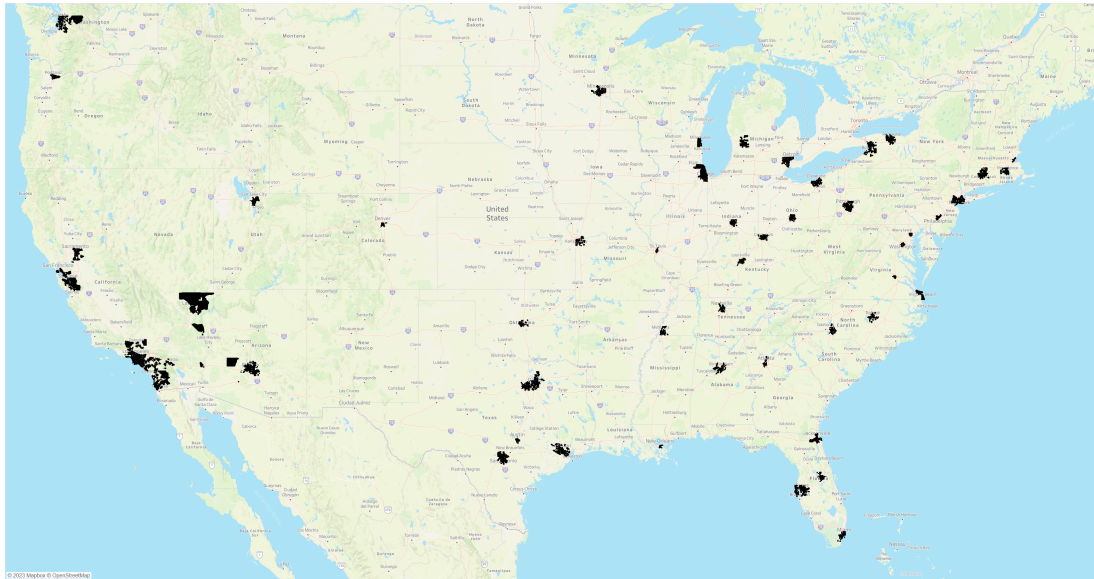


Figure 4.1: Urban Census Block Groups representing all the large metropolitan areas used in our analysis.

associated to each majority race. To delve deeper into the analysis, Figure 4.3 shows an example of the racial composition across CBGs for the Bay area and for New York City.

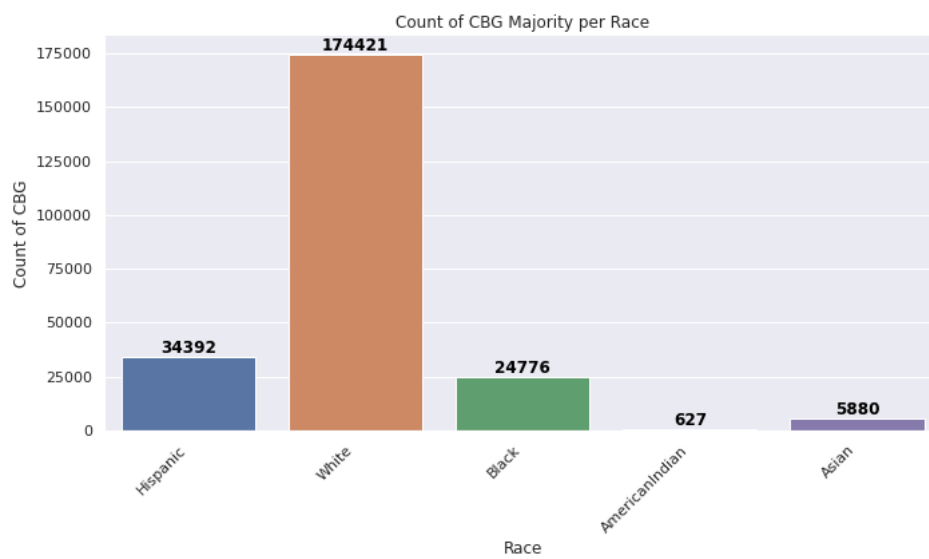


Figure 4.2: Count of CBGs associated to each majority race.

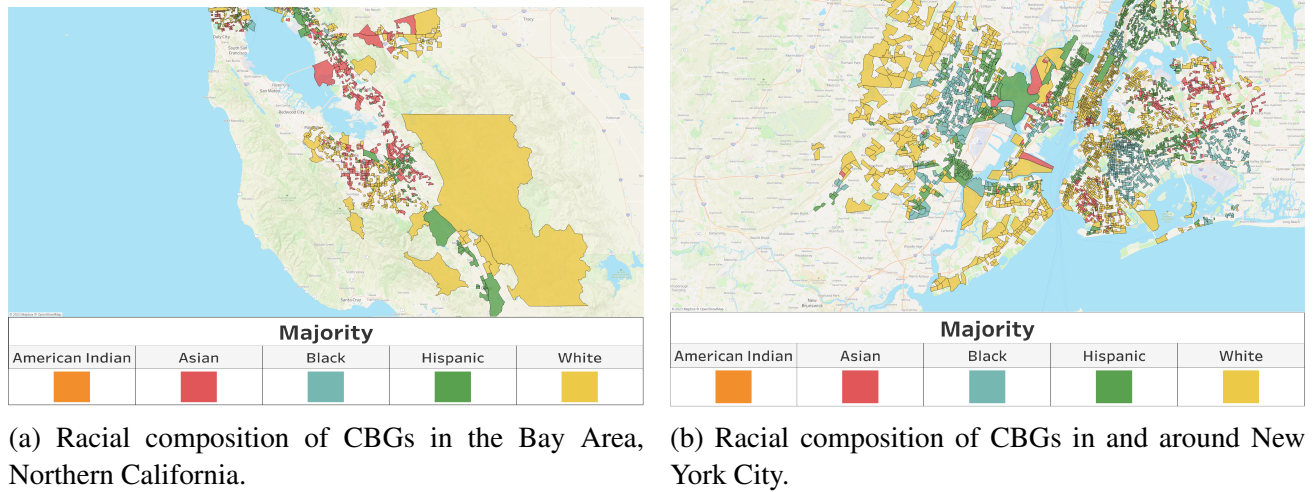


Figure 4.3: Racial composition at selected urban locations.

4.2.3 Why CBGs as a proxy measure of localized impact?

Firstly, CBG allows for a localized impact assessment. It provides a finer level of granularity than larger geographic units, such as census tracts or counties. By focusing on CBGs, researchers can zoom in on specific neighborhoods or areas within a city or region. This localized approach enables a more precise understanding of the effects of interventions, policies, or demographic changes on smaller-scale populations.

Secondly, CBG is particularly useful when studying highly homogenous populations. In some cases, neighborhoods or communities may exhibit a high degree of homogeneity in terms of race, ethnicity, income level, or other demographic characteristics. By analyzing CBGs within these areas, researchers can gain insights into the dynamics and changes occurring within these homogeneous populations. This level of analysis can help identify trends, patterns, or disparities that may not be evident when examining larger, more heterogeneous geographic units.

Lastly, the reliance on CBGs as the unit of analysis is prevalent in most quantitative studies of neighborhood racial change [29, 136, 144]. Researchers often utilize CBGs because they offer a

balance between geographic precision and sample size. CBGs generally have a population size that is manageable for analysis while still providing enough data points to draw statistically significant conclusions. This widespread usage of CBGs as a unit of analysis allows for comparability and consistency across studies, facilitating the accumulation of knowledge and the identification of broader trends.

4.3 Methodological Approach

For each urban CBG in the US, we compute a time series with the daily percentage change in aggregated visits to small restaurants within that CBG. The daily percentage change reflects changes in volumes of visits during the pandemic with respect to the predefined pre-pandemic visitation patterns. To investigate the recovery patterns across ethnoracial groups, we carry out separate analyses for each race and ethnicity by exclusively looking into the daily percentage change time series of census block groups with the same ethnoracial majority. For instance, to examine the pattern specific to the Asian demographic subgroup, we focus solely on the census block groups - and their corresponding daily visitation time series - where Asians constitute the majority population.

The proposed methodological approach will allow us to examine overall changes in restaurant visitation patterns following the pandemic, and whether disparities in visitation patterns exist across racial and ethnic groups, seeking to uncover potential inequities in the recovery process across metropolitan areas. Nevertheless, a potential limitation of this approach is that it will not reveal the reasons behind changes in the visitation patterns *i.e.*, small urban restaurants in specific ethnoracial CBGs could have received fewer visits during the pandemic either because residents of that CBG visited less their local restaurants, or because non-residents were visiting less that CBG. While the

former would point to the impact of a specific ethnorace's visiting behaviors to their local restaurants; the latter would signal changes in visiting behaviors from non-residents to specific ethnoracial CBGs. In addition, changes in visitation patterns for specific ethnoraces could also be due to higher volumes of business closings.

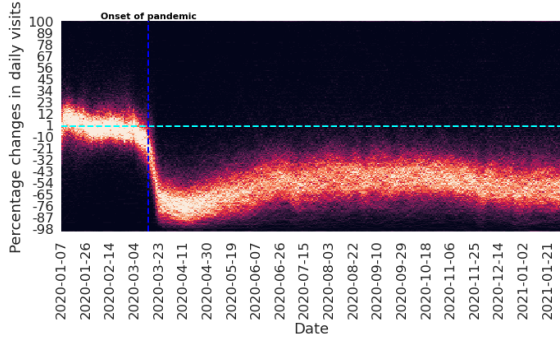
To evaluate overall small urban restaurants' recovery patterns, and changes across ethnoracial groups, we will also analyze the average percentage change in volumes of visits during the pandemic with respect to their corresponding pre-pandemic baselines, and we carry out this analysis for each ethnoracial group across three phases: phase I or social distancing (from mid March till mid May), phase II or management of the pandemic post-onset (from mid May till the end of November), and phase III or vaccination phase, to characterize the period when vaccines became available for certain social groups (from November onwards).

To evaluate statistical differences in recovery patterns across ethnoracial groups, we compute a two-sample Kolmogorov-Smirnov test [124] between each pair of ethnoracial average percentage change time series, with each average time series computed as the average of all time series across CBGs of the same majority ethnorace. The KS test, which is non-parametric, is used to capture statistically significant similarities and differences in both location and scale between pairs of ethnoracial time series. This focused approach allows us to analyze and compare the recovery trajectories among various racial communities. By employing this method, our objective is to identify any disparities or variations in post-pandemic recovery patterns based on racial composition within urban settings.

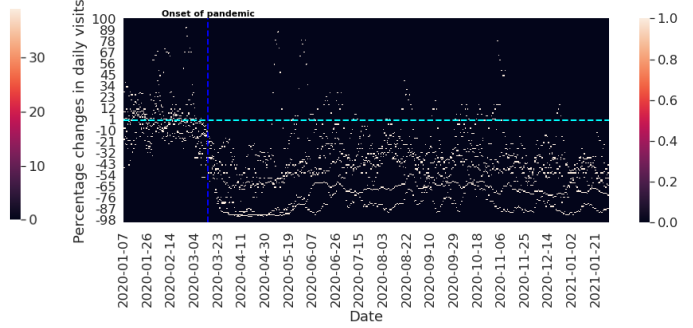
4.4 Results

Figure 4.4 shows heatmap matrices computed with the percentage change time series for the CBGs in our analysis. Each CBG is represented by a time series from January 7, 2020 to January 31, 2021; and every CBG contributes to a single pixel per day in the heatmap matrix. The heatmap matrix's density reflects the combined contribution of each CBG, enabling a comprehensive visual representation of the variations in visitation patterns over time and across ethnoraces. While 4.4(a) displays the heatmap matrix for all the CBGs, Figures 4.4(b–f) show the percentage change time series represented as heatmap matrices for each ethnicity or race in our analysis. Figure 4.5, on the other hand, represents the average percentage change time series across all CBGs with the same majority race. In other words, for each heatmap matrix in Figures 4.4(b–f), we compute its average time series, which allows for a straight forward comparison of changes in restaurant visitation patterns across ethnoraces.

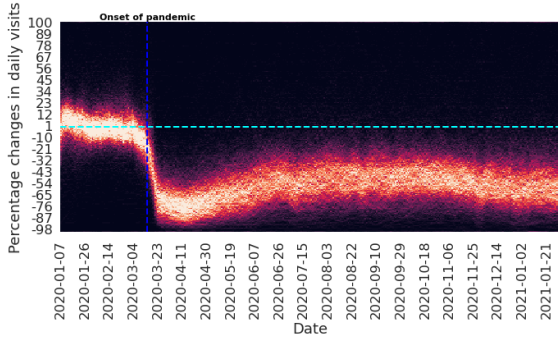
As we can observe in Figure 4.4, right after the global onset of the pandemic (March 15, 2020) the overall volume of visits to small urban restaurants as well as the volumes of visits disaggregated by ethnorace heavily decreased and did not go back to normal even after one year into the pandemic. To analyze disparities in visitation patterns and trajectory recoveries across ethnoraces, Figure 4.5 shows one time series per ethnorace, representing the average percentage changes in visitation patterns for small restaurants located in CBGs of a specific ethnorace. Overall, we observe that all CBGs, independently of their majority race were heavily affected by the pandemic with large percentage decreases of visitation patterns when compared to the baseline. We also observe that all time series reflect smooth trends across phases except for the American Indian time series, that suffers from more oscillations probably due to the smaller number of majority CBGs of that race in our dataset. Despite



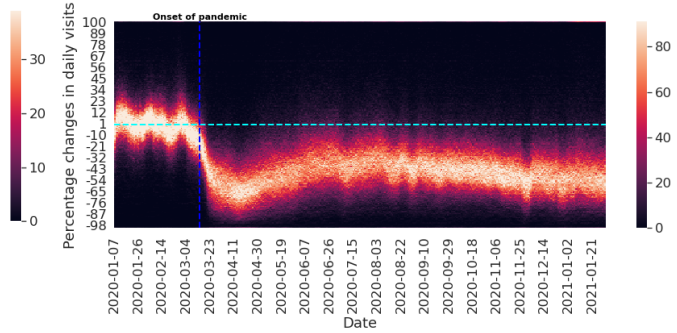
(a) Heatmap matrix for CBGs across all groups



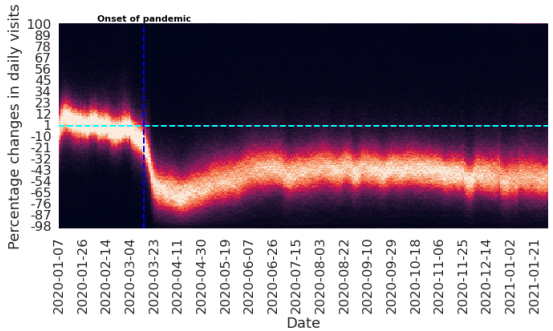
(b) Heatmap matrix for CBGs with American Indian Majority



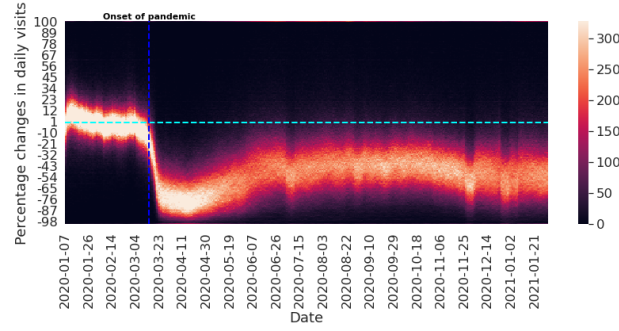
(c) Heatmap matrix for CBGs with Asian Majority



(d) Heatmap matrix for CBGs with Black Majority



(e) Heatmap matrix for CBGs with Hispanic Majority



(f) Heatmap matrix for CBGs with White Majority

Figure 4.4: Heatmap matrices for percentage changes in visit patterns

the large pandemic impact on visitation patterns, there exist nuanced differences across ethnoraces, with majority Asian and majority American Indian being not only the most impacted, but also the ones with the slowest recovery even after one year into the pandemic.

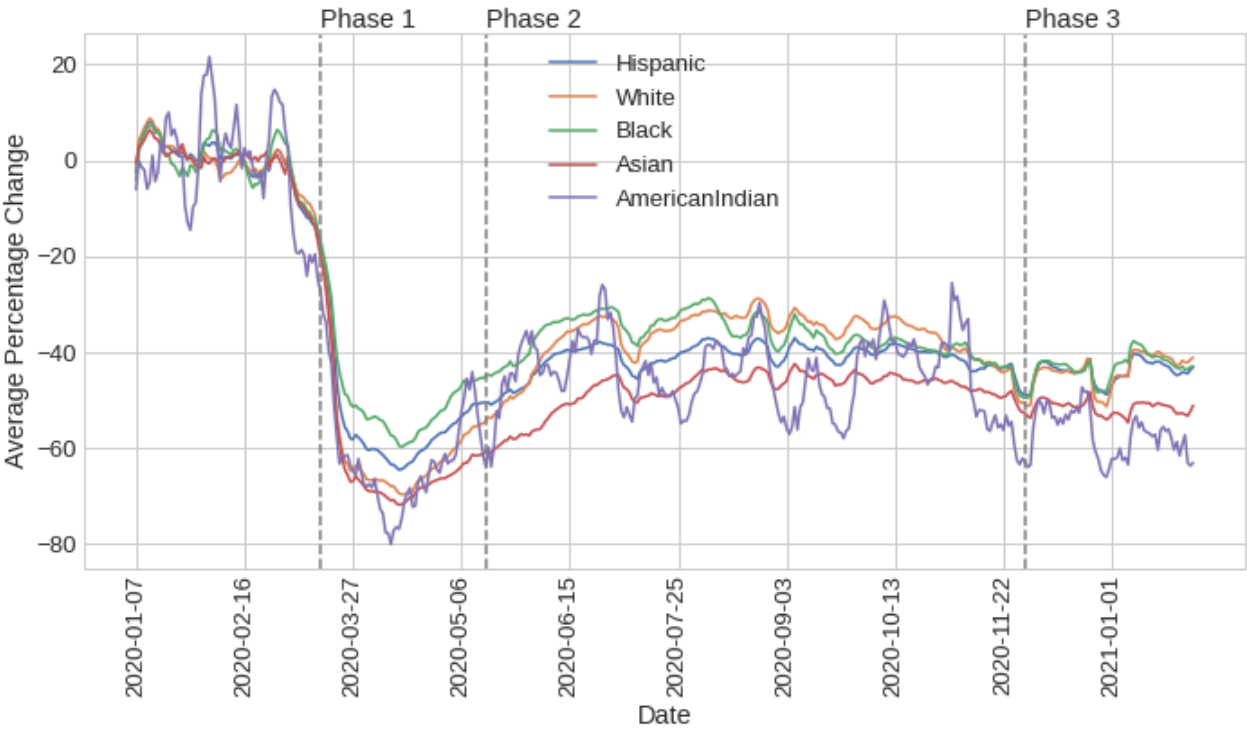


Figure 4.5: Average percentage change time series across all CBGs for each ethnicity or race. The vertical dashed lines separate the pre-pandemic phase from phases 1, 2 and 3.

To quantify these differences, Table 4.1 displays the average percentage changes in visitation patterns per ethnorace and across the three phases explained earlier. We can observe that Asian and American Indian CBGs are the ones most impacted, with larger average percentage decreases in visits with respect to pre-pandemic times of 42.7% and 41.7%, respectively, across all phases; reaching maximum decreases of 63% and 62% respectively for phase 1 (onset of the pandemic). Table 4.2 shows the KS statistics and p-values for all combinations of ethnoracial average percentage change time series. The null hypothesis was rejected for all pairs of distinct time series ($p_value = 0$) and the

Asian average percentage change time series is the one most different to the others (with the highest ks statistic values). Put together with the findings in Table 4.1, this statistical analysis shows that Asian CBGs are not only the least visited across pandemic phases, but that they are also the most statistically significantly different from the other ethnorace's distributions, with ks statistics in the 0.48 – 0.58 range when compared to majority Hispanic, White and Black CBGs, albeit more similar to the American Indian time series ($ks = 0.22$).

	American Indian	Asian	Black	Hispanic	White
Phase 1	-61.9199 (11.42972)	-63.0853 (11.52541)	-49.3982 (9.574766)	-54.6981 (10.45263)	-60.1125 (11.40951)
Phase 2	-44.7563 (8.186457)	-47.9976 (4.229237)	-36.8484 (4.287543)	-41.2918 (3.107061)	-37.1945 (5.706389)
Phase 3	-57.1392 (4.839321)	-51.597 (1.347966)	-43.0362 (2.738588)	-44.043 (2.244181)	-43.8665 (3.330658)
All Phases	-41.7721 (21.8725)	-42.709 (20.79114)	-33.4509 (16.75843)	-36.7284 (18.04656)	-35.4661 (19.03601)

Table 4.1: Average percentage change and the standard deviation (std) in visitation patterns per ethnorace over all the time period under study as well as for each phase.

KS_Test	American Indian	Asian	Black	Hispanic	White
American Indian	(0.0, 1.0)	(0.21995, 0.0)	(0.38107, 0.0)	(0.30179, 0.0)	(0.31969, 0.0)
Asian	(0.21995, 0.0)	(0.0, 1.0)	(0.58312, 0.0)	(0.48593, 0.0)	(0.50639, 0.0)
Black	(0.38107, 0.0)	(0.58312, 0.0)	(0.0, 1.0)	(0.27877, 0.0)	(0.10997, 0.01761)
Hispanic	(0.30179, 0.0)	(0.48593, 0.0)	(0.27877, 0.0)	(0.0, 1.0)	(0.32481, 0.0)
White	(0.31969, 0.0)	(0.50639, 0.0)	(0.10997, 0.01761)	(0.32481, 0.0)	(0.0, 1.0)

Table 4.2: Two-sample Kolmogorov-Smirnov tests between each pair of ethnoracial average percentage change time series. For each pair of distributions we report the ks statistic and its significance (p-value).

Chapter 5: Do Footfalls Reliably Mirror Economic Activity?

5.1 Introduction

Understanding the health and resilience of businesses in real time has become increasingly critical for policymakers, investors, and researchers seeking to respond to economic disruptions and guide recovery efforts. Numerous survey-based studies document business trajectories [17, 62, 63], including large efforts such as the Small Business Pulse Survey [20], but these instruments typically lack temporal granularity and are difficult to scale to large populations. In contrast, high-frequency human mobility data—digital “footfalls” or visits to commercial establishments—offer fine-grained temporal and spatial resolution, enabling near real-time monitoring of business activity and recovery dynamics [86, 103, 125, 159]. A growing literature uses mobility as a leading indicator of economic resilience, showing that place visits track disruption, adaptation, and rebound in local economies [140, 177, 178]. Together, these advances motivate the use of mobility as a high-resolution proxy for business health in settings where timely decisions matter most.

However, translating mobility patterns into meaningful economic insight requires a nuanced understanding of how foot traffic relates to actual spending, because visit–spend relationships are heterogeneous across sectors and business types [31]. The strength of these relationships varies markedly across economic categories of POIs—essential goods, discretionary retail, routine services, and social/experiential establishments—which reflect different trip purposes, basket sizes, and purchase

channels. The rapid expansion of digital commerce and omnichannel strategies [12] (curbside pickup, mobile ordering, delivery) has further weakened any simple one-to-one mapping between a physical visit and realized spend, a shift accelerated during COVID-19. Restaurants increasingly decouple physical visits from spending via delivery platforms and mobile ordering, so foot traffic alone can underestimate business performance [157]. Retailers adopting “buy online, pick up in store” models may generate single visits tied to multiple online transactions, while showrooming—visiting a store to examine products before purchasing online—can inflate foot traffic without commensurate in-store sales [141].

Given these complexities, a central question is how the visit–spend relationship has evolved amid rapid changes in consumer behavior. Existing work on visit–spend relationships [31, 40] consistently finds positive co-movements between visits and spending, but leaves several gaps. First, most analyses operate at regional or broad business-type aggregates, masking heterogeneity that is critical for economic monitoring; stable aggregates may conceal offsetting instabilities across categories. Second, these studies provide limited evidence on the *temporal stability* of visit–spend elasticities across years, despite the unprecedented disruptions of the COVID-19 period. Third, little attention has been paid to how neighborhood demographic characteristics moderate visit–spend relationships by business type, even though income, education, and racial/ethnic composition are known to shape consumer behavior and spending patterns [44, 92].

More broadly, our work is situated within a larger literature that uses high-frequency mobility from CDRs and smartphone traces to nowcast macroeconomic conditions and other high-stakes outcomes at national, regional, and establishment scales [9, 86, 103, 112, 125, 127, 129, 140, 152, 159, 166, 168, 177, 178, 179], as well as work that benchmarks mobility-based proxies against population, spending, or official statistics [31, 40, 49, 66, 137, 153]. We contribute to this literature by bring-

ing the evaluation down to the business-category level and quantifying, at weekly resolution, both the temporal stability and demographic moderation of visit–spend elasticities across multiple metropolitan areas.

To do so, we focus on visit–spend elasticity, defined as the percentage change in spending associated with a percentage change in visits, and pose two complementary research questions. First, we study the temporal stability of category-specific elasticities: **RQ1:** *Do visit–spend elasticities remain stable over time within business categories, and which categories serve as the most reliable proxies for economic activity? (Proxy validity).* We trace how elasticities evolved from 2019–2024, identify the most stable categories, and characterize how COVID-19 differentially affected visit–spend relationships by business type. Second, we examine demographic moderation: **RQ2:** *How do neighborhood demographic characteristics moderate the visit–spend relationship across business categories?* This analysis quantifies how income, education, and racial/ethnic composition shape spending responsiveness to visits across different business types.

Our analysis leverages a comprehensive dataset spanning 12.5 million POI-week observations from 52,051 establishments across six major metropolitan areas and twelve business categories from January 2019 through December 2024. We estimate category-specific panel regressions with POI fixed effects and year interaction terms to capture the temporal evolution of elasticities (RQ1), and demographic interaction models to assess moderation effects (RQ2). Our goal is to clarify the conditions under which mobility data serves as a reliable proxy for economic activity, an issue of growing importance given the increasing reliance on mobility for real-time economic monitoring. The main contributions of this paper are:

- **Systematic assessment of category-specific proxy reliability across time and geography.**

We evaluate which business categories sustain robust visit–spend relationships and quantify their temporal stability and cross-MSA consistency. Restaurants emerge as the most reliable proxy, combining moderate elasticities with exceptional temporal stability and geographic consistency. Discretionary retail categories exhibit the highest responsiveness but more variable stability across years and markets, whereas routine services show weak and statistically unreliable relationships, with frequent non-significant estimates. We also document systematic structural breaks induced by the pandemic.

- **Evidence of systematic demographic moderation and implications for equitable economic monitoring.** We show that neighborhood demographics systematically moderate visit–spend elasticities. Elasticities are consistently lower in neighborhoods with higher Hispanic population shares across most categories, with grocery stores and restaurants showing particularly strong attenuation. In contrast, income and education exhibit predominantly positive moderation effects. Applying uniform elasticities would therefore tend to overstate spending losses and understate true activity in communities with higher minority population shares, risking misdiagnosis and misallocation of recovery resources.

5.2 Related Work

5.2.1 Mobility Data as a proxy for downstream tasks

A growing literature exploits high-frequency mobility data from CDRs and smartphone traces to monitor high-stakes outcomes and nowcast macroeconomic conditions [103, 125, 127, 129, 152, 166, 168, 177]. At the regional level, Arhipova et.al. [9] uses aggregated, GDPR-compliant CDRs

to classify Latvian municipalities by weekday/weekend patterns, construct an efficiency index, and visualize commuting flows for ongoing monitoring. At the national scale, Spelta et.al.[159] combines Facebook commuting and Google mobility data in a dynamic-factor model to produce daily nowcasts of industrial production. At the establishment level, Goolsbee et.al. [86] use SafeGraph foot traffic to compare businesses across county policy borders within the same commuting zone, showing that legal shutdowns explain only a modest share of the collapse in visits, with most decline due to voluntary distancing.

Beyond nowcasting, Pappalardo et.al. [137] builds a reproducible CDR pipeline that aggregates mobility and social measures to estimate municipal deprivation and per-capita income, finding mobility diversity to be the most informative predictor. Leslie et.al. [112] releases a weekly, sector-resolved U.S. mobility dataset spanning 2019–2023 that summarizes visitors, distance, and dwell time for millions of POIs at 2- and 4-digit NAICS levels, providing useful infrastructure for tracking sectoral activity. Additionally, Yabe et.al. [179] constructs behavior-based dependency networks from same-day POI co-visits and show that these long-distance, category-biased links significantly improve predictions of business resilience to shocks.

5.2.2 Evaluation of mobility data as a proxy

Multiple studies evaluate mobility data as a proxy for economic outcomes [66]. Couture et.al. [49] provides a general-purpose measurement toolkit and public dataset that validates mobility signals against population, spending, and transport benchmarks. Using five months of CDR data, Scepanovic et.al. [153] infer users’ home and work locations to construct mobility metrics that proxy regional socio-economic conditions. They validate these proxies through high correlations with census pop-

ulation density and poverty indices, and by showing that commuting patterns match economic hubs and infrastructure. Matsumura et.al. [125] maps smartphone GPS to POIs and factories, demonstrating that service-sector foot traffic tracks official visitor and sales data, while factory foot traffic improves industrial-production nowcasts in labor-intensive industries with lower RMSE than random-walk baselines. Using UK district data, Cepparulo et.al. [40] estimates sizable elasticities between Google mobility and in-store card spending, with results robust during lockdowns and reopening, pinpointing where mobility proxies are most informative. Campos et.al. [31] exploits a state-by-week panel across Mexico’s 32 states to show that POS spending co-moves strongly with mobility, arguing that mobility indicators provide credible real-time proxies for consumption in settings with limited e-commerce and financial inclusion. Despite the important contributions of these studies, there still exists a gap in modeling the temporal stability of visit-spend relationships at a granular level, as well as in understanding how neighborhood demographics moderate these relationships, particularly given the behavioral shifts accelerated by the COVID-19 pandemic.

5.3 Material and Methods

We employ a *panel regression framework* that leverages the longitudinal structure of our mobility and spending data. Our empirical analysis utilizes a panel data structure, which combines cross-sectional variation across establishments with time-series variation within each establishment over multiple time periods. This design allows us to control for unobserved characteristics of individual businesses while tracking how relationships evolve.

We employ logarithmic transformations for both our dependent and independent variables, where $\log(\text{spend}_{i,t})$ denotes the logarithm of weekly spending at POI i in week t , and $\log(\text{visits}_{i,t})$ denotes

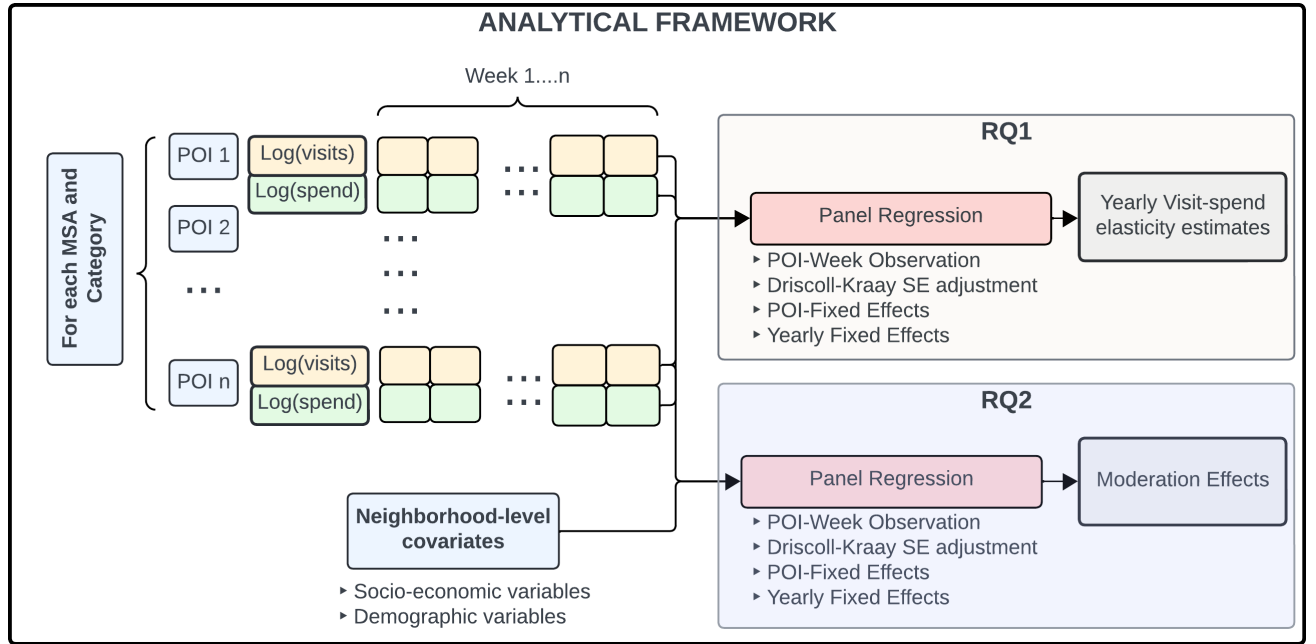


Figure 5.1: Analytical framework for estimating visit–spend elasticities and moderation effects. For each MSA and business category, we construct a weekly POI-level panel with $\log(\text{visits})$ and $\log(\text{spend})$. RQ1 estimates yearly visit–spend elasticities using POI-week observations with POI fixed effects, yearly fixed effects, and Driscoll–Kraay standard errors to account for cross-sectional(spatial) & temporal dependence, as defined in eq. (5.1). RQ2 extends this specification by interacting visit–spend relationships with neighborhood socio-demographic characteristics to quantify moderation effects; the full specification is provided in eq. (5.8).

the logarithm of weekly visitor counts to the same establishment. This logarithmic specification serves two important purposes: first, it allows us to interpret our coefficients as *elasticities* – a fundamental economic concept that measures the percentage change in one variable (spending) that results from a 1% change in another variable (visits) [130]. For example, an elasticity of 0.8 indicates that a 10% increase in visits corresponds to an 8% increase in spending. Second, logarithmic transformation helps address the right-skewed distribution of both spending and visit variables, improving the statistical properties of our estimations. This elasticity measure is particularly valuable for economic analysis because it provides a standardized metric that allows meaningful comparisons across different business categories, time periods, and geographic markets, regardless of the underlying scale of spending

or visit volumes. The framework is summarized in Figure 5.1.

5.3.1 Data Sources and Preprocessing

Our empirical analysis leverages a comprehensive dataset that uniquely enables the systematic examination of visit-spend elasticity patterns across business categories, time periods, and demographic contexts. The dataset combines high-frequency consumer visits to businesses with corresponding spending data, allowing us to directly assess the temporal stability of these relationships (**RQ1**) and examine how neighborhood demographic characteristics moderate elasticity patterns across different business types (**RQ2**).

We integrate three primary data sources to construct our analytical dataset. (1) SafeGraph Spend(<https://docs.safegraph.com/docs/spend>) provides daily aggregated and anonymized credit and debit transaction data associated with specific points of interest (POIs), capturing the spending component of our visit-spend relationship. (2) Cuebiq(<https://docs.spectus.ai/>) contributes fine-grained aggregated mobility data from anonymized users who have opted-in to provide access to their location data anonymously, through a CCPA and GDPR-compliant framework. Through its Social Impact program, Cuebiq provides mobility insights for academic research and humanitarian initiatives. The Cuebiq responsible data sharing framework enables research partners to query anonymized and privacy-enhanced data, by providing access to an auditable, on-premise Data Cleanroom environment. All final outputs provided to partners are aggregated in order to preserve privacy. From this mobility data, we derive daily visit counts to POIs using visit attribution methods detailed in Algorithm 1, which are then up-scaled to population estimates by applying a county-level adjustment based on the Cuebiq device sample. (3) SafeGraph Places(<https://docs.safegraph.com/docs/places>) provides daily aggregated and anonymized location data associated with specific points of interest (POIs), capturing the visit component of our visit-spend relationship.

`//docs.safegraph.com/docs/places`) supplies comprehensive POI geometry, identification and categorization, enabling precise business type classification essential for our category-specific analysis.

To address potential measurement error from systematic reporting delays documented by SafeGraph – particularly weekend spending data often attributed to Monday – we aggregate daily observations into weekly panels using custom temporal boundaries running Wednesday to Tuesday rather than standard calendar weeks. This aggregation preserves the high-frequency nature of our data while ensuring robust measurement of the visit-spend relationship across our 313-week panel spanning January 2019 through December 2024. This temporal coverage is crucial for RQ1, as it allows us to evaluate how visit-spend relationships evolved during unprecedented economic shocks and whether categories maintained stable elasticity patterns or experienced systematic disruptions that would compromise their reliability as economic indicators.

Our sample encompasses twelve representative business categories that capture the breadth of consumer economic activity, organized into four theoretically-motivated groups based on consumer engagement patterns and spending behaviors: (1) **Essential Goods**: *Grocery and Convenience Stores*, (2) **Routine Services**: *Personal Care Services, Dry Cleaning and Laundry Services*, (3) **Social & Experiential**: *Restaurants and Other Eating Places, Drinking Places, Specialty Food Stores*, and (4) **Discretionary Retail**: *Clothing Stores, Shoe Stores, Electronics and Appliance Stores, Home Furnishings Stores, Sporting Goods and Musical Instrument Stores, Department Stores*. This categorization reflects distinct consumer trip purposes, purchase patterns, and price points that theory suggests should exhibit different visit-spend elasticity patterns. The diversity across categories enables our comparative analysis of which business types serve as the most reliable economic proxies (RQ1), while the variation in discretionary versus essential spending allows examination of how demographic characteristics

differentially moderate these relationships (RQ2).

We focus our analysis on **six** major metropolitan statistical areas (MSAs) representing diverse economic structures and geographic regions: Chicago-Naperville-Elgin (**CHI**), Houston-Pasadena-The Woodlands (**HOU**), Atlanta-Sandy Springs-Roswell (**ATL**), Washington-Arlington-Alexandria (**WAS**), Miami-Fort Lauderdale-West Palm Beach (**MIA**), and Philadelphia-Camden-Wilmington (**PHI**). This geographic diversity provides substantial variation in economic bases, regulatory environments, and demographic compositions, enabling identification of generalizable patterns in visit-spend relationships while testing robustness across different metropolitan contexts.

To enable our demographic moderation analysis (RQ2), we integrate detailed neighborhood socioeconomic characteristics from the American Community Survey (ACS) five-year estimates. Each POI is matched to census block group (CBG) demographic characteristics, including median household income percentile, educational attainment (percentage with bachelor's degrees or higher), and racial/ethnic composition.

To ensure data quality and sufficient within-establishment variation for reliable analysis, we implement three sequential quality filters.

1. Minimum 50 total observations per POI across the entire study period
2. Presence in at least 4 out of 6 years (2019-2024) to enable proxy validity analysis
3. Minimum 10 POI-week observations per active year when POI is present

Our final analysis dataset comprises **12.5 million POI-week observations** from **52,051 unique POIs** across the six metropolitan areas and twelve business categories, as detailed in Table 5.1. This structure yields 72 separate MSA-category combinations for analysis (6 MSAs $\times$ 12 categories). Detailed statistics of demographic distribution is provided in Table C.1.

Business Category	CHI	HOU	ATL	MIA	WAS	PHI	Total
Essential Goods							
Grocery/Convenience	323,667 (1,201)	150,549 (630)	144,809 (544)	217,368 (779)	266,519 (958)	221,658 (813)	1,324,570 (4,925)
Routine Services							
Personal Care	261,610 (1,314)	177,168 (942)	186,383 (925)	153,320 (787)	154,274 (773)	160,341 (810)	1,093,096 (5,551)
Drycleaning/Laundry	24,475 (122)	16,523 (89)	7,227 (38)	12,451 (62)	15,826 (79)	5,440 (31)	81,942 (421)
Social & Experiential							
Restaurants	1,964,557 (7,992)	1,406,618 (5,633)	1,354,348 (5,350)	1,127,195 (4,609)	1,072,619 (4,473)	938,458 (3,951)	7,863,795 (32,008)
Bars/Pubs	60,852 (296)	40,868 (192)	28,822 (122)	25,702 (123)	25,360 (124)	23,893 (118)	205,497 (975)
Specialty Food	27,859 (132)	24,968 (115)	21,705 (102)	16,993 (82)	14,502 (68)	25,728 (117)	131,755 (616)
Discretionary Retail							
Clothing	181,679 (789)	133,452 (572)	121,302 (506)	137,564 (592)	120,958 (522)	125,475 (544)	820,430 (3,525)
Electronics/Appliance	19,269 (81)	11,133 (50)	12,980 (59)	12,766 (58)	17,033 (73)	11,362 (49)	84,543 (370)
Department Stores	57,497 (204)	41,602 (148)	43,626 (159)	48,352 (178)	33,076 (120)	34,161 (124)	258,314 (933)
Home Furnishings Stores	20,226 (82)	14,532 (57)	15,140 (58)	14,685 (60)	17,700 (71)	12,189 (56)	94,472 (384)
Shoe Stores	45,085 (188)	39,402 (159)	35,250 (144)	36,954 (159)	26,668 (116)	29,028 (126)	212,387 (892)
Sporting/Hobby/Musical	85,183 (353)	55,089 (228)	62,940 (255)	45,815 (202)	47,906 (206)	47,960 (207)	344,893 (1,451)
Total	3,071,959 (12,754)	2,111,904 (8,815)	2,034,532 (8,262)	1,849,165 (7,691)	1,812,441 (7,583)	1,635,693 (6,946)	12,515,694 (52,051)

Table 5.1: **Sample Size Distribution by MSA and Business Category after quality filters.** Table shows POI-week observations with number of unique POIs in parentheses. Sample covers 2019-2024 period across *six* major metropolitan areas and *twelve* business categories grouped by economic function.

5.3.2 RQ1: Proxy Validity and Temporal Stability Assessment

To evaluate which business categories serve as the most reliable proxies for economic activity, we assess both the strength of visit-spend relationships and their temporal stability (*proxy validity*) across different economic conditions. The elasticity estimates themselves indicate proxy validity—categories with strong, significant elasticities demonstrating that visit patterns effectively co-move with spending patterns, making them suitable economic indicators. Additionally, we evaluate whether these relationships remain stable over time, which is crucial for reliable economic monitoring. We estimate separate panel regressions for each MSA-category combination, extract elasticity estimates across time and geographic markets, and systematically analyze both their magnitude/significance and their temporal consistency (stability) via the proxy validation metrics. For each category c in MSA m , we estimate:

$$\log(\text{spend}_{i,t}) = \alpha_i + \beta_1 \log(\text{visits}_{i,t}) + \sum_{\tau=2020}^{2024} \gamma_{\tau} [\log(\text{visits}_{i,t}) \times \mathbf{1}[\text{Year} = \tau]] + \lambda_{\tau(\text{year})} + \varepsilon_{i,t} \quad (5.1)$$

where α_i represents POI fixed effects for POI i , β_1 captures the baseline elasticity using 2019 as the reference year, γ_{τ} measures year-specific deviations from this baseline, and λ_{τ} represents year fixed effects that control for aggregate temporal trends common to all POIs in year τ . The year-specific elasticity for any given year τ is obtained as $\beta_1 + \gamma_{\tau}$, allowing us to trace how the visit-spend relationship evolves within each category. This specification enables us to identify both the average elasticity level and its temporal stability, with particular attention to disruptions during the COVID-19 period and subsequent recovery patterns.

5.3.2.1 Proxy Validation metrics

From the category-specific panel regressions, we extract **five key metrics** to comprehensively evaluate the **proxy quality** of visit-spend elasticities across multiple dimensions: responsiveness strength, temporal consistency or stability, statistical robustness, COVID-19 resilience, and geographic transferability. The metrics are based on the estimated coefficients where β_1 represents the baseline elasticity (2019 reference year) and γ_{τ} captures the year-specific deviation from baseline for year τ .

1. **Responsiveness (Mean Elasticity):** We calculate the average elasticity across all years for each category:

$$\text{Responsiveness} = \frac{1}{T} \left[\beta_1 + (\beta_1 + \gamma_{2020}) + (\beta_1 + \gamma_{2021}) + \dots \right] = \beta_1 + \frac{1}{T} \sum_{\tau=2020}^{2024} \gamma_{\tau}. \quad (5.2)$$

where, $Y = \{2019...2024\}$ and $T = |Y| = 6$.

This metric quantifies how strongly spending responds to changes in foot traffic, with higher values indicating greater economic sensitivity.

2. **Temporal Stability:** We measure temporal stability using the coefficient of variation of year-specific elasticities:

$$\text{Stability} = 1 - \frac{\sigma(\{\beta_1, \beta_1 + \gamma_{2020}, \dots, \beta_1 + \gamma_{2024}\})}{\mu(\{\beta_1, \beta_1 + \gamma_{2020}, \dots, \beta_1 + \gamma_{2024}\})} \quad (5.3)$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote the standard deviation and mean of the elasticity estimates across years, respectively.

3. **Reliability (Statistical Significance Rate):** We calculate the proportion of statistically significant elasticity estimates across all years:

$$\text{Reliability} = \frac{\text{Number of significant elasticity estimates}}{\text{Total number of year-specific estimates}} \quad (5.4)$$

where significance is determined at the $\alpha = 0.05$ level. This metric reflects the statistical robustness of the visit-spend relationship. Categories achieving $\geq 95\%$ significance rates demonstrate excellent reliability, while those below 70% show poor statistical consistency, indicating weak or inconsistent economic relationships.

4. **COVID Disruption:** To assess how the COVID-19 pandemic affected visit-spend relationships, we measure statistically significant changes in elasticities during the acute pandemic period (2020-2021) relative to the pre-pandemic baseline (2019). COVID disruption captures both the magnitude and direction of pandemic-induced changes.

$$\text{COVID Disruption} = \frac{1}{|S|} \sum_{t \in S} [(\beta_1 + \gamma_t) - (\beta_1)] = \frac{1}{|S|} \sum_{t \in S} \gamma_t \quad (5.5)$$

where S is the set of COVID years (2020, 2021) for which the total elasticity shows statistically significant difference from the 2019 baseline, determined by non-overlapping 95% confidence intervals:

$$S = \{t \in \{2020, 2021\} : CI_{95}(\beta_1 + \gamma_t) \cap CI_{95}(\beta_1) = \emptyset\} \quad (5.6)$$

Positive values indicate that COVID strengthened the visit-spend relationship (higher elasticity), while negative values indicate weakening (lower elasticity). The metric is only calculated when the baseline 2019 elasticity is statistically significant and at least one COVID year shows significant difference from baseline. If no statistically significant COVID impact is detected, we set $\text{COVID Disruption} = 0$.

5. Geographic Consistency (Cross-MSA Stability)

To evaluate whether business categories serve as reliable economic indicators across different MSAs, we assess the geographic consistency of visit-spend elasticities. This metric measures how similar each category's temporal behavior is across the six MSAs in our sample, which is crucial for identifying categories that can serve as robust economic indicators regardless of local

market conditions.

For each business category c , we calculate geographic consistency across the set of MSAs where the category meets our reliability threshold (M_c), defined as metropolitan areas where category c has at least 4 out of 6 statistically significant yearly elasticity estimates. Geographic consistency is measured using the coefficient of variation of responsiveness values across qualifying MSAs:

$$\text{Geographic Consistency}_c = \max \left(0, 1 - \frac{\sigma(\text{Responsiveness}_{c,m \in M_c})}{\mu(\text{Responsiveness}_{c,m \in M_c})} \right) \quad (5.7)$$

where $\text{Responsiveness}_{c,m}$ represents the mean elasticity across significant years for category c in MSA m , $\sigma(\cdot)$ denotes the standard deviation across MSAs, and $\mu(\cdot)$ denotes the mean across MSAs. Geographic consistency is only calculated for categories that qualify in at least three MSAs ($|M_c| \geq 3$) to ensure meaningful variation estimates. Higher values indicate greater geographic consistency. Values approaching 1.0 indicate that the category exhibits highly similar visit-spend relationships across metropolitan markets, while values near 0 indicate substantial cross-market variation that may limit the category's utility as a universal economic indicator.

5.3.3 RQ2: Demographic Moderation of Elasticities

Paying attention to the heterogenous nature of responses to economic shocks, it is also important to investigate how neighborhood demographic characteristics influence visit-spend elasticity across different business categories. For that, we employ a demographic moderation framework that estimates category-specific panel regressions with interaction terms. For each category c in MSA m , we estimate:

$$\log(\text{spend}_{it}) = \alpha_i + \gamma_{\text{year}(t)} + \beta_1 \log(\text{visits}_{it}) + \sum_{k \in \mathcal{M}} \theta_k (\log(\text{visits}_{it}) \cdot D_{k,i}) + \varepsilon_{it}, \quad (5.8)$$

where i indexes POIs and t indexes weeks. α_i are POI fixed effects, and $\gamma_{\text{year}(t)}$ are year fixed effects.

The moderator set is

$$\mathcal{M} = \{\text{IncomePct}, \text{BAPct}, \text{BlackPct}, \text{HispanicPct}, \text{AsianPct}\},$$

and $D_{k,i}$ denotes POI i 's (time-invariant) neighborhood value for moderator $k \in \mathcal{M}$.

The POI fixed effects (α_i) absorb all time-invariant characteristics of establishments, including the main effects of neighborhood demographics. Therefore, we can only identify the moderation effects through the interaction coefficients θ_k . Each θ_k represents how the visit-spend elasticity changes with respect to demographic variable k . For example, θ_{BlackPct} (the Black percentage interaction coefficient) indicates how much the visit-spend elasticity increases or decreases for each percentage point increase in Black population share, relative to neighborhoods with higher White population shares. A positive θ_{BlackPct} suggests that POIs in neighborhoods with higher Black population shares exhibit greater spending responsiveness to visitor fluctuations, while a negative coefficient indicates lower responsiveness.

For all panel regressions in RQ1 and RQ2, we use Driscoll–Kraay standard errors as our primary inference procedure. This estimator is robust to heteroskedasticity, serial correlation, and cross-sectional dependence in panel settings, and is well-suited to our high-frequency POI–week data. Within each category–MSA panel, weekly observations for a given establishment are likely to exhibit

temporal dependence, and shocks may be correlated across establishments in the same local market; Driscoll–Kraay standard errors help account for these features of the error structure.

5.4 Analysis of Results

5.4.1 RQ1: Proxy Validity and Temporal Stability Assessment

Using the panel regression framework, we estimate visit–spend elasticities across business categories and years to assess how well mobility serves as a proxy for economic activity. In what follows, we summarize seven key findings on the magnitude and temporal stability of these elasticities, emphasizing which categories and contexts support reliable mobility-based inference and where caution is warranted. Full year- and MSA-specific elasticity estimates are reported in Table C.2, and the proxy validity metrics (responsiveness, stability, reliability, COVID disruption, and geographic consistency) are summarized in Table C.3.

Finding 1: Predominantly positive visit–spend elasticities across MSAs and categories

Figs 5.2 and 5.3 show that, among statistically significant elasticity estimates, the vast majority are positive: increased visitor traffic is generally associated with higher spending across business categories. A few exceptions appear in *Drycleaning and Laundry Services* (e.g., ATL 2019: $\beta = -0.0583, p < 0.05$; PHI 2022: $\beta = -0.0715, p < 0.01$), primarily in combinations with small samples (38 POIs in ATL, 31 in PHI; 7,227 and 5,440 POI-weeks, respectively; see Table 5.1). Overall, this confirms the expectation that visits and spending typically move in the same direction, but the *strength* and *stability* of this relationship vary widely by category. Proxy validity therefore must be evaluated holistically, combining mean elasticities, stability scores, and significance rates: a cate-

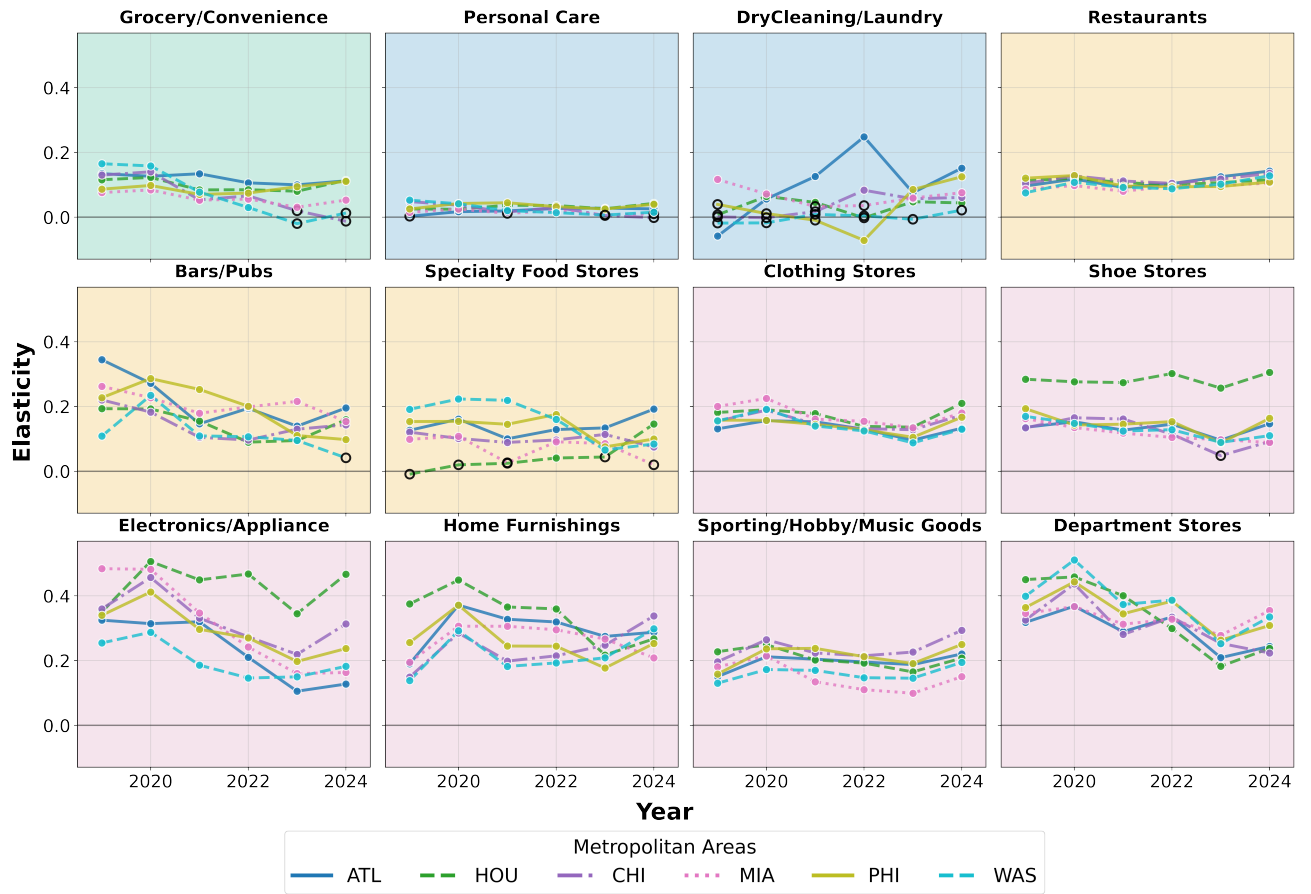


Figure 5.2: **Temporal Evolution of Visit-Spend Elasticities Across Business Categories and Metropolitan Areas (2019–2024).** Year-specific elasticity estimates from Eq. 5.1 for twelve business categories across six MSAs. For 2019, elasticity is given by the baseline coefficient β_1 ; for 2020–2024, elasticities are $\beta_1 + \gamma_\tau$, capturing deviations from the 2019 baseline. Each panel is a business category, with background colors indicating economic type (**Essential Goods**, **Routine Services**, **Social & Experiential**, **Discretionary Retail**). Colored lines denote MSAs; *black outlined circles* denote non-significant estimates ($p > 0.05$).

gory can have high responsiveness but poor temporal consistency, or low responsiveness but reliable stability.

Finding 2: Discretionary retail shows strong visit-spend co-movements with uneven temporal stability

Discretionary retail categories exhibit the strongest visit-spend elasticities. In these sectors, an additional visit tends to be associated with substantial same-week spending, plausibly reflecting larger

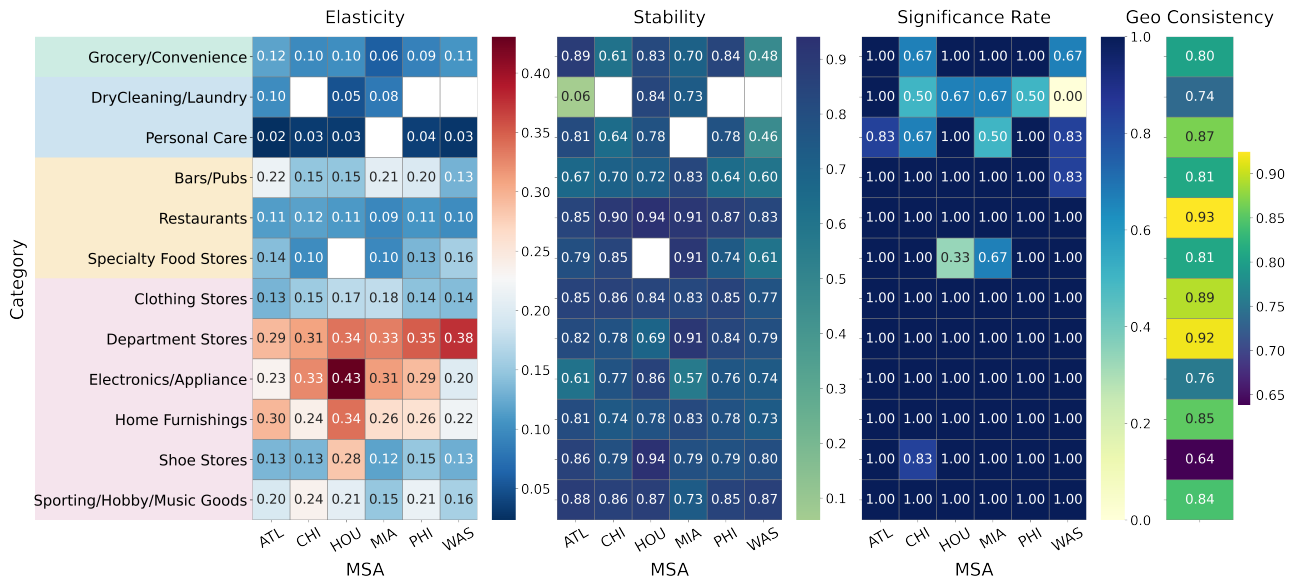


Figure 5.3: **Comprehensive Assessment of Proxy Validity Metrics Across Business Categories and Metropolitan Areas.** Each row is a business category (grouped as **Essential Goods**, **Routine Services**, **Social & Experiential**, **Discretionary Retail**); columns are MSAs (ATL, CHI, HOU, MIA, PHI, WAS). The **Mean Responsiveness** panel shows average elasticities (2019–2024; darker red = stronger relationships). The **Temporal Stability** panel shows the coefficient of variation across years (darker blue = higher stability). The **Significance Rate** panel shows the share of years with statistically significant elasticities (darker blue = more consistently significant). The **Geo Consistency** metric (single value per category, reported in Table C.3 captures cross-MSA agreement. White cells in the Elasticity and Stability panels indicate that significance rates did not meet the inclusion threshold.

baskets and a higher likelihood of completing a purchase when trips occur. However, the temporal stability and cross-market consistency of these elasticities differ across categories.

Within discretionary retail, *Department Stores* are the most responsive (mean elasticities around 0.30–0.38), but stability varies by MSA. *Clothing Stores* combine strong elasticities (roughly 0.13–0.18) with the most stable temporal profiles in this group. *Electronics & Appliances* are highly responsive but the least stable over time, with especially low stability in some markets (e.g., MIA). This pattern is consistent with episodic, big-ticket purchases that are sensitive to showrooming and seasonal dynamics [61]. *Home Furnishings* and *Shoe Stores* also show high responsiveness with moderate-to-

strong stability, though *Shoe Stores* exhibit lower geo-consistency, indicating stronger local context effects in some MSAs. *Sporting Goods/Hobby/Music* occupy a mid-range in responsiveness with generally robust stability.

In summary, discretionary retail provides powerful, high-signal proxies for nowcasting within a given period, but the elasticity itself is less stable over time. *Clothing Stores* and *Home Furnishings Stores* stand out as particularly dependable within this group, while *Department Stores* and *Electronics & Appliances* offer high responsiveness but require more frequent recalibration and careful interpretation, especially for cross-MSA extrapolation.

Finding 3: Restaurants provide consistent and reliable visit–spend elasticities

Restaurants exhibit somewhat lower elasticities than many discretionary categories—consistent with smaller per-visit spending changes than big-ticket retail—but they show the most consistent patterns across all proxy validity metrics. Temporal stability is exceptionally high (with several MSAs exceeding 0.90), and geo-consistency is likewise strong (cross-MSA consistency around 0.93), indicating that restaurant visit–spend relationships are both temporally and spatially robust.

These properties make *restaurant visits* an especially reliable proxy for economic activity. They are suitable for both high-frequency nowcasting and longer-horizon monitoring, and require less frequent recalibration than other categories because their elasticity estimates are stable across time and across markets.

Finding 4: Grocery and Convenience Stores: moderate coupling with uneven stability and reliability

Grocery and Convenience Stores show moderate elasticities (roughly 0.06–0.12), implying more modest spending responses to visit changes than in discretionary retail. Stability and statistical reliability differ sharply across MSAs. PHI and ATL display strong temporal stability with consistently

significant estimates, whereas WAS and CHI show weak stability, with sharp disruptions during the pandemic period (e.g., large 2021 drops, non-significant estimates in 2023–2024). These patterns align with evidence of e-grocery adoption and channel substitution decoupling visits and spending [49, 97].

Despite this heterogeneity, cross-market agreement is reasonably high when estimates are significant (Geo-consistency ≈ 0.80), suggesting that the underlying relationship is similar where it exists. From a policy standpoint, grocery visits can serve as a useful real-time signal in markets with stable elasticities (e.g., PHI, ATL), but should generally be treated as a *supplementary indicator*. Robust monitoring requires validation against traditional economic measures and periodic recalibration, particularly in MSAs with unstable or non-significant estimates.

Finding 5: Routine services provide the weakest and least reliable visit–spend signal

Routine services—*Drycleaning and Laundry* and *Personal Care Services*—provide the weakest visit–spend coupling. *Drycleaning and Laundry* has especially weak elasticities: three MSAs fail our reliability threshold entirely, while the remaining markets show only moderate geo-consistency and include counterintuitive negative elasticities (e.g., PHI 2022, ATL 2019) and low temporal stability. Even when some MSAs (e.g., MIA, HOU) show seemingly stable profiles, these are based on a subset of significant years and must be interpreted with caution.

Personal Care Services fare somewhat better, showing good geographic consistency but the lowest mean responsiveness of any category (elasticities around 0.02–0.04). Temporal stability ranges from high (e.g., ATL) to poor (e.g., WAS), with clear pandemic-era disruptions in several MSAs (WAS and CHI). Business models involving subscriptions, service packages, or prepayment systems might explain why contemporaneous visits and spending show weaker correlation in this category.

For these categories, mobility-based indicators should be used *cautiously*. Small, unstable, and

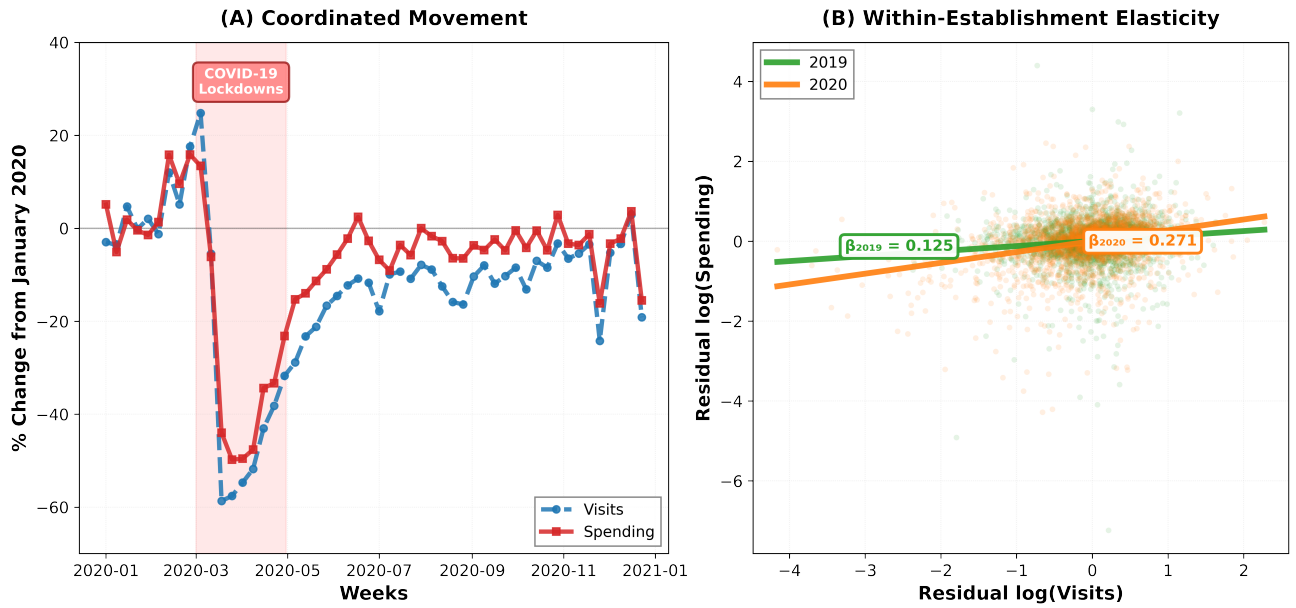


Figure 5.4: **ATL–Restaurants: co-movement and within-POI elasticities.** **Panel A:** Weekly percent change from the January 2020 average in visits (blue, dashed) and card spending (red, solid) for restaurants in ATL; the shaded band marks initial lockdown weeks. **Panel B:** Within-POI relationship between residual log(spend) and residual log(visits) after removing POI and year fixed effects. Lines are OLS fits by year. Estimated elasticities: $\beta_{2019} = 0.125$ and $\beta_{2020} = 0.271$, indicating steeper coupling in 2020. Formal RQ1 estimates from the full specification appear in Table C.2.

sometimes non-significant elasticities imply that visits are a noisy proxy for underlying economic activity. Where used, they should be triangulated with complementary data sources.

Finding 6: Pandemic onset (2020) saw elevated elasticities in selected categories

Relative to 2019, visit–spend elasticities rose sharply in 2020 across many business categories, indicating that a given percentage change in visits translated into larger percentage changes in spending during the onset of COVID-19. Increases were most consistent in discretionary retail (e.g., *Department Stores*, *Home Furnishings*, *Sporting/Hobby/Music*, *Clothing*) and in *Restaurants*, with several other categories (e.g., *Electronics/Appliances*, *Grocery and Convenience*, *Specialty Food*) showing increases in most MSAs. Magnitudes ranged from modest lifts in some grocery markets to large surges in *Home Furnishings*; the complete set of 2019-2020 elasticity changes by category and MSA is reported in

Table C.4.

Fig 5.4 illustrates this mechanism for ATL restaurants: visits and spending collapse and rebound together (Panel A), and the within-POI log–log slope more than doubles between 2019 and 2020 (Panel B). These elevated elasticities likely arise because both visits and spending respond strongly and synchronously to a common shock. However, several category–MSA combinations exhibit *lower* elasticities in 2020 (e.g., *Shoe Stores* and *Bars/Pubs* in multiple MSAs; see Table C.4), despite aggregate co-movement in visits and spending, underscoring that not all sectors respond to shocks in the same way.

For mobility-based monitoring, these findings have two implications. First, for categories with elevated pandemic elasticities, mobility data can be particularly informative during crises when traditional indicators lag, supporting prior work that uses such data for shock analysis [26, 140, 177]. Second, shock-era elasticities should *not* be extrapolated to normal times: doing so would overstate the economic consequences of visit changes in non-crisis periods. Effective monitoring therefore requires estimating and applying separate elasticities for crisis versus normal regimes.

Finding 7: Branded vs. non-branded POIs show heterogeneous elasticity trajectories

Branded and non-branded establishments systematically differ in their operating models, digital infrastructure, pricing strategies, and supply chain arrangements, all of which can affect how changes in visits translate into changes in spending. To test whether this translates into different elasticity paths, we estimate separate elasticities for branded and non-branded *Restaurants* (Table C.5) using the same panel specification as Equation (5.1). We focus on *Restaurants* because this category offers a larger sample and a more balanced split between panel and non-panel establishments.

Looking at Fig 5.5, we see, in 2019, non-branded POIs generally exhibit higher elasticities

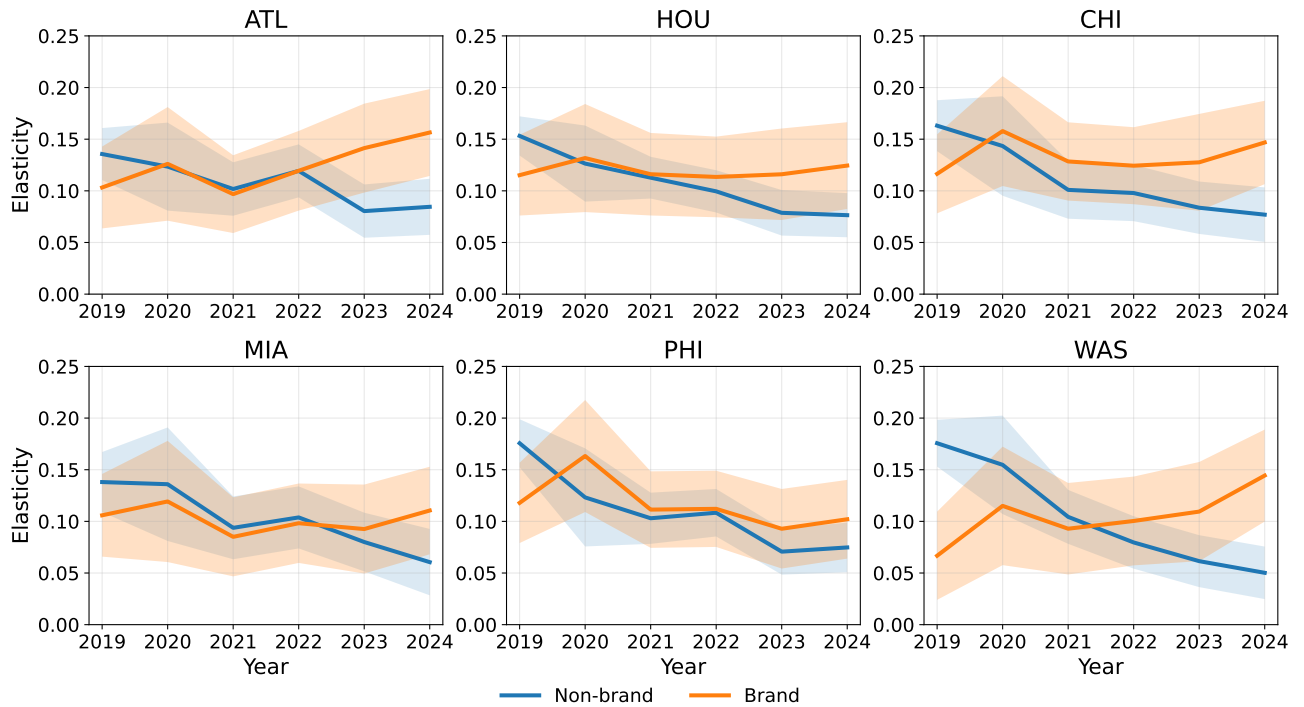


Figure 5.5: **Branded vs. non-branded restaurant elasticities, 2019–2024.** Annual within-POI spend–visit elasticities estimated separately for branded and non-branded restaurants using Eq. 5.1 with POI and year fixed effects; shaded bands show 95% confidence intervals.

than branded ones. In 2020, this pattern reverses: branded elasticities increase in every MSA, while non-branded elasticities decline everywhere. In several markets (e.g., ATL, CHI, WAS), the brand segment’s elasticity rises noticeably while the non-brand segment’s elasticity falls. By 2023–2024, branded and non-branded confidence intervals no longer overlap in ATL, CHI, and WAS, indicating statistically higher elasticities, whereas HOU, MIA, and PHI retain overlapping intervals.

Non-branded elasticities also trend downward over the full period, falling below 2019 levels by 2024 in all MSAs, with mostly monotonic declines in HOU, CHI, and WAS and small temporary rebounds in others. Overall, branded and non-branded restaurants remain similar in magnitude but follow different temporal trajectories, suggesting that chain affiliation can meaningfully alter how visit changes translate into spending during and after shocks.

5.4.2 RQ2: Do neighborhood-level demographic data moderate the visit-spend elasticities?

Figs 5.6 and 5.7 summarize moderation effects across all 72 MSA–category combinations (cells). Each cell reports an interaction coefficient (θ_k) which captures how a one–unit increase in a neighborhood characteristic shifts the baseline visit–spend elasticity (see Equation (5.8)). Positive coefficients (blue) indicate that higher values of the neighborhood characteristics are associated with *stronger* visit–spend relationships (amplification), whereas negative coefficients (red) indicate *weaker* relationships (attenuation). Coefficients are in elasticity units: income is measured in percentile points (0–100), while education and racial/ethnic variables are in percentage points. Thus, a value of 0.002 implies a +0.002 change in elasticity per one-point increase in the moderator (holding other moderators, POI fixed effects, and year effects constant); moving from the 25th to the 75th percentile would therefore raise elasticity by $50 \times 0.002 = 0.1$. In what follows, we discuss broad patterns for socioeconomic (income, education) and racial/ethnic (Black, Hispanic, Asian) moderators; full coefficient tables for all MSA–category cells appear in the Table C.6.

5.4.2.1 Socioeconomic Moderation (Income and Education)

Income moderates visit–spend elasticities in 64% of MSA–category cells (46 of 72), with amplification effects (27 cases) slightly outweighing attenuation effects (19 cases). The strongest and most systematic amplification occurs in discretionary retail. *Department Stores* exhibit uniformly positive income effects where significant, and *Electronics/Appliance Stores* show particularly strong responses (e.g., ATL $\beta = 0.0052$, $p < 0.001$), implying an elasticity increase of about 0.26 when moving

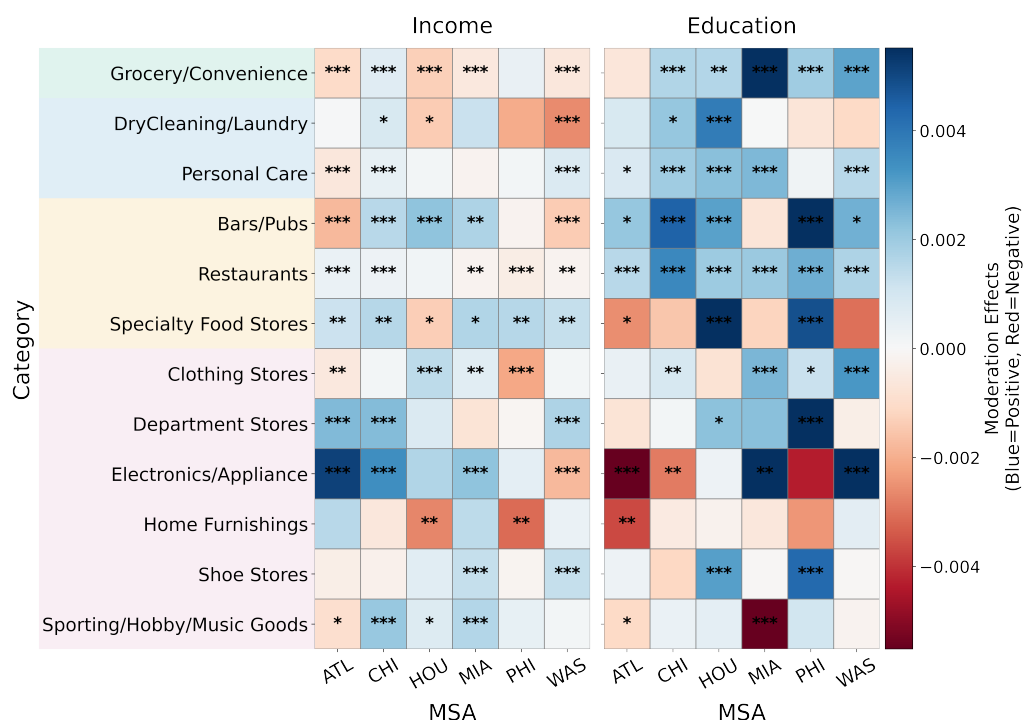


Figure 5.6: **Socioeconomic Moderation of Visit-Spend Elasticities Across Business Categories and Metropolitan Areas.** Each panel displays coefficients for a specific demographic variable: Income (median household income percentile) and Education (percentage with bachelor's degrees or higher). Rows represent business categories organized by economic type: **Essential Goods**, **Routine Services**, **Social & Experiential**, and **Discretionary Retail**. Columns within each panel show the six MSAs. Color intensity indicates coefficient magnitude and direction, with blue shades representing positive moderation effects (higher demographic characteristic associated with stronger visit-spend relationship) and red/orange shades representing the converse. Asterisks denote statistical significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, while cells without asterisks indicate non-significant coefficients.

from the 25th to the 75th income percentile. *Specialty Food Stores* likewise predominantly amplify visit-spend coupling in more affluent neighborhoods.

However, this pattern reverses for *Grocery and Convenience Stores*, which systematically *weaken* visit-spend elasticity at higher income levels: four of five significant income coefficients are negative. This attenuation aligns with post-COVID evidence that online grocery purchases substitute for in-person visits among higher-income households with better access to delivery and digital platforms

[97]. *Restaurants* show mixed income effects that vary by MSA, consistent with evidence that digital food ordering and eating out often act as complements rather than pure substitutes [97]. Overall, income-driven amplification is most consistent in discretionary categories, while essential categories such as grocery demonstrate substitution-driven attenuation.

Education, measured as the percentage of residents with at least a bachelor's degree at the CBG level, presents more consistent moderation than income. Educational attainment moderates elasticities in 57% of cells (41 of 72), with a strong positive skew (35 positive vs. 6 negative significant effects). Education effects cluster strongly in a group of categories where visit–spend relationships become systematically stronger in more educated neighborhoods. *Restaurants* display universal, significant amplification across all six MSAs (peaking in CHI: around $\beta \approx 0.0035$, $p < 0.001$), indicating that spending becomes more responsive to visit fluctuations in higher-education areas. *Personal Care Services* and *Bars/Pubs* show similar patterns, with amplification in most MSAs. Crucially, *Grocery/Convenience* exhibit positive education effects in all MSAs where coefficients are significant, in direct contrast to the predominantly negative income effects for the same category. This suggests that higher education is associated with stronger visit–spend coupling even where higher income is associated with online substitution.

Discretionary retail categories exhibit more mixed education patterns, though *Clothing Stores* are consistently positive across the MSAs where estimates are significant. Metro-level variation in education effects is limited: most MSAs show broadly positive amplification, with attenuations largely confined to a subset of discretionary categories.

5.4.2.2 Racial/Ethnic Moderation (Black, Hispanic and Asian Groups)

Racial and ethnic composition also moderates visit–spend elasticities in systematic ways, with especially strong and pervasive attenuation in neighborhoods with higher Black and Hispanic population shares.

Black population share moderates elasticities in 53% of cells (38 of 72), with a slight negative skew (22 negative vs. 16 positive significant effects). Several categories exhibit weaker visit–spend coupling in neighborhoods with higher Black shares. Most notably, *Grocery/Convenience* attenuate in both of the cells with significant effects, with coefficients ranging from -0.0020 to -0.0013 ; *Drycleaning and Laundry* show similar attenuation. *Restaurants* mostly attenuate, with a small positive exception in ATL. Within Discretionary Retail, *Department Stores* tend to show negative moderation, while *Electronics/Appliances* and *Shoe Stores* sometimes amplify, underscoring that not all categories behave uniformly. At the MSA level, patterns range from predominantly negative profiles to more mixed-to-positive profiles in a subset of markets.

Hispanic population share displays the most consistently negative moderation pattern among all demographics, moderating in 58% of cells (42 of 72) with a pronounced negative skew (32 negative vs. 10 positive significant effects). Category-level patterns in Fig 5.7 show consistent attenuation patterns. *Restaurants* exhibit universal negative coefficients, in sharp contrast to the uniformly positive education effects for the same category. *Grocery/Convenience Stores* and *Specialty Food Stores* are also uniformly negative wherever significant. Most other categories tilt negative across MSAs, with only a small set of positive exceptions (notably in *Electronics/Appliance Stores* and a few instances in *Bars/Pubs* and *Sporting/Hobby/Music Goods*). Metro-level profiles are broadly net negative, indicating that neighborhoods with higher Hispanic shares consistently exhibit weaker visit–spend coupling.

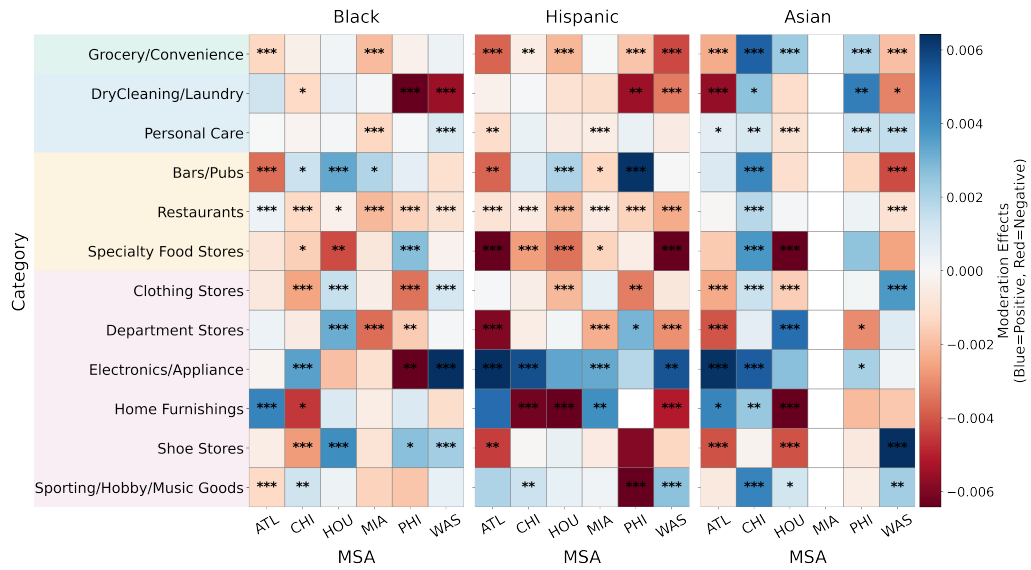


Figure 5.7: Racial Moderation of Visit-Spend Elasticities Across Business Categories and Metropolitan Areas. Each panel displays coefficients for a specific demographic variable: Racial/Ethnic Composition (percentage Black, Hispanic, and Asian populations, with White population as the reference category). Rows represent business categories organized by economic type: **Essential Goods**, **Routine Services**, **Social & Experiential**, and **Discretionary Retail**. Color intensity indicates coefficient magnitude and direction, with blue shades representing positive moderation effects (higher demographic characteristic associated with stronger visit-spend relationship) and red/orange shades representing the converse. Asterisks denote statistical significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, while cells without asterisks indicate non-significant coefficients.

Asian population share exhibits the most geographically polarized moderation pattern. It is significant in 65% of available cells (39 of 60, with MIA excluded due to low variation) and is roughly tilted towards amplification (24 positive vs. 15 negative). Unlike Hispanic shares, which lean clearly negative, Asian-share effects are highly context-dependent. Some categories, such as *Grocery/Convenience*, *Personal Care*, and *Home Furnishings*, tend to show positive moderation in several MSAs, while others (e.g., *Shoe Stores*) lean negative. Metro-level profiles range from mostly positive in some MSAs to predominantly negative in others, underscoring that the direction and strength of Asian-share moderation depends strongly on local context and business category.

Implications for Using Mobility-Based Indicators

Taken together, these socioeconomic and racial/ethnic moderation patterns demonstrate that mobility-based indicators can systematically misrepresent business health across neighborhoods. In particular, attenuated visit–spend elasticities in neighborhoods with higher Black and Hispanic population shares imply that a given decline in visits corresponds to a *smaller* proportional decline in spending than in other areas. For example, *Grocery/Convenience Stores* in higher-Black neighborhoods show consistent attenuation, and *Restaurants* exhibit uniformly weaker elasticities in higher-Hispanic neighborhoods. Applying a single, average elasticity across all neighborhoods would therefore *overstate* spending losses and *understate* true activity in these communities, risking misdiagnosis and misallocation of recovery resources. Monitoring frameworks should incorporate category-specific demographic adjustment factors and be localized by MSA, and—where feasible—visit–spend elasticities should be periodically validated and recalibrated against card-spend or other ground-truth data as new information becomes available.

5.5 Discussion

The central aim of this study was to assess whether high-frequency mobility data reliably mirrors economic activity. We first establish a predominantly positive visit–spend elasticity across most business categories and metropolitan areas, confirming the basic intuition that more footfall is usually associated with higher spending. Within this general pattern, our most robust result is the exceptional reliability of *Restaurants* as an economic proxy: this category exhibits moderate elasticities that are highly stable over time and consistent across MSAs. A likely mechanism is a high visit-to-purchase conversion rate; unlike many retail categories, restaurant visits almost always culminate in a transaction.

By contrast, this reliability does not extend uniformly to other categories. We uncover a sharp divide between discretionary retail and routine services. Discretionary categories such as *Department Stores* and *Electronics/Appliances* show very high visit–spend elasticities, but these relationships are comparatively volatile. This instability is consistent with the growth of non-transactional behaviors (e.g., showrooming, window shopping) and seasonal purchasing dynamics that decouple visits from immediate purchases [21, 61]. For policymakers, these categories are therefore powerful for within-period nowcasting but require periodic recalibration for longer-horizon tracking as the elasticity itself evolves.

The COVID-19 pandemic functions as a clear structural break [16, 46], elevating elasticities across a majority of categories. This surge likely reflects a combination of lockdown-induced hoarding [156] and tightly synchronized collapses and rebounds in both visits and spending during the shock period. Mobility-based indicators thus appear especially informative during crises, when visits and spending move in tandem. However, our results also imply that elasticities estimated in shock-era data may be inflated relative to “normal” conditions, underscoring the need to recalibrate visit–spend relationships as economies transition into post-shock regimes.

A central contribution of our work is to challenge the assumption that visit–spend relationships are demographically uniform. In our RQ2 results, we show that elasticities are significantly moderated by neighborhood characteristics, with neighborhoods containing higher shares of Hispanic and Black residents exhibiting consistent attenuation across many categories. In practical terms, an identical change in visits can correspond to very different spending responses across neighborhoods—for instance, between areas with high Asian population shares and those with high Hispanic shares. This directly undermines “one-size-fits-all” applications of mobility data and shows that demographic context is not a second-order nuance, but a first-order determinant of how well visits track economic

activity.

However, these conclusions remain subject to several important limitations. First, we analyze nominal card spending without explicitly adjusting for inflation. Year fixed effects absorb common *annual* shifts in price levels and macroeconomic conditions, but they cannot capture *within-year* inflation or relative price changes across categories and locations. As a result, some time-variation in elasticities may partly reflect changing prices rather than purely behavioral shifts.

Second, our outcome measure captures only electronic credit and debit card transactions. The DCPC [67] provides national, *annual* payment shares for credit and debit usage, but does not offer the POI- and week-level granularity needed to impute cash usage at our resolution. In our specification, POI fixed effects absorb time-invariant differences in cash intensity across establishments, and year fixed effects capture broad shifts in payment method adoption. However, within-year changes in card adoption at the POI or neighborhood level—especially if correlated with visits—are not fully controlled. Although electronic payments dominate in metropolitan areas [67], our estimates may understate true economic activity in cash-intensive businesses or areas with lower electronic payment penetration. We therefore treat the omission of cash as a scope limitation: the elasticities we report characterize *electronic spending* responsiveness to visits.

Third, our measurement of visits from mobility data is inherently imperfect. Complex geographies (e.g., malls, food courts, multi-tenant buildings) and behaviors such as curbside or parking-lot pickup can generate misattribution, leading to false positives or missed visits. Our visit-attribution algorithm (see Algorithm 1) addresses part of this problem, but residual error likely remains.

Finally, because SafeGraph Spend is derived from a permissioned cardholder panel rather than a census, the geographic density of panel customers may exhibit modest over- and under-indexing that varies across states and time. To test this, in Section C.2, we benchmark state-level panel customer

counts against ACS annual population estimates and find consistently high correlations ($r \geq 0.82$ in all months), indicating that panel density closely tracks population density overall, while still exhibiting modest state-month over- and under-indexing reflected in the distribution of absolute state bias (See Figure C.1 & Figure C.2. Developing and validating a reweighting scheme would require external ground truth on how panel coverage maps to merchant revenue across states, categories, and time. We treat this as an important direction for future work and, in the interim, provide diagnostics showing that panel density closely tracks population density overall.

Chapter 6: Evaluating LLM Priors About Neighborhood-Level Mobility

6.1 Introduction

Human mobility is a fundamental dimension of urban life, reflecting how people access opportunities, navigate infrastructure, and experience cities in everyday life [18, 58, 90, 181]. Because movement patterns shape access to jobs and services, pandemic spread, environmental burdens, and broader forms of social interaction, mobility data have become central to understanding cities and informing urban policy [30, 42, 58, 176, 179]. The growing availability of passively collected digital traces, especially smartphone- and mobile-phone-based location data, has substantially expanded this potential. Compared with traditional travel surveys [28] and other coarse data sources, such data provide much finer spatial and temporal resolution, often capturing behavior continuously over longer periods and at much larger scales [77, 83]. This has enabled researchers to characterize everyday activity patterns, travel extent, accessibility, and behavioral heterogeneity with unprecedented detail, greatly enlarging the empirical toolkit available to urban science and policy research.

Yet the smartphone-based location data that enable these analyses are largely proprietary, access-restricted, and subject to privacy constraints that limit their availability to many researchers and public agencies, raising the question of whether models that encode broad knowledge about cities and human behavior could serve as useful alternatives in data-scarce settings. Large language models (LLMs) are a natural candidate for this role. Trained on vast corpora encompassing urban descriptions, census

and demographic data, transportation research, and accounts of everyday life across cities, LLMs may encode the kinds of contextual knowledge that underlie mobility patterns. Recent studies suggest they do. Prior work has used LLMs for tasks such as next-location prediction and zero-shot mobility inference, showing that LLMs can sometimes act as mobility predictors even without task-specific training [22, 121, 167]. Other work has explored whether prompted LLMs can generate synthetic travel-survey responses for urban mobility assessment, producing plausible diary-like behavior from background knowledge alone [24]. More broadly, benchmarks such as CityBench evaluate LLMs across a range of urban tasks and show that, while advanced models can be competitive on tasks grounded in commonsense and semantic understanding, they remain less reliable on tasks requiring stronger urban reasoning [64]. However, all of these studies share a common target of evaluation, assessing whether LLMs can produce plausible outputs about individual-level mobility, such as a person’s next trip, a synthetic travel diary, or a single device’s behavioral signature.

This individual focus is valuable in its own right, but it does not address the forecasting needs of many real-world systems, where the key quantity of interest is aggregate mobility patterns. Transportation agencies, urban planners, emergency managers, and public health officials typically require predictions of origin-destination flows, demand surges, and population redistribution patterns rather than individual trajectories [147, 186]. Aggregate flow prediction is not simply the sum of many next-location forecasts; it reflects emergent collective dynamics shaped by network structure, temporal rhythms, external events, and policy interventions. In this work, we position LLM-based aggregate mobility pattern forecasting as a distinct and underexplored problem. While LLMs have demonstrated proficiency in mining individual histories, it remains an open question whether they can effectively interpret contextual urban inputs to predict aggregate, neighborhood-level mobility patterns.

In addition, even if LLMs can recover aggregate neighborhood-level outcomes, prediction ac-

curacy alone does not establish that they have learned the empirical relationships that structure mobility across places. Plausible performance may instead reflect shallow heuristics, broad stereotypes [123, 128], or memorization [91] of patterns seen during pretraining, all of which could align with outcomes on average while diverging from the specific relationships that actually structure how context shapes mobility. Thus, whether LLM predictions recover the structural relationships between neighborhood context and travel patterns that empirical data reveal remains an open question.

To address these gaps, we evaluate whether LLMs encode useful priors about urban mobility by testing their ability to infer neighborhood-level mobility outcomes from urban context, assessing the structural alignment of those predictions with empirical context-mobility relationships, and examining whether this alignment holds across cities. Our analysis is grounded in large-scale, real-world mobility trajectories from the United States, provided by Spectus, which amasses anonymized location data from nearly 70 million mobile devices, covering roughly 20% of the U.S. population. We aggregate these data to the Census Block Group (CBG) level and derive multiple mobility indicators capturing three complementary dimensions of behavior: (i) point-level, (ii) trajectory-level, and (iii) temporal-level mobility. Given built-environment and neighborhood sociodemographic characteristics as input, we ask three research questions:

RQ1: To what extent can LLMs predict neighborhood-level mobility outcomes given built-environment and sociodemographic context?

RQ2: To what extent do LLM predictions align with empirically observed relationships between neighborhood characteristics and mobility behavior?

RQ3: Do the mobility-relevant priors encoded by LLMs transfer consistently across different cities and mobility outcomes?

6.2 Related Work

6.2.1 LLMs for Mobility Generation, Prediction, and Urban Tasks

A growing body of work uses large language models (LLMs) to generate, predict, and simulate mobility and urban phenomena. In mobility modeling, several studies adapt LLMs to represent individual trajectories, activity chains, and travel behavior. Ma et al. [121] propose a foundation model for universal human mobility patterns that fuses cross-domain data and semantically enriches GPS traces with survey-informed knowledge, achieving robust performance in activity inference and POI classification across settings such as Los Angeles and Egypt. Wang et al. [165] introduce LLMOB, an LLM-agent framework that generates personal mobility trajectories by extracting patterns from historical traces and reasoning about daily motivations, producing realistic movement sequences even during abnormal periods such as the pandemic. Shao et al. [155] similarly explore the use of LLMs for human mobility generation through context-aware reasoning, emphasizing that language models can explain the intentions behind movement segments rather than merely imitate observed sequences. [114] propose GEO-LLAMA, a fine-tuned framework for generating mobility trajectories under explicit spatiotemporal constraints, showing that LLMs can produce coherent synthetic movements while satisfying visit-level requirements.

Other work focuses on generating richer behavioral records, such as travel diaries and activity schedules. Li et al. [115] develop MOBAGENT, which constructs fine-grained profiles and uses recursive reasoning to generate realistic travel diaries aligned with empirical mobility patterns and road-network constraints. Bhandari et al. [24] evaluate whether LLMs can generate synthetic travel survey data resembling the National Household Travel Survey (NHTS), finding that fine-tuned models

effectively capture complex activity chains and transition probabilities across major U.S. metropolitan areas. Liu et al. [118] present a retrieval-augmented framework with a feedback loop for generating daily activity chains under limited information, showing that the approach can reproduce coordinated household behaviors such as joint shopping trips or shared meal times. Extending this direction, Liu et al. [117] studies discrete travel demand modeling with persona-conditioned LLMs, demonstrating that behavioral traits and socioeconomic context improve simulation of individual travel choices.

A related stream uses LLMs for mobility prediction and recommendation rather than full-sequence generation. Li et al. [113] propose LLM4POI, a next-POI recommendation framework that fine-tunes LLMs on check-in data and combines this with key-query similarity to retain collaborative and contextual information. Gong et al. [82] introduce MOBILITY-LLM, a reprogramming framework that uses behavioral prompts and semantic location representations to improve next-location prediction, arrival-time estimation, and user-link prediction. Chen et al. [45] develop QT-MOB, which introduces semantic location tokenization to encode coordinates as semantically meaningful discrete tokens, substantially improving next-location prediction and trajectory recovery. At the urban scale, Li et al. [116] present URBANGPT, which integrates spatiotemporal encoding with instruction tuning to align urban interdependencies with LLM reasoning, demonstrating strong zero-shot generalization across prediction tasks such as taxi flows, bike flows, and crime rates.

Taken together, this literature shows that LLMs can serve as flexible models for synthetic mobility generation, travel simulation, recommendation, and urban prediction. It also suggests that these models encode nontrivial mobility-relevant signals. There are a few aggregate-oriented examples: URBANGPT studies regional spatiotemporal prediction tasks such as taxi flows and bike flows, and Bhandari et al. [24] evaluate generated surveys using aggregate pattern-level mobility metrics. Still, the

dominant targets of evaluation remain individual trajectories, travel diaries, next-location prediction, or task-specific urban outputs rather than direct recovery of neighborhood-level mobility indicators from contextual urban features, which we address in RQ1.

6.2.2 Evaluating Geospatial and Mobility Knowledge in LLMs

A second line of work asks what kinds of geographic and mobility knowledge LLMs already encode. Manvi et al. [122] introduce GEO-LLM, which fine-tunes LLMs on reverse-geocoded map data and shows that the resulting models can accurately predict geospatial socioeconomic indicators such as population density and asset wealth. Luo et al. [120] propose TSI-LLM, a framework for trajectory semantic inference that uses context-rich prompting and chain-of-thought reasoning to infer occupation categories, activity sequences, and detailed textual descriptions from raw movement traces. Asano et al. [11] introduce MOBQA, a benchmark of 5,800 question-answer pairs for evaluating LLMs on mobility-related factual retrieval, semantic inference, and interpretive explanation. Their results suggest that while models perform well on straightforward factual extraction, they struggle with more complex semantic reasoning over long GPS trajectories.

This evaluation-oriented literature also highlights important limitations and biases in LLMs’ geographic knowledge. Manvi et al. [123] show that although LLMs can make reasonably accurate zero-shot geospatial predictions for objective topics, they exhibit systematic geographic bias on subjective judgments, rating wealthier regions more favorably on attributes such as intelligence or work ethic. Wu and Wang [175] probe demographic and geographic biases in LLM-predicted POI visits, finding that models reflect and amplify race and gender stereotypes in the types of places they associate with different groups. Moayeri et al. [128] further demonstrates broad disparities in factual

recall through WORLDBENCH, showing that state-of-the-art LLMs perform substantially worse for countries in low-income and non-Western regions than for Western countries.

Recent work has also begun evaluating LLMs as broader models of urban knowledge rather than only as task-specific predictors. Zhang et al. [184] introduce AI4US, which tests whether LLMs can reproduce foundational urban-science relationships such as scaling laws, distance decay, and urban vitality. They find that LLMs often recover broad aggregate patterns with high fidelity, but also oversimplify urban complexity, showing limited diversity and weaker causal depth than real urban systems.

Our RQ2 work is most closely related to this evaluation-oriented literature and extends it in a new direction. Rather than examining factual geospatial recall, mobility question answering, or broad urban-theory reproduction, we probe whether LLMs encode empirically meaningful priors about neighborhood-level mobility behavior from built-environment and sociodemographic context.

6.3 Method

6.3.1 Problem Formulation

Let i index Census Block Groups (CBGs), let $\mathbf{x}_i \in \mathbb{R}^p$ denote the vector of contextual predictors for CBG i , and let $m \in \mathcal{M}$ denote a mobility outcome. For each outcome m , let $z_i^{(m)} \in \mathbb{R}$ denote the continuous CBG-level mobility value derived from aggregated individual mobility traces. We formulate CBG-level mobility outcome prediction as an outcome-specific three-class classification task in which the goal is to infer a discretized mobility label

$$y_i^{(m)} \in \{\text{low}, \text{neutral}, \text{high}\} \quad (6.1)$$

from neighborhood context $\mathbf{x}_i$. More formally, for each outcome m , we evaluate a predictor

$$F^{(m)} : \mathbf{x}_i \mapsto y_i^{(m)}. \quad (6.2)$$

Here, $F^{(m)}$ may represent either an LLM-based zero-shot predictor, which reasons over textual descriptions of neighborhood context and outcome semantics, or a supervised baseline trained directly on labeled CBG examples. This formulation allows us to test whether socio-demographic and built-environment context alone is sufficient to recover structured variation in aggregate mobility behavior across neighborhoods.

Our analysis considers two complementary evaluation objectives. The first is *predictive performance*: whether a model can correctly classify the mobility level of a CBG for a given outcome m . The second is *directional alignment*: whether the model captures empirically meaningful relationships between contextual predictors and mobility outcomes. Specifically, for contextual predictor x_{ij} and mobility outcome m , we assess whether the directional implication of the model’s reasoning agrees with the empirical direction estimated from observed data. This second objective is important because accurate class prediction alone does not establish that a model has recovered the underlying monotonic structure relating urban context to mobility behavior.

The remainder of this section presents the methodological components required to instantiate this formulation. We first describe how CBG-level mobility outcomes are constructed from individual mobility traces, then introduce the prediction framework for both LLM-based and supervised models, and finally present the directional alignment framework used to compare model-implied relationships with empirical patterns in the data.

6.3.2 Mobility Outcome Construction

In this paper, we construct CBG-level mobility outcomes from anonymized individual mobility traces. The goal is to derive aggregate measures that characterize different dimensions of resident mobility behavior and that can be paired with socio-demographic and built-environment predictors at the same spatial scale.

6.3.2.1 From Individual Mobility Data to CBG-Level Aggregates

The mobility data used in this study consist of anonymized individual mobility traces collected over 2021. From these traces, we first derive user-level mobility measures described in [6.3.2.2](#), that summarize annual movement behavior. These measures are computed from observed stops and travel episodes and capture complementary aspects of mobility, including spatial extent, movement structure, and temporal variability.

Because our prediction task is defined at the CBG level, we aggregate the user-level mobility measures to the CBG scale. For each user, we identify the associated home CBG and assign the user-level mobility measures to that area. Then, for each CBG, we summarize the mobility behavior of all associated users to obtain aggregate mobility outcomes representing the typical mobility profile of residents in that CBG.

This aggregation serves two purposes. First, it aligns the mobility outcomes with the same spatial unit used for the contextual predictors in the classification task. Second, it preserves privacy by ensuring that the prediction targets are aggregate behavioral summaries rather than individual trajectories. To improve reliability, we retain only CBGs with sufficient user support in the mobility data. The resulting dataset therefore, consists of robust CBG-level mobility outcomes that serve as the ground-

truth targets throughout the paper.

6.3.2.2 Mobility Outcome Families and Definitions

Following Wu et al. [173] taxonomy of mobility characterization from individual trajectories, we organize the mobility outcomes into three families: **point-level**, **trajectory-level**, and **temporal-level** outcomes. Point-level outcomes summarize the spatial footprint and organization of visited locations; trajectory-level outcomes characterize movement over observed travel sequences; and temporal-level outcomes quantify the temporal regularity of activity patterns.

For each user, we represent annual mobility as an ordered sequence of stops and travel segments,

$$s_1 \xrightarrow{\text{trv}_1} s_2 \xrightarrow{\text{trv}_2} \dots \xrightarrow{\text{trv}_{N-1}} s_N, \quad (6.3)$$

where each stop s_i is associated with a location $\mathbf{x}_i \in \mathbb{R}^2$, a dwell time τ_i , and an observation day d_i , and each travel segment trv_i connecting s_i and s_{i+1} is associated with a travel length L_i and a travel duration a_i . Unless otherwise noted, all outcomes are first computed at the user-year level and then aggregated to the Census Block Group (CBG) level. We further define normalized dwell-time weights as

$$w_i = \frac{\tau_i}{\sum_{k=1}^N \tau_k}. \quad (6.4)$$

(1) Point-level outcomes. Point-level outcomes quantify the extent, dispersion, and spatial organization of visited locations. Let a user have stops indexed by $i = 1, \dots, n$, where stop i has coordinates

s_i and dwell time d_i . Define normalized dwell-time weights as

$$w_i = \frac{d_i}{\sum_{k=1}^n d_k}. \quad (6.5)$$

Stay-points entropy measures how evenly dwell time is distributed across distinct visited locations. Let $\mathcal{L}$ denote the set of unique stay locations, and let p_l be the proportion of total dwell time spent at location $l \in \mathcal{L}$. Then

$$H_{\text{stay}} = - \sum_{l \in \mathcal{L}} p_l \log_2 p_l. \quad (6.6)$$

Higher values indicate that activity time is distributed more evenly across locations, whereas lower values indicate concentration in a small number of places.

Radius of gyration measures the characteristic distance of visited locations from the user's center of activity. Let

$$\bar{\mathbf{x}} = \sum_{i=1}^N w_i \mathbf{x}_i \quad (6.7)$$

denote the weighted centroid, and let $\delta(\mathbf{x}_i, \bar{\mathbf{x}})$ denote the haversine distance between stop i and the centroid. The radius of gyration is

$$r_g = \sqrt{\sum_{i=1}^N w_i \delta(\mathbf{x}_i, \bar{\mathbf{x}})^2}. \quad (6.8)$$

Larger values indicate more spatially dispersed activity patterns.

Convex hull diameter quantifies the maximum spatial extent of the activity space. Let $\mathcal{H}$ denote

the set of vertices of the convex hull formed by the observed stop locations. Then,

$$D_{\text{hull}} = \max_{\mathbf{u}, \mathbf{v} \in \mathcal{H}} d(\mathbf{u}, \mathbf{v}), \quad (6.9)$$

where $d(\cdot, \cdot)$ is the haversine distance. This outcome represents the largest distance between any two locations in a user's activity space.

Ellipse area measures the size of an elliptical approximation of the activity space. Let

$$\mathbf{C} = \sum_{i=1}^n w_i (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^\top \quad (6.10)$$

be the weighted covariance matrix of stop coordinates, and let $\lambda_1 \geq \lambda_2$ be its eigenvalues. The ellipse area is

$$A_{\text{ellipse}} = \pi \sqrt{\lambda_1} \sqrt{\lambda_2} c^2, \quad (6.11)$$

where c is a scale factor. Larger values indicate a broader activity space.

(2) Trajectory-level outcomes. Trajectory-level outcomes characterize mobility over sequences of trips and capture the extent and diversity of travel behavior.

Total travel length measures the cumulative distance traveled over all observed trips. Let trajectory segments be indexed by $t = 1, \dots, T$, and let L_t denote the length of segment t . Then,

$$L_{\text{total}} = \sum_{t=1}^T L_t. \quad (6.12)$$

Higher values indicate more extensive travel over the observation period.

Travel entropy measures the diversity of routes used by a user. Let $\mathcal{R}$ denote the set of unique

route signatures, and let p_r be the proportion of trips corresponding to route $r \in \mathcal{R}$. Then,

$$H_{\text{travel}} = - \sum_{r \in \mathcal{R}} p_r \log_2 p_r. \quad (6.13)$$

Higher values indicate more varied routing behavior, while lower values indicate repetitive travel patterns.

Average duration measures the mean duration of observed trips. Let a_i denote the duration of travel segment trv_i , $i = 1, \dots, N - 1$. Then

$$\bar{a} = \frac{1}{N - 1} \sum_{i=1}^{N-1} a_i. \quad (6.14)$$

Higher values indicate longer average trip durations over the observation period.

(3) Temporal-level outcomes. Temporal-level outcomes characterize the temporal organization and irregularity of mobility behavior.

Daily temporal fragmentation measures within-day variability in dwell durations. For each observed day $d \in \mathcal{D}$, let

$$\sigma_d^2 = \text{Var}\{\tau_i : d_i = d\} \quad (6.15)$$

denote the variance of dwell times across all stops observed on that day. We define daily temporal fragmentation as

$$F_{\text{daily}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sigma_d^2. \quad (6.16)$$

Higher values indicate greater within-day temporal irregularity, reflected in a wider mix of short and long stays across stops within observed days.

These three types of mobility outcomes provide complementary views of aggregate mobility behavior at the CBG level. In the remainder of the paper, we use this organization to structure the prediction tasks, the directional alignment analysis, and the presentation of results.

6.3.3 Mobility Outcome Prediction Framework

We evaluate two classes of predictors for CBG-level mobility outcome prediction: **LLM-based zero-shot predictors** and **supervised baseline models**. Both use the same contextual predictors for each CBG, but they differ in how the mapping from neighborhood context to mobility outcomes is obtained. The LLM setting evaluates zero-shot contextual inference from textual descriptions, whereas the supervised baselines learn the mapping directly from labeled examples.

Recall from Section 6.3.2.2 that each CBG i is represented by a contextual predictor vector $\mathbf{x}_i \in \mathbb{R}^p$, and that for each mobility outcome $m \in \mathcal{M}$, the prediction target is the discretized class label $y_i^{(m)} \in \{\text{low}, \text{neutral}, \text{high}\}$. The prediction task is therefore outcome-specific: for each m , we construct a predictor $F^{(m)}$ that maps neighborhood context $\mathbf{x}_i$ to a three-class mobility label.

6.3.3.1 LLM-Based Prediction

For each CBG i and mobility outcome m , we construct an outcome-specific prompt $P(i, m)$ that describes the local contextual characteristics of the CBG and asks the model to predict the corresponding mobility class (0-shot). Each prompt contains four components: (1) the Core-Based Statistical Area (CBSA) in which the CBG is located; (2) a natural-language definition of the target mobility outcome m ; (3) the CBSA-specific tertile thresholds used to discretize the continuous outcome into low, neutral, and high classes; and (4) a textual profile of the CBG’s contextual predictors. The

associated system prompt is provided in [D.1.1](#)

User Prompt Template

```
You are given Census Block Group (CBG) characteristics for CBSA: [CBSA name].

## Outcome Description
Mobility outcome: [outcome name]
Definition: [outcome definition]

Tertile cutpoints within [CBSA name]: low <= [c1]; neutral ([c1], [c2]);
high > [c2].

## Predictor Descriptor
- Percentiles are computed within the CBSA.
- Predictor buckets correspond to CBSA-wide quintiles: very_low | low | neutral | high | very_high.
- Outcome labels are tertiles: low | neutral | high.

## CBG Features
- [feature 1]; [raw value] (CBSA percentile Pxx, [bucket])
- [feature 2]; [raw value] (CBSA percentile Pxx, [bucket])
- ...
- [feature p]; [raw value] (CBSA percentile Pxx, [bucket])

## Question
Predict the mobility outcome category for this CBG.
A. low
B. neutral
C. high
```

The contextual predictors are expressed in a percentile-based format designed to convey both absolute and relative neighborhood context. For each predictor x_{ij} , the prompt includes its raw value,

its within-CBSA percentile rank r_{ij} , and a quintile-based categorical bucket

$$b_{ij} \in \{\text{very_low}, \text{low}, \text{neutral}, \text{high}, \text{very_high}\}.$$

Thus, each CBG is represented as a structured textual profile summarizing how its socio-demographic and built-environment characteristics compare with those of other CBGs in the same metropolitan area. This prompt design encourages the model to reason over relative urban context rather than relying only on raw feature magnitudes.

The task is outcome-specific. For each mobility outcome m , a separate prompt is constructed using the outcome definition and the corresponding CBSA-specific tertile thresholds, while the contextual predictor profile of the CBG remains fixed. The model is then asked to output one of three class labels, $\{\text{low}, \text{neutral}, \text{high}\}$. In addition to the class prediction, the model is instructed to provide brief reasoning and a ranked assessment of influential predictors.

For comparison, we also train a set of supervised baseline models on the same contextual predictors and outcome labels. Specifically, for each mobility outcome m , we fit the following multiclass classifiers: `logit_multinomial`, `random_forest`, `decision_tree`, `hist_gradient_boosting`, and `xgboost`. These trained baselines provide data-driven reference points for the classification task and allow us to compare zero-shot contextual inference by LLMs against conventional supervised learning approaches trained directly on labeled CBG examples.

6.3.3.2 Evaluation Setup

Prompts are generated for all available CBGs and mobility outcomes. For predictive evaluation, however, we report LLM performance on the same held-out 20% test split used for the supervised

baselines, so that both model classes are evaluated on an identical subset of CBGs allowing a direct comparison between zero-shot LLM-based contextual inference and conventional supervised prediction for CBG-level mobility outcome classification.

6.3.4 Directional Alignment Framework

Since classification accuracy alone does not indicate whether a model captures the empirical structure relating neighborhood context to mobility behavior, we also examine whether correct predictions may still be based on misleading or weakly grounded reasoning. To address this, we introduce a complementary *directional alignment* analysis that evaluates whether the directional implications of model reasoning are consistent with empirical relationships observed in the data.

Empirical Relationships/Priors. Let x_{ij} denote contextual predictor j for CBG i , and let $z_i^{(m)}$ denote the continuous value of mobility outcome m . For each CBSA c , mobility outcome m , and contextual predictor j , we estimate an empirical directional relationship between x_{ij} and $z_i^{(m)}$ using two complementary approaches. First, we fit an ordinary least squares (OLS) regression in which the continuous mobility outcome is regressed on the contextual predictors. The sign of the standardized coefficient for predictor j , denoted β_{cjm}^{OLS} , provides a parametric estimate of direction:

$$\text{dir}_{cjm}^{\text{OLS}} = \text{sign}(\beta_{cjm}^{\text{OLS}}) \in \{-1, +1\}. \quad (6.17)$$

Second, we use the Jonckheere–Terpstra (JT) trend test to assess whether the observed mobility outcome exhibits a monotonic trend across ordered predictor buckets. This provides a nonparametric check on whether the empirical relationship is directionally increasing or decreasing.

User Prompt Template for Directional Alignment

```
You are given Census Block Group (CBG) characteristics for CBSA: {CBSA_name}.

## Outcome Description
Mobility Outcome: [outcome_name]
Definition: [outcome definition]

## Predictor Scale
- Predictor buckets correspond to CBSA-wide quintiles: very_low | low | neutral | high | very_high.
- Percentiles (Pxx) indicate the CBG's standing relative to the entire CBSA.

## Selected Census Block Group Features
- [feature 1]; [raw value] (CBSA percentile Pxx, [bucket])
- [feature 2]; [raw value] (CBSA percentile Pxx, [bucket])
- ...
- [feature p]; [raw value] (CBSA percentile Pxx, [bucket])

## Task
Based on the outcome definition provided, assess the directional push (+, -, 0) for each feature
in this specific CBG.
```

LLM-implied Relationships. We then derive an LLM-implied directional signal from the model outputs. In the prompt, each contextual predictor is described using a quintile-based bucket,

$$b_{ij} \in \{\text{very_low}, \text{low}, \text{neutral}, \text{high}, \text{very_high}\},$$

and the model is additionally asked to provide a ranked assessment of influential predictors together with whether each predictor pushes the mobility outcome downward, upward, or neither. We encode this directional push as

$$p_{ijm} \in \{-1, 0, +1\}, \quad (6.18)$$

where -1 , 0 , and $+1$ denote negative, neutral, and positive influence on outcome m , respectively.

To assess whether the model captures the empirical direction for predictor j , we examine how the model-implied push changes across the ordered predictor buckets. Intuitively, if higher values of predictor j are empirically associated with higher mobility outcomes, then the model should increasingly assign positive pushes as the predictor bucket moves from `very_low` to `very_high`; conversely, if the empirical relationship is negative, then the model should increasingly assign negative pushes. Formally, after orienting the LLM-implied push by the empirical regression sign, we test for monotonic directional agreement across predictor buckets using two-sided Jonckheere-Terpstra (JT) tests. This yields a directional label for each (c, j, m) combination: `aligned_increasing`, `opposite_trend`, `no_clear_trend`.

This analysis allows us to distinguish between two different notions of model quality. Predictive performance evaluates whether a model can correctly classify neighborhood mobility levels, whereas directional alignment evaluates whether the model captures empirically meaningful monotonic relationships between contextual predictors and mobility outcomes. Taken together, these analyses provide a more complete assessment of whether LLMs encode useful priors about urban mobility behavior.

6.4 Experiments

6.4.1 Experimental Setting

6.4.1.1 Data

Our experiments integrate three data sources: large-scale human mobility data, socio-demographic context from the American Community Survey (ACS), and built-environment indicators from the EPA

Table 6.1: Summary of CBG coverage and device counts by core-based statistical area (CBSA).

CBSA	# CBGs	Total devices	Mean devices/CBG	Median [Q1, Q3]
Atlanta–Sandy Springs–Roswell, GA	1051	209099	198.95	146 [89, 266]
Los Angeles–Long Beach–Anaheim, CA	4196	359987	85.79	73 [52, 104]
Miami–Fort Lauderdale–West Palm Beach, FL	1730	273187	157.91	126 [80, 195]
San Francisco–Oakland–Berkeley, CA	1779	134379	75.54	57 [37, 90]

Table 6.2: Coverage statistics across the study CBSAs. The table reports the number of CBGs covered, the total number of devices, and the distribution of devices per CBG.

Smart Location Database (SLD).

Mobility data. We derive the mobility outcomes from anonymized and aggregated mobility traces provided by Spectus¹. The data cover four Core-Based Statistical Areas (CBSAs) in the United States: Atlanta–Sandy Springs–Roswell, GA; Los Angeles–Long Beach–Anaheim, CA; Miami–Fort Lauderdale–West Palm Beach, FL; and San Francisco–Oakland–Berkeley, CA. We use data from calendar year 2021 and retain only users observed on more than 60 days during the year. To ensure reliable CBG-level aggregation, we further restrict the analysis to CBGs with at least 20 devices. Under these criteria, the final dataset covers 8,756 CBGs² and 976,652 devices across the four CBSAs. At the CBG level, device support varies across study areas, with median devices per CBG ranging from 57 in San Francisco to 144 in Atlanta. These mobility traces are used to construct the continuous user-level mobility measures described in Section 6.3.2.2, which are then aggregated to the CBG level and discretized into outcome classes for prediction. Table 6.2 and Figure 6.1 show the distribution of the sample sizes as well as the CBGs covered.

¹<https://docs.spectus.ai/>

²CBGs corresponding to military locations are excluded in accordance with Spectus data-sharing policies.



Figure 6.1: Sample size distributions for CBGs across four CBSAs.

Socio-demographic predictors. We obtain socio-demographic variables from the 2019 American Community Survey (ACS) 5-year estimates. These predictors characterize the population and house-

hold composition of each CBG and include measures such as median income, racial and ethnic composition, age structure, educational attainment, and vehicle ownership. In the prediction framework, these variables serve as contextual signals describing the demographic and socioeconomic profile of each neighborhood. For the OLS, we drop some of the correlated features.

Built-environment predictors. Built-environment variables are derived from the EPA Smart Location Database (SLD), which provides CBG-level indicators related to density, street-network structure, transit accessibility, and land-use mix. Because the raw SLD variables are numerous and highly correlated, we summarize them using principal component analysis (PCA). For each conceptual built-environment domain, we retain the first principal component and use it as a composite index in the downstream analysis. This yields four interpretable built-environment predictors corresponding to density, connectivity, transit access, and land-use mix.

Predictor standardization. To support both the LLM prompting framework and the supervised baselines, all predictors are harmonized at the CBG level. For the LLM-based experiments, each predictor is represented using its raw value, within-CBSA percentile, and a quintile-based bucket. For the supervised baselines, the same underlying predictor values are used directly as numeric features.

6.4.1.2 LLM Prediction Settings

We evaluate all LLMs in a zero-shot setting using the prompt template described in Section 6.3.3. Open-weight models are served with v_{LLM} , while proprietary and larger hosted models are accessed through AWS cloud infrastructure. Across all models, decoding is deterministic with temperature set to 0.0. For each CBG and mobility outcome, we issue one outcome-specific prompt and parse

the response into a class prediction and predictor-level directional assessments. LLM performance is reported on the same held-out 20% test split used for the supervised baselines.

6.4.2 RQ1: Predictive Performance

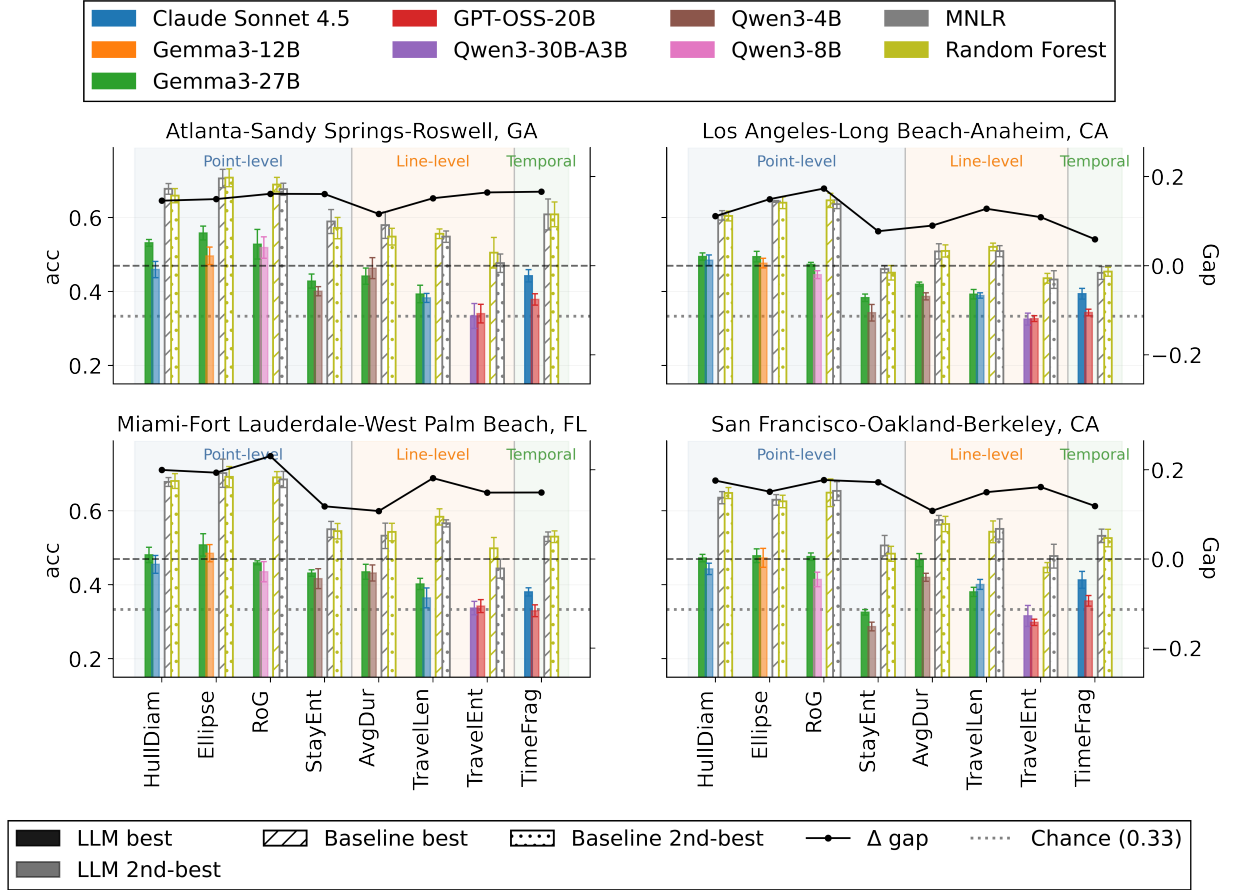


Figure 6.2: Mean 3-class prediction accuracy across four CBSAs for mobility outcomes inferred from socio-demographic and built-environment predictors. Outcomes are grouped into point-level, line-level, and temporal mobility features; bars show the top two LLMs and top two supervised baselines for each outcome, averaged over repeated 20% held-out test splits, with error bars denoting standard deviation across 5 repetitions. The horizontal dashed line marks chance accuracy (1/3), and the black line shows the performance gap between the best baseline and best LLM for each outcome. Across cities, supervised baselines consistently outperform zero-shot LLMs, although LLMs remain above chance on many tasks, with the strongest performance typically observed for point-level mobility outcomes.

RQ1 asks whether LLMs can infer CBG-level mobility outcome classes from contextual urban

information alone, and how their performance compares with supervised baselines trained on labeled data. Figure 6.2 shows the mean 3-class prediction accuracy across four CBSAs for mobility outcomes inferred from socio-demographic and built-environment predictors. As expected, the supervised baselines consistently outperform zero-shot LLMs across all four CBSAs and all mobility outcomes. Nevertheless, the task is clearly learnable from neighborhood context: baseline accuracies are well above chance across outcomes, and the strongest LLMs also exceed the random three-class baseline of 0.33 for most tasks.

A first result is that zero-shot LLM performance is substantially above chance for nearly all outcomes, indicating that LLMs do encode non-trivial prior knowledge about the relationship between neighborhood context and aggregate mobility behavior. This pattern is especially visible for the point-level outcomes, where the best LLMs achieve accuracies in the range of roughly 0.48–0.56 across cities. Performance is weaker for the line-level and temporal outcomes, but still generally above chance. The main exception is *travel entropy*, which is the most difficult outcome for LLMs and remains close to chance in several cities, suggesting that route-diversity behavior is harder to infer from static neighborhood context alone.

Across the evaluated LLMs, Gemma3–27B is the strongest and most consistent model. It is the top-performing LLM for nearly all point-level outcomes (*convex hull diameter*, *ellipse area*, *radius of gyration*, and often *stay-points entropy*) across all four CBSAs, and it also performs best on some line-level outcomes, especially in Los Angeles. The two main exceptions are *travel entropy* and *time fragmented degree*, where other models emerge as stronger: GPT–OSS–20B or Qwen3–30B–A3B for *travel entropy*, and Claude Sonnet 4.5 for *time fragmented degree*. This suggests that while Gemma3–27B provides the strongest overall zero-shot contextual reasoning, performance remains outcome-dependent, and different models appear to have comparative advantages on more behaviorally

complex tasks.

Despite these encouraging LLM results, the gap to supervised learning remains substantial. For every city–outcome pair, the best baseline exceeds the best LLM, with the performance gap typically ranging from about 0.06 to more than 0.20 accuracy points. The largest gaps appear for outcomes such as *radius of gyration*, *time fragmented degree*, and several Miami outcomes, where the best baseline outperforms the best LLM by around 0.17–0.23. Even for the strongest LLM settings, the gap remains non-trivial; for example, in Atlanta the best LLM accuracy for *ellipse area* is 0.558, compared with 0.708 for the best baseline, and in Miami the best LLM accuracy for *radius of gyration* is 0.460, compared with 0.690 for the best baseline. These differences indicate that supervised models trained on labeled examples are much better able to exploit the available contextual predictors than zero-shot LLM prompting.

The strongest supervised baselines are generally multinomial logistic regression and random forest, which appear most often as the top-performing or second-best models across cities and outcomes. Their consistent advantage indicates that the relationship between contextual predictors and mobility outcomes is sufficiently structured to be learned effectively from labeled data. In turn, this strengthens the interpretation of the LLM results: the fact that LLMs perform above chance on most tasks is meaningful because the task itself contains substantial learnable signal, even though LLMs do not recover that signal as effectively as supervised baselines.

Taken together, the results provide a nuanced answer to RQ1. Zero-shot LLMs are able to infer broad mobility patterns from neighborhood context at levels above chance, which suggests that they encode some useful urban and mobility-related priors. However, these priors remain substantially weaker than the predictive relationships captured by supervised baselines trained on labeled CBG examples. Thus, while LLMs show evidence of contextual urban reasoning, they do not yet match

conventional supervised approaches for CBG-level mobility outcome classification.

6.4.3 RQ2: Directional Alignment Results

RQ2 asks whether LLMs recover empirically meaningful directional relationships between neighborhood context and mobility outcomes. Figure 6.3 compares the empirical directions estimated from OLS regression (left panels) with the LLM-based directional alignment results from the Jonckheere–Terpstra analysis (right panels). Overall, the results suggest that LLMs often capture the *broad sign* of context–mobility relationships, but do so in a coarse and simplified way that does not fully reflect metropolitan variation or outcome-specific complexity.

A first pattern is that alignment is strongest for the point-level outcomes. For *convex hull diameter*, *ellipse area*, and *radius of gyration*, the LLM-derived directions are frequently consistent with the empirical coefficients, especially for median income, age 18–29, and several built-environment predictors such as density, transit access, and land-use mix. These variables tend to exhibit positive empirical associations with broader spatial activity spaces, and the LLMs often recover these same positive directions. This suggests that the models do encode a meaningful high-level prior linking neighborhood advantage and accessibility to greater mobility extent.

At the same time, the alignment results reveal a clear tendency toward simplified directional reasoning. Across cities, the LLMs often assign positive relationships to many socioeconomic and built-environment predictors, with relatively limited adaptation across metropolitan areas. In other words, the directional patterns implied by the models are often similar from Atlanta to Los Angeles to Miami to San Francisco, even when the empirical relationships are more context-dependent. This indicates that the models rely on a relatively stable directional template rather than fully city-specific

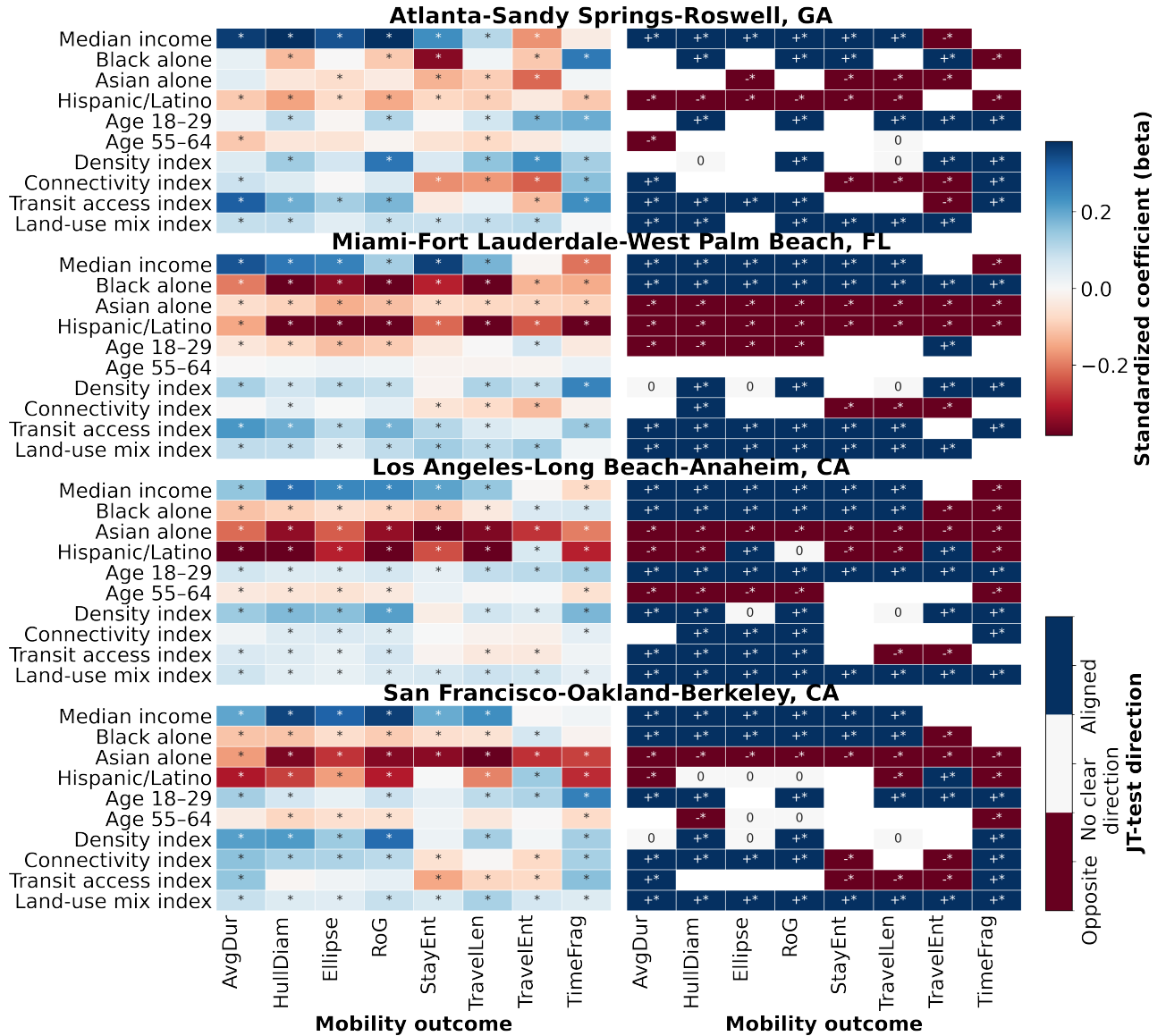


Figure 6.3: Empirical and LLM-implied directional relationships between contextual predictors and mobility outcomes across the four study CBSAs. Left panels show standardized OLS coefficients for each predictor–outcome pair, where red indicates a negative empirical association and blue indicates a positive empirical association; only statistically significant coefficients are marked. Right panels show the corresponding LLM-based directional alignment results from the Jonckheere–Terpstra analysis, restricted to predictor–outcome pairs with significant OLS effects. Blue cells indicate that the LLM-implied direction is aligned with the empirical relationship, red cells indicate an opposite trend, and white cells indicate no clear monotonic trend.

reasoning.

The demographic predictors illustrate this limitation particularly clearly. Black population share is consistently assigned a negative direction by the LLMs and is often aligned with the empirical results. In contrast, the relationships involving Hispanic/Latino and Asian population shares are more frequently misaligned or unstable. These predictors do not exhibit a single uniform pattern across outcomes and CBSAs in the empirical analysis, yet the LLMs appear less able to adapt their directional reasoning accordingly. This suggests that the models are better at recovering broad and persistent associations than at reasoning through demographic relationships that vary more across metropolitan contexts.

A further pattern is that directional alignment weakens substantially for the entropy-based and temporal outcomes. While *stay-points entropy* still shows some meaningful alignment, the models struggle much more on *travel entropy* and *time fragmented degree*. These outcomes measure diversity and irregularity rather than simple mobility magnitude: *travel entropy* reflects variation in route usage, and *time fragmented degree* reflects within-day irregularity in dwell durations. Such properties are harder to infer from static neighborhood context alone, and they appear to be precisely where the LLMs are least reliable. This result is important because entropy-based measures are especially relevant for realistic mobility generation: matching travel extent without capturing diversity, repetition, and temporal fragmentation would still produce behaviorally unrealistic trajectories.

Importantly, the dominant failure mode is usually not a systematic reversal of empirical direction, but rather the absence of a clear monotonic trend. That is, many predictor–outcome pairs are labeled as `no_clear_trend` rather than `opposite_trend`. When the LLMs do recover a directional signal, it is often the correct one; the larger limitation is that this signal is incomplete, overly general, and weaker for more behaviorally complex outcomes.

Taken together, the RQ2 results show that LLMs recover a non-trivial share of the empirical directional structure relating neighborhood context to mobility behavior, particularly for the more spatially interpretable point-level outcomes. However, this capability remains uneven. The models tend to rely on broad, stable directional priors, show limited metropolitan adaptation, and struggle most with entropy-based and temporal outcomes that depend on richer behavioral heterogeneity. Thus, the directional alignment analysis suggests that current LLMs capture coarse urban regularities, but not the full complexity of how contextual factors shape mobility patterns in real trajectory data.

Chapter 7: Conclusion

7.1 Conclusion and Summary of Contribution

This dissertation has argued that behavioral and contextual data must be critically audited for trustworthiness across dimensions of utility, fairness, and validity before being trusted in high-stakes decision-making. The four studies presented here collectively substantiate that argument, showing that the same pattern of uneven, context-dependent, and demographically moderated performance recurs across prediction, fairness, proxy validity, and contextual alternatives alike. The adoption of any data source or model for high-stakes inference should therefore be preceded by systematic evaluation of when it works, for whom it works, and what assumptions it silently encodes.

7.1.1 Study 1: Evaluating the Impact of Mobility Data on COVID-19 Case Prediction.

The value of mobility data for county-level COVID-19 case prediction is highly conditional rather than universal. Mobility data improves prediction performance in at most 60% of counties; in the remaining 40%, adding it actively hurts performance relative to a non-mobility baseline. Where improvements occur, they are modest and most reliably realized at longer forecast horizons. More importantly, where mobility data fails to help, that failure is not random; counties serving higher

proportions of Black, Hispanic, rural, and older residents are systematically less likely to benefit, pointing to underrepresentation of these groups in mobility datasets as the likely mechanism. Linear regression models outperform ARIMAX models on both the proportion of counties that benefit and the magnitude of improvements, and differences across mobility dataset providers are small, meaning the choice of dataset matters less than the decision of whether and when to use mobility data at all.

7.1.2 Study 2: Fairness Audit of the U.S. COVID-19 Forecast Hub

Even when prediction models perform well on average, their errors are not distributed equitably. Hispanic counties consistently exhibit statistically significantly higher prediction errors than majority-White counties across model types, with the disparity reaching approximately 24.6% at 14-day forecast horizons and persisting at 16.8% at 28-day horizons. Less urbanized areas face systematically higher errors than large metropolitan areas, with the disparity worsening as areas become more rural. Critically, the counties where mobility data is least likely to improve prediction in Study 1 are largely the same counties that face the highest prediction errors here; data underrepresentation and model inequity are not separate problems but the same problem observed at different stages of the pipeline. Interaction analyses show that deep learning models produce the smallest disparities across groups, while compartmental and statistical models are associated with the largest; and that the inclusion of mobility data in model training is associated with reduced prediction error disparities, suggesting that behavioral data, despite its own representational limitations, provides information that partially compensates for structural inequities in COVID-19 case reporting.

7.1.3 Study 3: Validity of Foot Traffic as an Economic Proxy

Foot traffic does mirror economic activity, but only contingently. Restaurants emerge as an exceptionally reliable proxy, exhibiting moderate elasticities that are highly stable across years and consistent across metropolitan areas. By contrast, discretionary retail categories show high but volatile elasticities, consistent with the growth of non-transactional behaviors such as showrooming, and routine service categories exhibit weak and unreliable relationships. The COVID-19 pandemic functions as a structural break, elevating elasticities across most categories, with effects persisting differently for branded versus non-branded establishments. Crucially, the demographic dimensions that drove fairness failures in epidemic prediction reappear here as moderators of the visit-spend relationship; elasticities are lower in neighborhoods with higher Hispanic population shares and higher in more affluent and educated neighborhoods. Uncalibrated adoption of foot traffic as an economic signal, therefore, risks systematic misdiagnosis of business health in precisely the communities where accurate monitoring matters most.

7.1.4 Study 4: LLMs as Priors for Urban Mobility

When behavioral traces are unavailable, and large language models are considered as contextual alternatives, they prove informationally non-trivial but structurally incomplete. Supervised baselines consistently outperform zero-shot LLMs across cities and mobility outcomes; however, LLMs perform above chance on many tasks, with the strongest performance observed for point-level mobility outcomes. The directional alignment analysis reveals a critical distinction between instance-level accuracy and structural validity; a model may achieve reasonable classification performance while still failing to recover the empirical ordering between neighborhood characteristics and mobility outcomes.

The mobility-relevant priors encoded by LLMs do not transfer consistently across cities or mobility outcome families, varying substantially with urban context and the dimension of movement being predicted. LLMs, therefore, represent a potentially useful starting point when direct behavioral traces are unavailable, but should not be treated as equivalent to empirically grounded models without careful validation of structural alignment in the contexts of intended use.

7.2 Implications

The findings of this dissertation carry practical implications across the research, data, institutional, and policy dimensions of behavioral data use. First, aggregate evaluation is insufficient as a reporting standard; average performance across regions or populations obscures the conditional and demographically patterned variation that matters most for equitable deployment, and fairness metrics should accompany accuracy as a default rather than an optional addition [1]. Second, the consistent association between demographic underrepresentation and degraded model performance across studies places a concrete obligation on mobility data providers; datasets that systematically undercount minority, rural, and older populations do not merely limit scientific generalizability but actively concentrate inferential harm in the communities least able to absorb it [1]. Third, the fairness analysis in Study 2 relied on regression analysis to evaluate associations between county-level prediction errors and racial and ethnic composition, rather than directly comparing predicted versus actual case counts stratified by race; the latter would have required race-stratified COVID-19 case data at the county level, which was not collected due to longstanding structural gaps in public health surveillance systems. Without such data, fairness audits of epidemic models remain necessarily indirect, and closing this gap requires a sustained institutional commitment to demographic disaggregation in public health reporting, not

merely improved data collection practices. Fourth, future epidemic modeling efforts should treat bias mitigation as a primary design objective; this includes exploring alternative outcome variables such as hospitalization and death counts where available at finer geographic granularities, refining model architectures to reduce disparities across social determinants, and incorporating fairness evaluation into the standard validation workflow of forecasting hubs. Fifth, for practitioners using mobility as an economic proxy, category-specific calibration, neighborhood-demographic adjustment, and temporal recalibration are not optional refinements but necessary conditions for avoiding systematic misdiagnosis of business health in diverse communities. Finally, LLMs should not be adopted as drop-in substitutes for behavioral data without structural validation; producing plausible outputs is not sufficient evidence of empirical grounding, and the inconsistent transferability of their priors across cities and mobility dimensions warrants caution in deployment. Taken together, these implications suggest that the infrastructure for high-stakes data-driven decision-making should include standardized audits of *utility, fairness, and validity* as preconditions for deployment; without such audits, the adoption of behavioral data risks cascading existing inequities through the apparent objectivity of quantitative models.

Appendix A: Analysis of Performance Improvements and Bias Associated with the Use of Human Mobility Data in COVID-19 Case Prediction Models

A.1 Appendix

Table A.1 shows the average performance - measured as average correlation between predicted and actual COVID-19 cases - across lookaheads and linear regression implementations for LTW and STW training-testing approaches, respectively. We can observe that, on average, Ridge is the best performing model for 3 out of the 5 lookaheads; and Elasticnet is the best performing model for 3 out of 5 lookaheads as well. Given their majority best performance, we select this as implementations for the analysis.

Lookaheads	STW				LTW			
	OLS	Ridge	Lasso	ElasticNet	OLS	Ridge	Lasso	ElasticNet
1	0.965098	0.965128	0.962530	0.963475	0.952180	0.952182	0.938414	0.945211
7	0.639711	0.640578	0.644599	0.647571	0.572311	0.572374	0.548012	0.560634
14	0.453315	0.454412	0.462288	0.465076	0.366340	0.366426	0.350444	0.361948
21	0.290965	0.292186	0.308287	0.308882	0.214946	0.215025	0.212894	0.218513
28	0.152364	0.153709	0.177438	0.174648	0.121256	0.121314	0.134021	0.132422

Table A.1: Average p_{corr} value for each regularized method. The best value for each window and each lookahead is bolded.

Features	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org
r2	0.0176	0.0157	0.0173	0.0065	0.0094
adjusted_r2	0.013	0.0117	0.0091	0.0032	0.006
const	0.0585	0.0528	0.0342	0.0411	0.0475
age_65+	-0.0493**	-0.0032	0.0078	-0.0052	-0.0022
low_income	0.0319*	-0.014	0.0097	-0.0007	0.0049
black	-0.0102	0.0343**	-0.0063	-0.0028	-0.0134
hispanic	0.0195	0.023	0.0032	0.0129	0.0021
race_other	-0.0935***	0.0313*	0.0038	-0.0184	-0.0044
rurality	-0.042*	0.0137	0.0046	-0.0093	-0.0035
age_65+:low_income	-0.039**	-0.0049	0.002	-0.002	-0.001
age_65+:black	-0.0006	-0.0042	0.017	-0.0022	-0.0062
age_65+:hispanic	-0.0238**	-0.0096	-0.0156*	-0.005	-0.0035
age_65+:race_other	0.0188	0.0003	0.0139	0.0085*	0.0024
age_65+:rurality	0.0396*	0.0026	-0.0214	0.0143	0.0052
low_income:black	0.0012	0.0259***	-0.0026	-0.0043	-0.0222***
low_income:hispanic	0.0156	0.0013	-0.0179*	0.0083	0.0026
low_income:race_other	-0.0381***	0.0129	0.0035	-0.0033	-0.0051
low_income:rurality	-0.0047	0.0053	-0.0147	-0.0011	-0.0006
black:hispanic	-0.0019	-0.0063*	-0.0055	-0.0025	-0.0077***
black:race_other	0.0019	0.0055*	-0.0023	0.0001	-0.0034
black:rurality	0.017*	-0.0022	-0.0059	0.0054	0.0088
hispanic:race_other	0.0109**	-0.0014	-0.0061	0.001	0.0036
hispanic:rurality	0.0155*	-0.0076	0.0003	0.0003	0.0094
race_other:rurality	0.0409***	-0.0223**	-0.0078	0.0075	-0.0025

Features	(SafeGraph) Restaurants	(SafeGraph) Schools	SafeGraph Inflow	SafeGraph Intraflow	SafeGraph Outflow
r2	0.0112	0.0111	0.0119	0.0083	0.0089
adjusted_r2	0.0079	0.0078	0.0081	0.0051	0.0055
const	0.0489	5.58e-2	0.0565	0.0568	0.0582
age_65+	-0.029*	-0.0075	-0.0298*	-0.0152	-0.0343*
low_income	-0.0106	0.0106	0.0075	0.0017	0.0097
black	0.0075	-0.0104	-0.0131	0.0036	0.0139
hispanic	0.0045	-0.022	-2.06e-2	-0.0055	-0.0109
race_other	-0.0201	-0.0289	-0.0373*	-0.0378**	-0.0302*
rurality	-0.0235*	-0.014	-0.0429***	-0.014	-0.0354**
age_65+:low_income	-0.0022	0.0017	-0.0025	-0.0043	-0.0158
age_65+:black	0.0058	0.0056	0.0091	0.0101	-0.0062
age_65+:hispanic	-0.0025	-0.0007	0.0043	0.0028	0.0019
age_65+:race_other	0.0038	0.0103*	0.0102*	0.0093*	0.007
age_65+:rurality	0.0519***	0.0122	0.0519***	0.0231	0.0449**
low_income:black	0.0132*	-0.0019	-0.0086	0.0061	0.0088
low_income:hispanic	0.0065	-0.0025	-0.0138	-0.0033	-0.0057
low_income:race_other	-0.0099	-8.30e-3	-0.0099	-0.0078	-0.0022
low_income:rurality	0.0106*	-0.0017	-0.0079	0.0029	-0.0043
black:hispanic	-0.0041	-0.0003	-0.0063*	-0.0043	-0.0051*
black:race_other	0.0017	0.0003	-0.0023	0.0019	0.0031
black:rurality	0	0.002	5.70e-3	-0.0007	0.0075
hispanic:race_other	-0.0029	0.0032	-0.0012	0.0008	0.0037
hispanic:rurality	0.0067	0.022***	0.0089	0.0051	0.0097
race_other:rurality	0.0096	0.0087	0.0214	0.0226*	0.0201*

Table A.2: Regression coefficient analysis for Arimax LTW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '***' 0.001, '**' 0.01, '*' 0.05

Features	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org
r2	0.0444	0.0617	0.0745	0.0693	0.056
adjusted_r2	0.0371	0.0567	0.0601	0.0649	0.0521
const	0.0743	0.0594	0.039	0.0361	0.0417
age_65+	-0.0192	-0.0071	-0.029*	-0.0101	0.0113
low_income	0.0093	-0.0115	0.0209*	-0.0002	-0.0175***
black	0.0449***	-0.0088	0.0179	0.0128*	0.0028
hispanic	-0.0032	-0.0142	0.0123	-0.0032	-0.0138*
race_other	0.0081	0.0089	0.0017	-0.0088	-0.0087
rurality	-0.0124	0.0306**	0.0029	0.0023	0.0188**
age_65+:low_income	-0.0241*	-0.0166*	-0.044***	-0.0116*	0.0038
age_65+:black	-0.0176	-0.0108	-0.0174*	-0.0044	-0.0136***
age_65+:hispanic	-0.0169**	-0.0032	-0.0126*	0.0009	0.0069**
age_65+:race_other	-0.0136	-0.0138*	-0.0161*	-0.0007	0.0048
age_65+:rurality	0.0221	-0.0035	0.0129	0.0094	-0.0132*
low_income:black	0.0157*	-0.0203***	-0.005	-0.0006	-0.0062
low_income:hispanic	-0.001	-0.0058	-0.0011	-0.0016	-0.0014
low_income:race_other	-0.006	-0.0034	-0.0073	-0.0171***	-0.0037
low_income:rurality	0.0084	0.0259***	0.01	0.0088**	0.0142***
black:hispanic	0.0055*	0.0048*	-0.0017	0.0002	0.002
black:race_other	-0.0046*	-0.002	-0.0023	-0.0004	-0.0016
black:rurality	-0.0001	0.0035	0.0041	-0.0008	0.0112**
hispanic:race_other	-0.0008	-0.0024	0.0032	0.0006	0.0002
hispanic:rurality	0.0191**	0.0089*	-0.0004	0.0061*	0.0063*
race_other:rurality	0.0028	0.0054	0.0078	-0.005	0.0009

Features	(SafeGraph) Restaurants	(SafeGraph) Schools	SafeGraph Inflow	SafeGraph Intraflow	SafeGraph Outflow
r2	0.0727	0.0345	0.0529	0.0323	0.0376
adjusted_r2	0.0679	0.03	0.0482	0.0274	0.033
const	0.0453	0.0479	0.0514	0.0474	0.0531
age_65+	0.0085	-0.002	0.0056	-0.0013	-0.0158
low_income	-0.0168**	-0.0049	-0.0114	-0.0033	0.0007
black	0.0003	-0.0014	0.0201*	0.0083	0.002
hispanic	-0.0058	-0.0274***	-0.0087	0.0008	-0.0211**
race_other	-0.0089	-0.0241**	-0.0273*	0.0001	-0.0104
rurality	0.0124	0.0013	0.0214**	0.0006	0.0019
age_65+:low_income	0.0028	-0.0024	-0.0119	-0.0058	-0.0145*
age_65+:black	0.0128**	-0.0023	-0.0116*	-0.0054	-0.0137**
age_65+:hispanic	0.0062*	0.0049	-0.0082*	-0.0049	-0.0011
age_65+:race_other	0.0035	0.0011	-0.0019	-0.0034	-0.003
age_65+:rurality	-0.0054	0.0024	-0.0091	0.0084	0.0182*
low_income:black	-0.0101**	0.0022	0.009*	0.0019	-0.0077
low_income:hispanic	0.005	-0.0048	-0.0064	-0.0022	-0.0063
low_income:race_other	-0.0037	-0.0079*	-0.0109**	-0.0057	-0.0092*
low_income:rurality	0.0136***	0.009**	0.0215***	0.0078*	0.0169***
black:hispanic	0.0007	0.0056***	-0.0001	-0.0001	0.0017
black:race_other	-0.0012	-0.0018	0.0017	-0.0009	-0.0032*
black:rurality	0.0133***	0.0098**	0.0086*	0.0079	0.0113**
hispanic:race_other	0.0034*	0.0023	0.0012	-0.0001	0.0007
hispanic:rurality	0.0073*	0.0184***	0.0162***	0.0064	0.0187***
race_other:rurality	0.0006	0.0138*	0.0194*	-0.0001	0.0051

Table A.3: Regression coefficient analysis for Arimax STW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '***' 0.001, '**' 0.01, '*' 0.05

Features	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org
r2	0.0297	0.0218	0.0054	0.0195	0.0102
adjusted_r2	0.0229	0.0184	-0.001	0.0164	0.0076
const	0.101***	0.064***	0.061***	0.071***	0.063***
age_65+	-0.027	-0.017	0.01	-0.018	-0.012
low_income	0.055*	0.033*	0.024	0.032*	0.015
black	-0.067	-0.013	-0.046	-0.042*	-0.003
hispanic	-0.072*	-0.017	0.02	-0.038*	-0.017
race_other	-0.049	-0.015	-0.082	-0.041	-0.031
rurality	-0.039	-0.031	-0.003	-0.005	-0.004
age_65+:low_income	-0.033	-0.016	-0.002	-0.02	-0.013
age_65+:black	0.038	-0.028*	0.002	0.028*	0
age_65+:hispanic	-0.029	0.022*	0.001	0.021*	0.007
age_65+:race_other	-0.036	-0.003	0.026	-0.003	-0.002
age_65+:rurality	0.039	0.021	-0.026	0.002	0.012
low_income:black	-0.026	-0.036***	-0.033	-0.015	-0.006
low_income:hispanic	-0.079***	-0.021	0.01	-0.025*	-0.018
low_income:race_other	-0.011	-0.004	-0.032	-0.013	-0.008
low_income:rurality	0.002	-0.011	-0.017	0.001	0.006
black:hispanic	-0.012	-0.009*	-0.006	-0.006	0
black:race_other	0.009	-0.002	0.007	0.004	0.003
black:rurality	0.024	0.031***	0.017	0.017	0.007
hispanic:race_other	0.028***	0.007	0.013	0.004	0
hispanic:rurality	0.035*	-0.008	-0.012	0.001	0.002
race_other:rurality	0.05*	0.013	0.013	0.026	0.022

Features	(SafeGraph) Restaurants	(SafeGraph) Schools	SafeGraph Inflow	SafeGraph Intraflow	SafeGraph Outflow
r2	0.0208	0.0056	0.0217	0.0154	0.0204
adjusted_r2	0.018	0.0025	0.0192	0.0131	0.0181
const	0.069***	0.058***	0.07***	0.07***	0.066***
age_65+	-0.041*	-0.01	-0.035*	-0.038*	-0.043*
low_income	0.047***	0.017	0.035**	0.037**	0.039***
black	-0.039*	0.01	-0.051**	-0.029	-0.04*
hispanic	-0.038*	-0.048*	-0.035*	-0.031*	-0.041**
race_other	-0.057**	-0.043	-0.023	-0.027	-0.033
rurality	-0.037*	0.006	-0.036*	-0.03	-0.019
age_65+:low_income	-0.04*	-0.021	-0.028*	-0.025*	0.033**
age_65+:black	0.003	-0.003	0.009	-0.011	0.011
age_65+:hispanic	0.011	0.015	0.014*	0.022***	0.014*
age_65+:race_other	-0.005	-0.005	-0.005	0.004	0.003
age_65+:rurality	0.033	-0.005	0.04*	0.03	0.029
low_income:black	-0.03**	-0.003	-0.034***	-0.031***	-0.029***
low_income:hispanic	-0.021*	-0.025*	-0.024*	-0.02*	-0.026**
low_income:race_other	-0.01	-0.004	-0.006	-0.012	-0.013
low_income:rurality	-0.004	0.008	-0.002	-0.007	0.004
black:hispanic	-0.001	-0.004	0	-0.003	0.002
black:race_other	0.001	0	0.001	0.003	0.004
black:rurality	0.024**	-0.003	0.03***	0.027**	0.016
hispanic:race_other	0.02***	0.012**	0.001	-0.005	-0.003
hispanic:rurality	0.013	0.011	0.008	0	0.01
race_other:rurality	0.042**	0.032	0.021	0.012	0.016

Table A.4: Regression coefficient analysis for Ridge Regression LTW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '***' 0.001, '**' 0.01, '*' 0.05

Features	Apple	Descartes	Google	(SafeGraph) Grocery Store	(SafeGraph) Religious Org	Features	(SafeGraph) Restaurants	(SafeGraph) Schools
r2	0.012576	0.022346	0.017356	0.008641	0.009302	r2	0.011075	0.009231
adjusted_r2	0.008654	0.01869	0.0085	0.00556	0.006092	adjusted_r2	0.007945	0.006016
const	0.1071***	0.0708***	0.0656***	0.0591***	0.0689***	const	0.0731***	0.1135***
age_65+	-0.0306	-0.0208	-0.0126	-0.0064	-3.10e-3	age_65+	-0.0167	0.029
low_income	0.0077	0.0098	-5.58e-5	-0.0054	0.0024	low_income	0.0111	-0.0298*
black	-0.0443*	0.0088	-0.0147	0.0134	-0.0199	black	-0.0171	0.0283
hispanic	0.0115	-0.01	-0.0134	0.0034	-0.0215	hispanic	-0.003	-0.0331
race_other	-0.0064	-0.0273*	0.0017	-0.0111	-0.0033	race_other	-0.0084	-0.0202
rurality	-0.0684**	0.0094	-0.0231	-0.0005	-0.0218	rurality	-2.07e-2	0.0155
age_65+:low_income	-0.0119	-0.0181	-0.0033	-0.0075	0.0044	age_65+:low_income	-0.0154	0.0209
age_65+:black	0.0303*	-0.0025	-0.0014	-0.0009	0.0015	age_65+:black	0.0046	0.0034
age_65+:hispanic	-0.0252**	0.0056	0.0138	-0.0012	0.0016	age_65+:hispanic	-0.0038	-0.0018
age_65+:race_other	-0.0119	0.0117	-0.0215	0.0056	0.0079	age_65+:race_other	0.003	0.008
age_65+:rurality	0.059**	0.0092	0.0288	0.008	0.0117	age_65+:rurality	0.02	-0.0186
low_income:black	-0.0069	-0.0133*	-0.0139	0.0046	-0.0145*	low_income:black	-0.0133	0.0224*
low_income:hispanic	0.013	-0.0193*	-0.0028	0.0018	-0.0096	low_income:hispanic	-0.0086	-0.007
low_income:race_other	-0.0035	-0.0055	-0.0092	1.60e-3	0.0009	low_income:race_other	-0.0004	0.0037
low_income:rurality	-0.0153	0.0134*	0.0011	0.0005	-0.004	low_income:rurality	-0.0047	0.0101
black:hispanic	0.0064	-0.0115***	-0.0022	-0.0023	-0.0035	black:hispanic	-0.0038	0.0018
black:race_other	0.0031	0.0039	0.0025	4.00e-4	0.0015	black:race_other	0.0024	-0.0004
black:rurality	0.0133	-0.0083	0.0105	-0.0015	0.0113	black:rurality	0.0109	-0.0157
hispanic:race_other	0.0085	0.0051	-0.0012	-0.0008	-1.70e-3	hispanic:race_other	-0.0008	-0.0037
hispanic:rurality	0.0189*	-0.0034	-0.0006	0.0017	0.0125*	hispanic:rurality	0.0044	0.022*
race_other:rurality	0.0095	0.0056	0.0115	0.008	-0.0028	race_other:rurality	0.006	0.0126

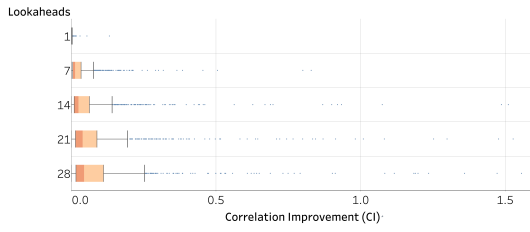
Table A.5: Regression coefficient analysis for ElasticNet Regression STW. Values highlighted in green represent statistically significant positive coefficients, while values highlighted in red represent statistically significant negative coefficients. Significance level (p-value smaller than): '***' 0.001, '**' 0.01, '*' 0.05

Dataset considered	All datasets	Google excluded
Chi-square	28.2873	9.6000
p-value	0.0009	0.2942

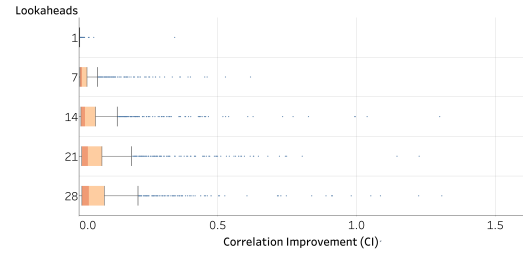
Table A.6: Non-parametric Friedman test analysis with distributions containing the percentage of counties where mobility data improved COVID-19 case prediction models without mobility data across datasets, lookaheads, training approaches and models.

Dataset considered	Chi-square	p-value
All datasets	182.7029	0.0
Restaurants Mobility excluded	164.0857	0.0
Religious Organization Mobility excluded	164.0603	0.0
Schools Mobility excluded	166.6444	0.0
Grocery Stores Mobility excluded	164.6127	0.0
SafeGraph Inflows Mobility excluded	164.1492	0.0
SafeGraph Outflows Mobility excluded	164.6000	0.0
SafeGraph Intraflows Mobility excluded	164.2508	0.0
Apple Mobility excluded	164.5937	0.0
Descartes Mobility excluded	164.0603	0.0
Google Mobility excluded	164.0794	0.0
Apple, Google and SafeGraph Inflow Mobility excluded	108.6286	0.0

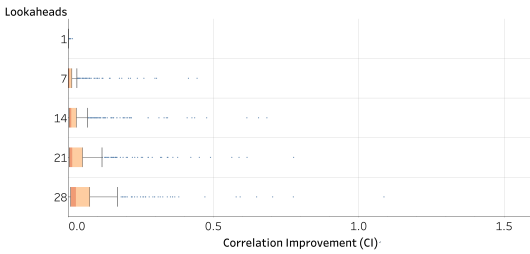
Table A.7: Non-parametric Friedman test analysis with distributions containing the median correlation improvement values of counties where mobility data improved COVID-19 case prediction models without mobility data across datasets, lookaheads, training approaches and models. Each test is run with one or a few mobility datasets excluded.



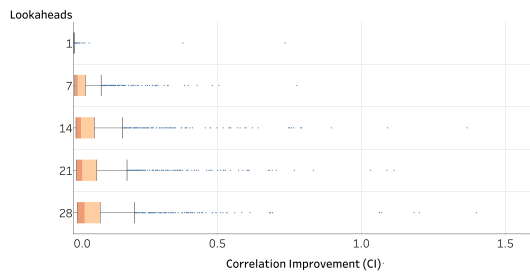
(a) CI for Apple Mobility



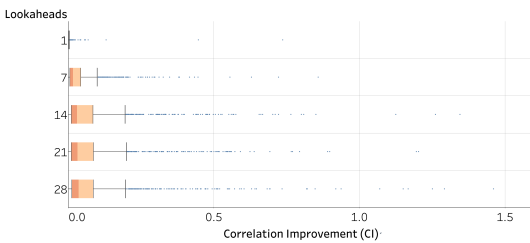
(c) CI for Descartes Mobility



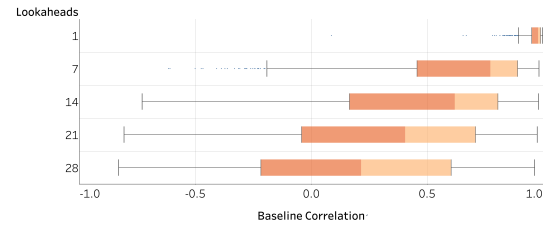
(e) CI for Google Mobility



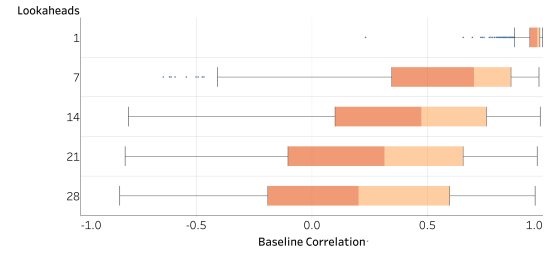
(g) CI for SafeGraph Inflows Mobility



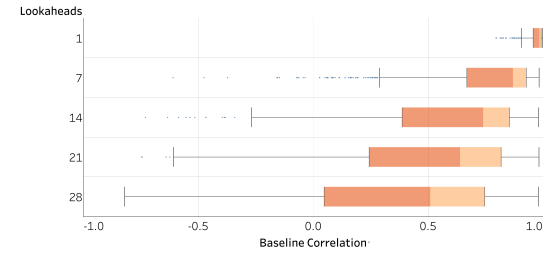
(i) CI for SafeGraph Outflows Mobility



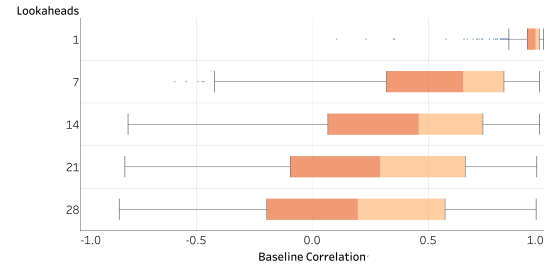
(b) Baseline Correlation for Apple Mobility



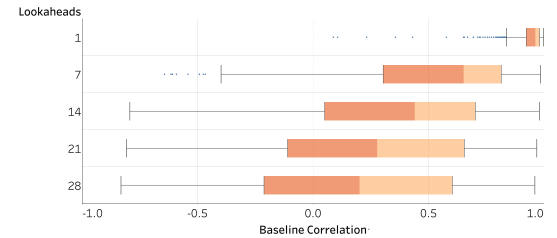
(d) Baseline Correlation for Descartes Mobility



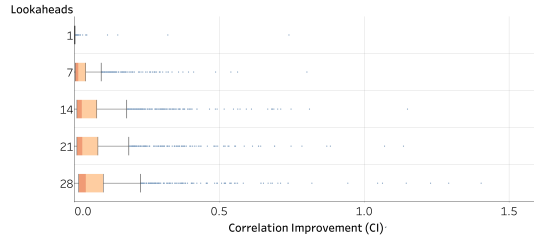
(f) Baseline Correlation for Google Mobility



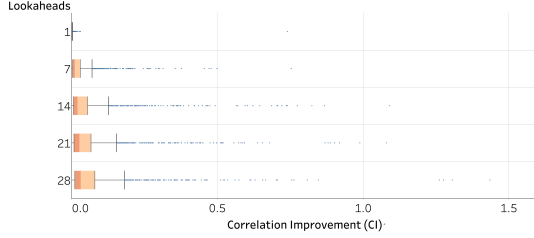
(h) Baseline Correlation for SafeGraph Inflows Mobility



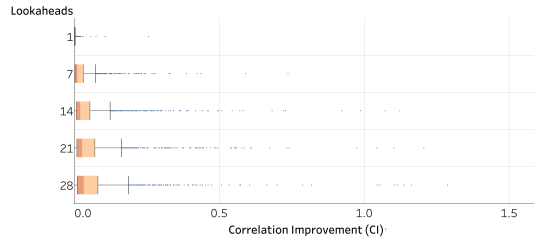
(j) Baseline Correlation for SafeGraph Outflows Mobility



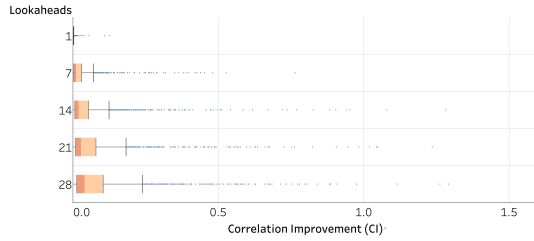
(k) CI for SafeGraph Intraflows Mobility



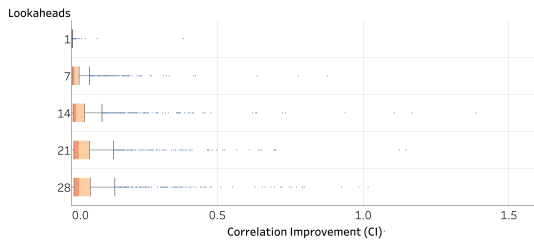
(m) CI for Restaurants Mobility



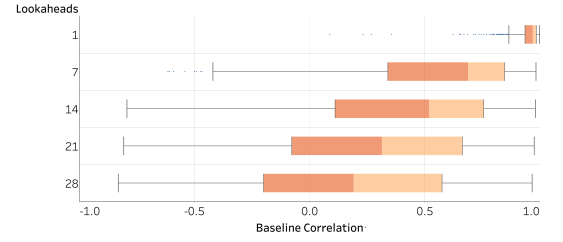
(o) CI for Religious Organization Mobility



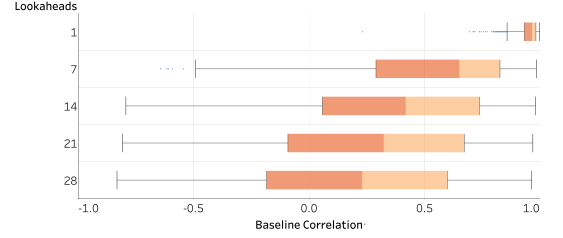
(q) CI for Schools Mobility



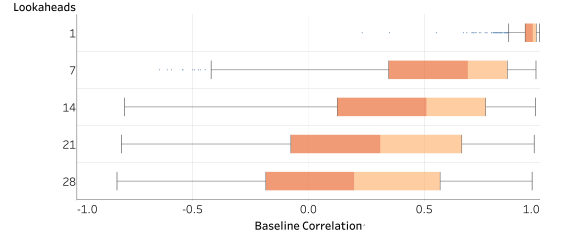
(s) CI for Grocery Stores Mobility



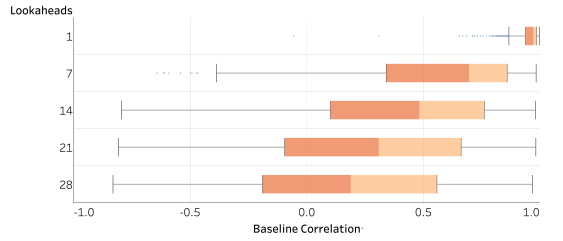
(l) Baseline Correlation for SafeGraph Intraflows Mobility



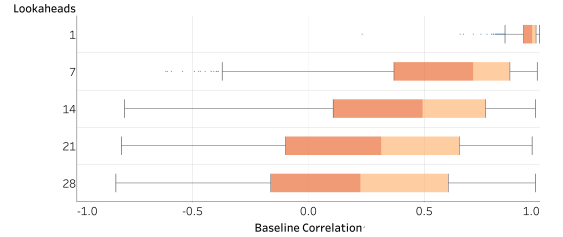
(n) Baseline Correlation for Restaurants Mobility



(p) Baseline Correlation for Religious Organization Mobility

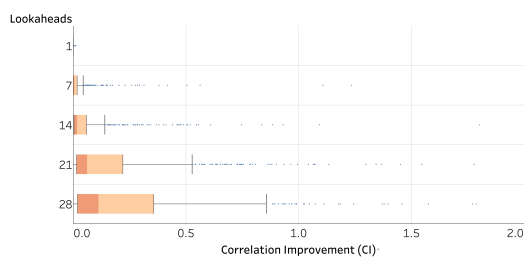


(r) Baseline Correlation for Schools Mobility

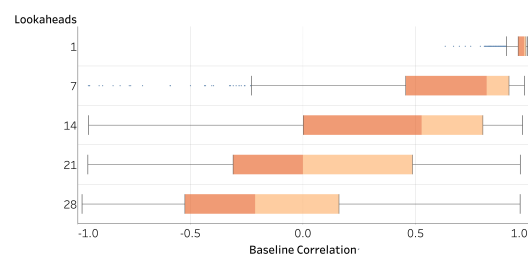


(t) Baseline Correlation for Grocery Stores Mobility

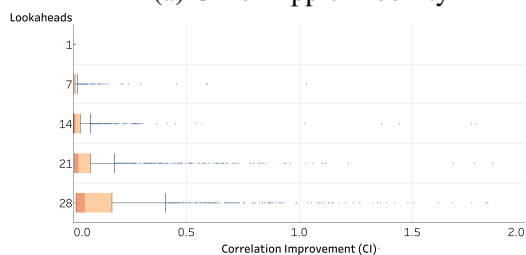
Figure A.1: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with LTW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).



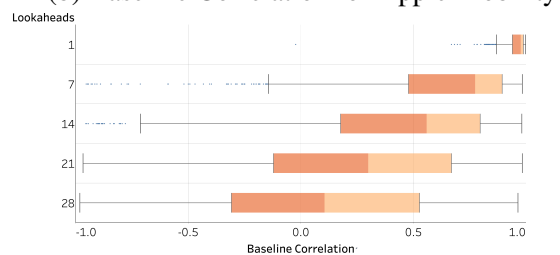
(a) CI for Apple Mobility



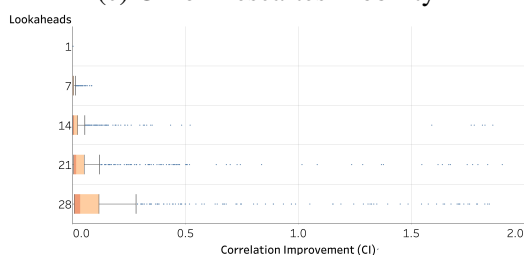
(b) Baseline Correlation for Apple Mobility



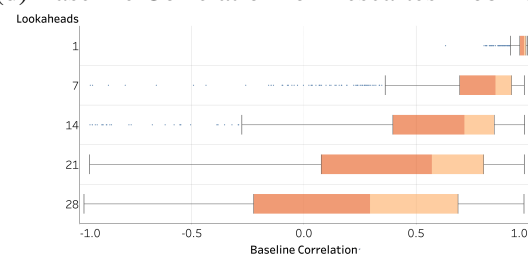
(c) CI for Descartes Mobility



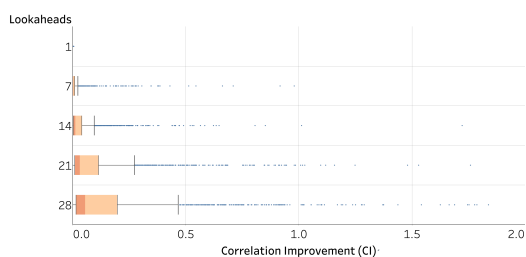
(d) Baseline Correlation for Descartes Mobility



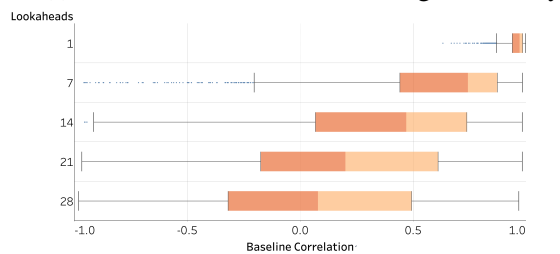
(e) CI for Google Mobility



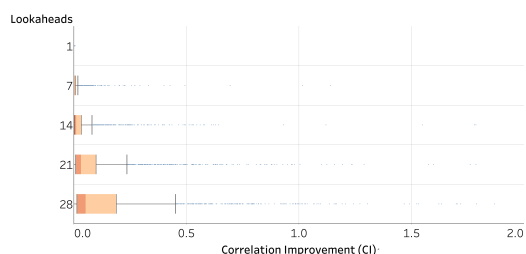
(f) Baseline Correlation for Google Mobility



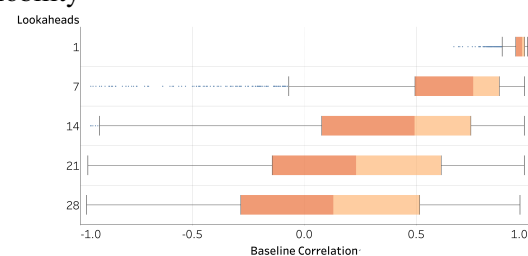
(g) CI for SafeGraph Inflows Mobility



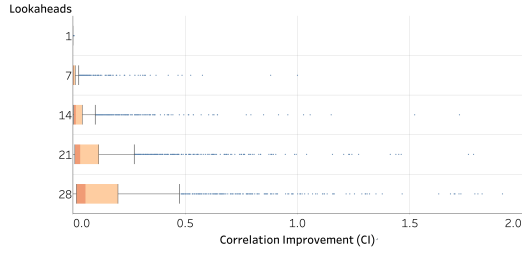
(h) Baseline Correlation for SafeGraph Inflows Mobility



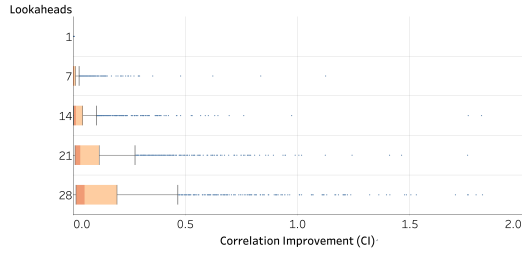
(i) CI for Outflows Mobility



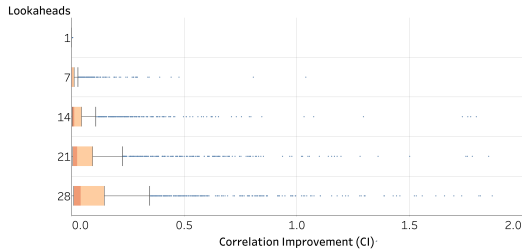
(j) Baseline Correlation for SafeGraph Outflows Mobility



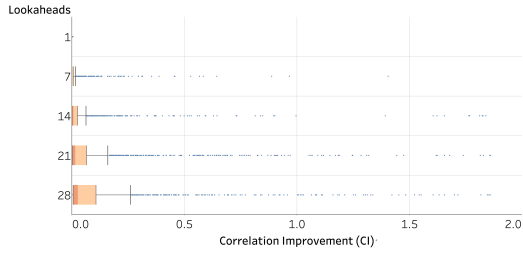
(k) CI for SafeGraph Intraflows Mobility



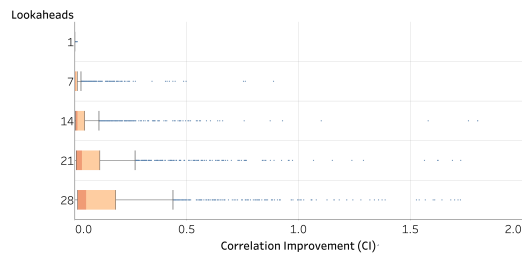
(m) CI for Restaurants Mobility



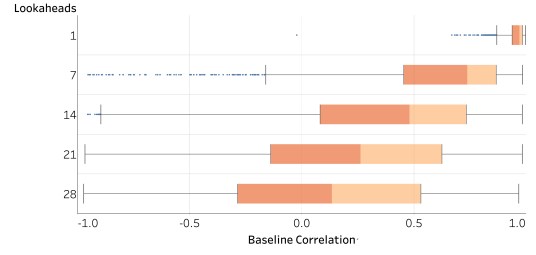
(o) CI for Religious Organization Mobility



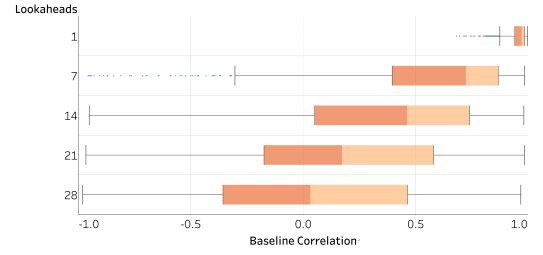
(q) CI for Schools Mobility



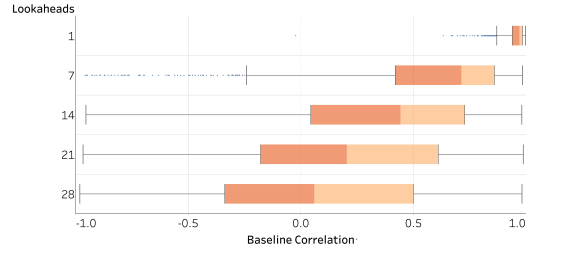
(s) CI for Grocery Stores Mobility



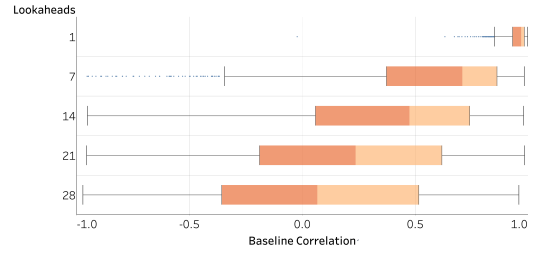
(l) Baseline Correlation for SafeGraph Intraflows Mobility



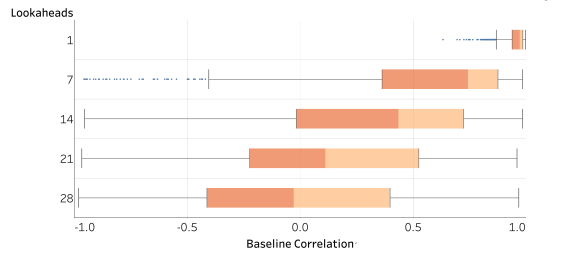
(n) Baseline Correlation for Restaurants Mobility



(p) Baseline Correlation for Religious Organization Mobility

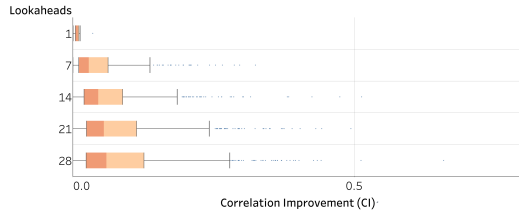


(r) Baseline Correlation for Schools Mobility

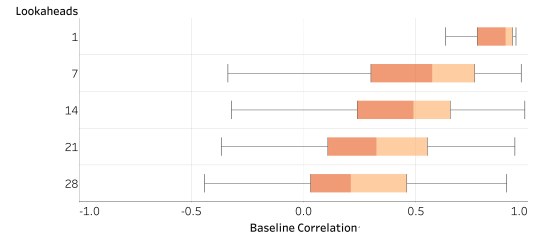


(t) Baseline Correlation for Grocery Stores Mobility

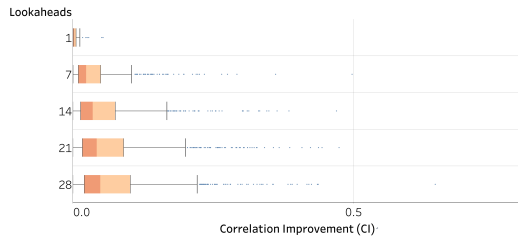
Figure A.2: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ridge regression with LTW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).



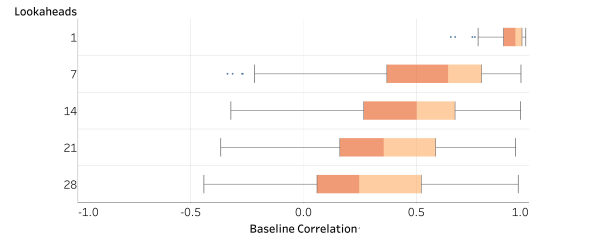
(a) CI for Apple Mobility



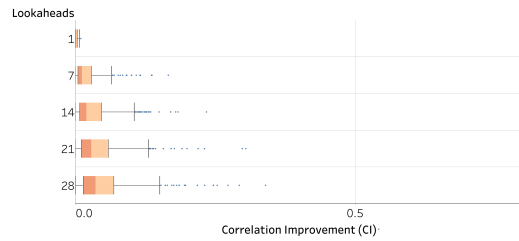
(b) Baseline Correlation for Apple Mobility



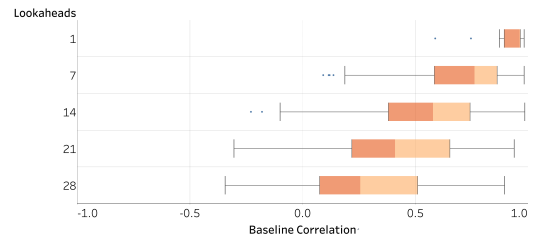
(c) CI for Descartes Mobility



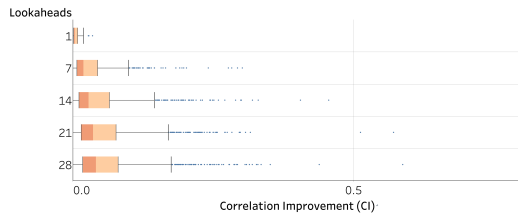
(d) Baseline Correlation for Descartes Mobility



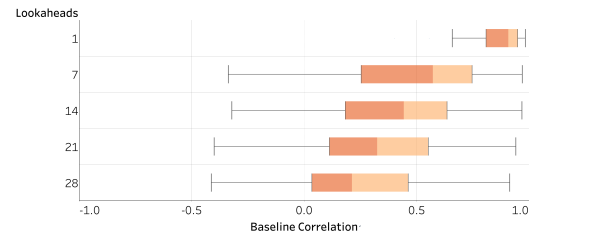
(e) CI for Google Mobility



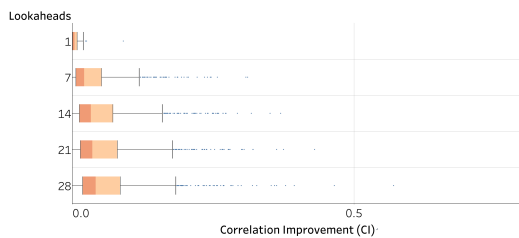
(f) Baseline Correlation for Google Mobility



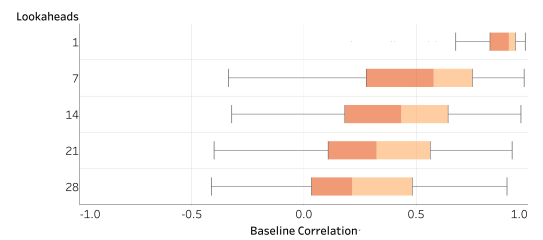
(g) CI for SafeGraph Inflows Mobility



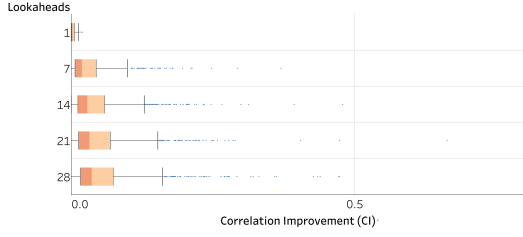
(h) Baseline Correlation for SafeGraph Inflows Mobility



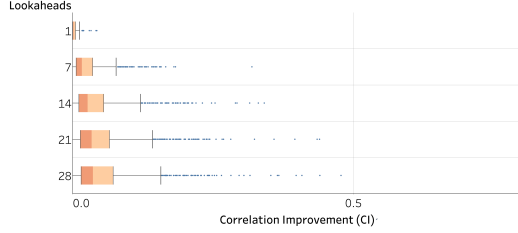
(i) CI for SafeGraph Outflows Mobility



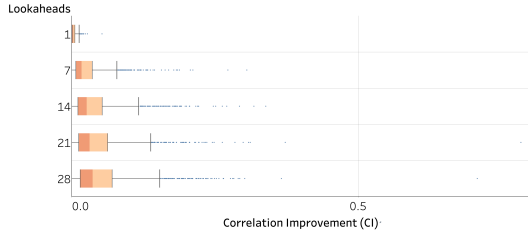
(j) Baseline Correlation for SafeGraph Outflows Mobility



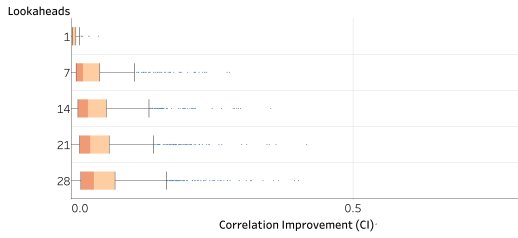
(k) CI for SafeGraph Intraflows Mobility



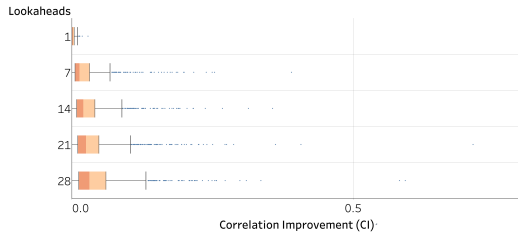
(m) CI for Restaurants Mobility



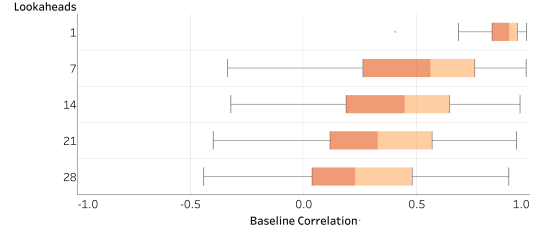
(o) CI for Religious Organization Mobility



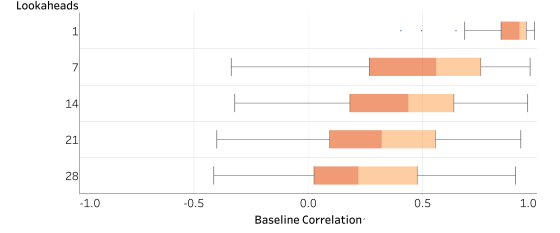
(q) CI for Schools Mobility



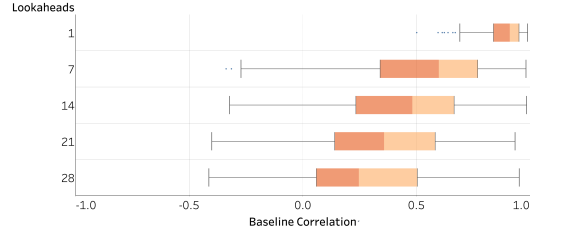
(s) CI for Grocery Stores Mobility



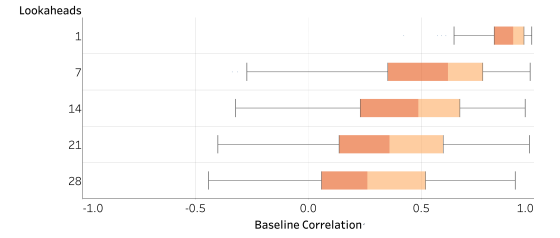
(l) Baseline Correlation for SafeGraph Intraflows Mobility



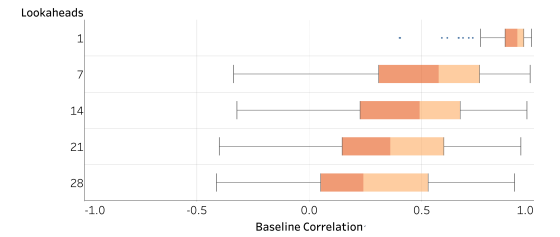
(n) Baseline Correlation for Restaurants Mobility



(p) Baseline Correlation for Religious Organization Mobility

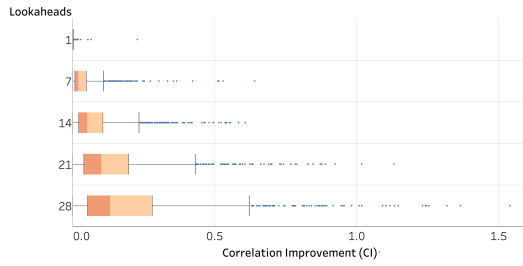


(r) Baseline Correlation for Schools Mobility

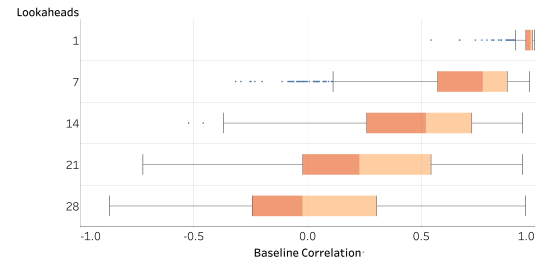


(t) Baseline Correlation for Grocery Stores Mobility

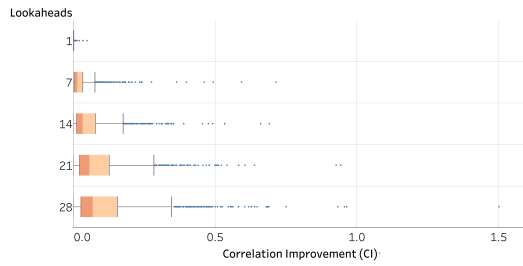
Figure A.3: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for ARIMAX with STW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).



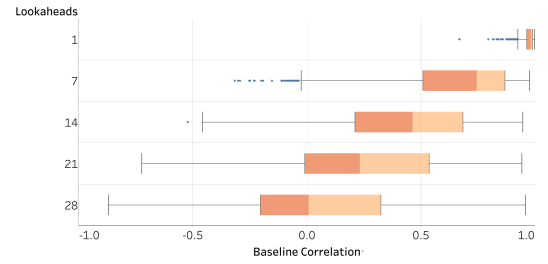
(a) CI for Apple Mobility.



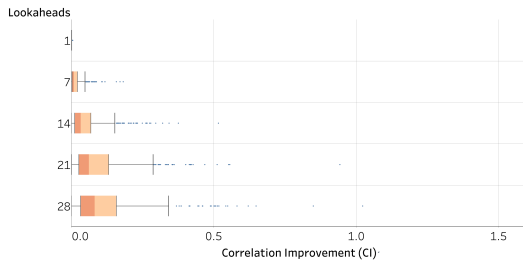
(b) Baseline Correlation for Apple Mobility.



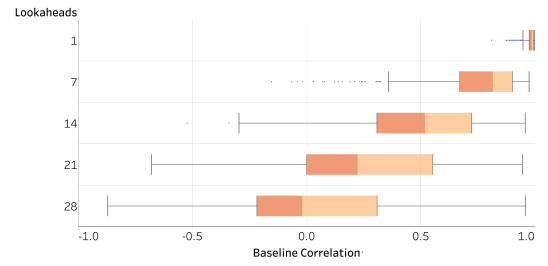
(c) CI for Descartes Mobility.



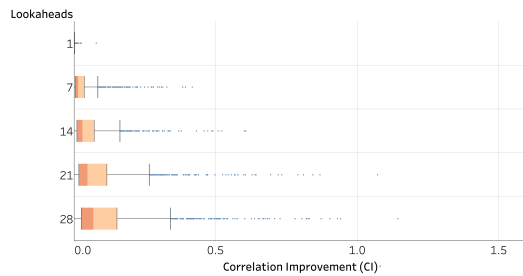
(d) Baseline Correlation for Descartes Mobility data.



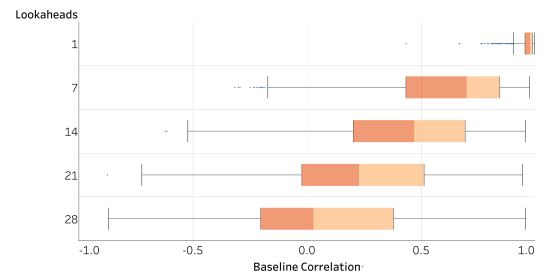
(e) CI for Google Mobility.



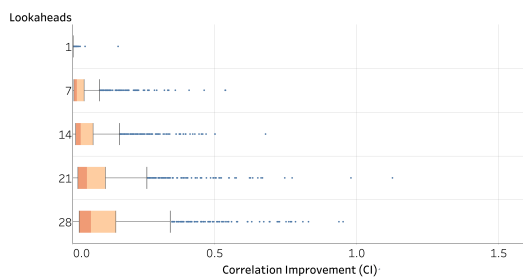
(f) Baseline Correlation for Google Mobility data.



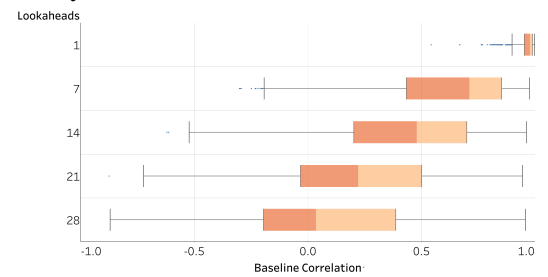
(g) CI for SafeGraph Inflows Mobility.



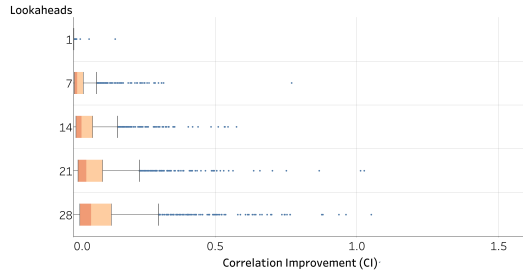
(h) Baseline Correlation for SafeGraph Inflows Mobility.



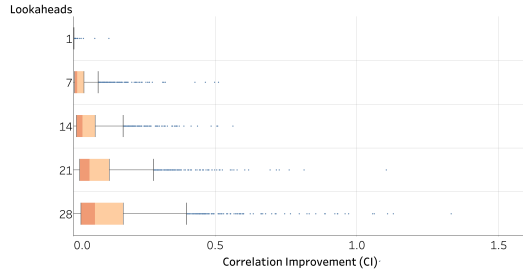
(i) CI for SafeGraph Outflows Mobility.



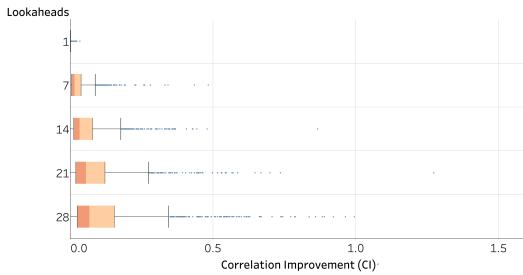
(j) Baseline Correlation for SafeGraph Outflows Mobility.



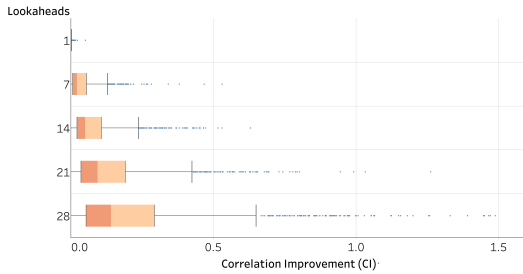
(k) CI for SafeGraph Intraflows Mobility.



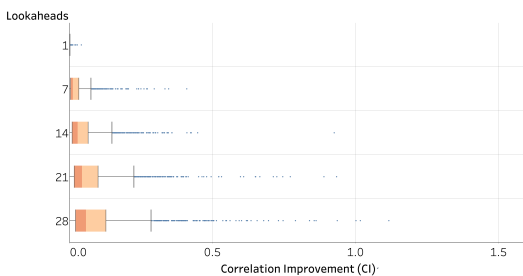
(m) CI for Restaurants Mobility.



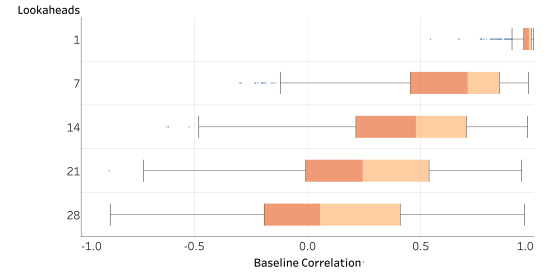
(o) CI for Religious Organization Mobility.



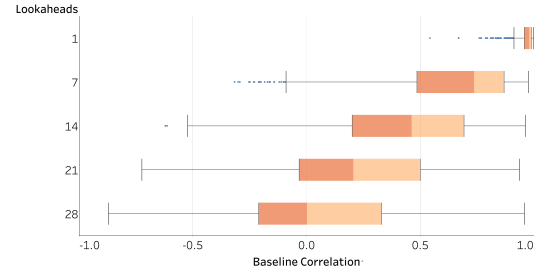
(q) CI for Schools Mobility.



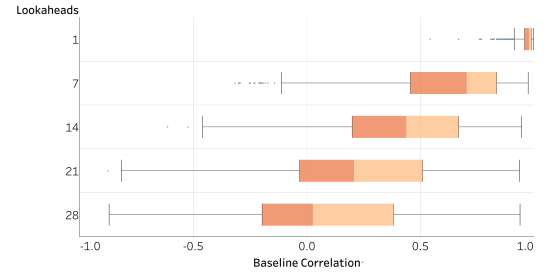
(s) CI for Grocery Stores Mobility.



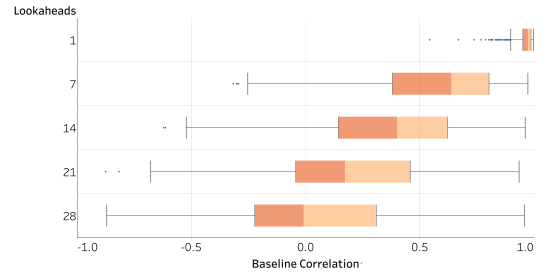
(l) Baseline Correlation for SafeGraph Intraflows Mobility.



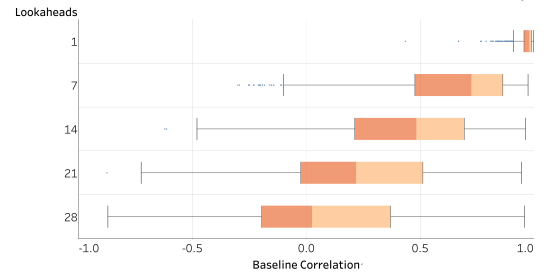
(n) Baseline Correlation for Restaurants Mobility.



(p) Baseline Correlation for Religious Organization Mobility.



(r) Baseline Correlation for Schools Mobility.



(t) Baseline Correlation for Grocery Stores Mobility.

Figure A.4: Correlation improvements (ci) and baseline correlation distributions (non-mobility) across lookaheads for elastic regression with STW approach using (a,b) Apple Mobility data, (c,d) Descartes Mobility data, (e,f) Google Mobility data, (g,h) SafeGraph Inflow Mobility data, (i,j) SafeGraph Outflow Mobility data, (k,l) SafeGraph Intraflow Mobility data, (m,n) Restaurants Mobility data (SafeGraph POI), (o,p) Religious Mobility data (SafeGraph POI), (q,r) Schools Mobility data (SafeGraph POI), (s,t) Grocery Stores Mobility data (SafeGraph POI).

ARIMAX (p,d,q)	# Counties	ARIMAX (p,d,q)	# Counties
(0, 0, 0)	13	(4, 0, 4)	3
(0, 1, 0)	1263	(4, 1, 0)	19
(0, 1, 1)	189	(4, 1, 1)	10
(0, 1, 2)	18	(4, 1, 2)	18
(0, 1, 3)	4	(4, 1, 3)	11
(0, 1, 4)	8	(4, 1, 4)	3
(0, 1, 5)	12	(4, 1, 5)	1
(0, 2, 0)	1	(4, 2, 0)	1
(0, 2, 1)	25	(4, 2, 1)	1
(0, 2, 2)	5	(4, 2, 2)	1
(0, 2, 3)	3	(4, 2, 3)	1
(1, 0, 0)	108	(4, 2, 4)	1
(1, 0, 1)	25	(5, 0, 0)	2
(1, 0, 2)	4	(5, 0, 1)	4
(1, 0, 3)	3	(5, 0, 2)	7
(1, 0, 4)	3	(5, 1, 0)	4
(1, 0, 5)	1	(5, 1, 1)	3
(1, 1, 0)	201	(5, 1, 2)	10
(1, 1, 1)	263	(5, 1, 3)	9
(1, 1, 2)	44	(5, 1, 4)	5
(1, 1, 3)	22	(5, 2, 2)	2
(1, 1, 4)	9	(5, 2, 3)	3
(1, 1, 5)	9	(5, 2, 5)	1
(1, 2, 1)	2	(6, 0, 3)	1
(1, 2, 2)	6	(6, 1, 3)	2
(1, 2, 3)	3	(6, 1, 4)	1
(1, 2, 4)	2	(6, 1, 5)	3
(2, 0, 0)	31	(7, 0, 2)	1
(2, 0, 1)	73	(7, 0, 3)	1
(2, 0, 2)	18	(7, 1, 0)	17
(2, 0, 3)	5	(7, 1, 1)	3
(2, 0, 4)	3	(7, 1, 2)	5
(2, 1, 0)	100	(7, 1, 3)	2
(2, 1, 1)	59	(7, 1, 4)	2
(2, 1, 2)	103	(7, 1, 5)	1
(2, 1, 3)	19	(7, 2, 1)	2
(2, 1, 4)	15	(7, 2, 3)	2
(2, 1, 5)	15	(8, 0, 0)	3
(2, 2, 1)	7	(8, 0, 2)	1
(2, 2, 3)	2	(8, 1, 0)	5
(2, 2, 4)	1	(8, 1, 1)	4
(2, 2, 5)	1	(8, 1, 2)	2
(3, 0, 0)	9	(8, 1, 3)	2
(3, 0, 1)	6	(8, 1, 4)	2
(3, 0, 2)	24	(9, 0, 0)	3
(3, 0, 3)	4	(9, 0, 1)	1
(3, 0, 4)	5	(9, 0, 3)	1
(3, 0, 5)	1	(9, 1, 1)	3
(3, 1, 0)	32	(9, 1, 2)	1
(3, 1, 1)	19	(9, 1, 4)	1
(3, 1, 2)	36	(10, 0, 0)	2
(3, 1, 3)	23	(10, 0, 4)	1
(3, 1, 4)	14	(10, 1, 1)	1
(3, 1, 5)	12	(10, 1, 2)	2
(3, 2, 0)	2	(10, 1, 3)	1
(3, 2, 1)	1	(11, 0, 0)	1
(3, 2, 2)	2	(11, 1, 0)	1
(3, 2, 5)	1	(12, 0, 3)	1
(4, 0, 0)	7	(12, 1, 0)	1
(4, 0, 1)	6	(12, 1, 1)	1
(4, 0, 2)	5	(14, 1, 0)	1
(4, 0, 3)	3	(15, 1, 0)	1

Table A.8: Summary of the combinations of p, d, q values identified for the ARIMAX models using grid search and AIC. The table represents the number of counties whose p, d, q , values are the ones listed on the left hand side column.

Appendix B: Auditing the Fairness of COVID-19 Forecast Hub’s Case Prediction Models

B.1 Supporting Information for Data

B.1.1 Exclusion of some counties and certain minority races

Our analysis excluded certain counties and racial/ethnic groups based on data availability and statistical considerations. At the county level, we excluded three geographical regions:

- Florida counties were excluded due to insufficient data in the PLACES Dataset, which did not meet the CDC’s minimum data requirements for inclusion¹.
- Connecticut counties were excluded due to a structural mismatch in reporting systems. Connecticut’s switch from county FIPS reporting to planning region reporting created incompatibilities between COVID-19 ForecastHub predictions, PLACES dataset, and ACS demographic data².
- Puerto Rico counties (FIPS Code: 72XXX) were excluded as the COVID-19 ForecastHub does not provide county-level forecasts for this territory.

Regarding racial and ethnic groups, we made methodological decisions based on population distributions across U.S. counties (see Table 1 in main manuscript). We excluded American Indian/Alaska

¹<https://www.cdc.gov/places/current-release-notes/index.html#:~:text=2023%20Release,were%20carried%20over%20for%20Florida>

²https://developer.ap.org/ap-elections-api/docs/CT_FIPS_Codes_forPlanningRegions.htm

Native (AIAN) and Native Hawaiian/Pacific Islander (NHPI) populations from our regression analysis due to limited variable distribution. Specifically, 97.69% of counties have NHPI populations under 1%, and 79.03% have AIAN populations under 1%, providing insufficient variation for meaningful regression analysis.

In contrast, over one-third of Asian counties, two-thirds of Black counties and almost all Hispanic counties have more than 1% of population presence. Additionally, while Asian populations show a relatively low median of 0.70% (compared to Black and Hispanic counties), their mean percentage (1.53%) is more than twice the median, indicating a right-skewed but analytically useful distribution. These statistics reveal an adequate statistical power for detecting potential effects in our regression analysis. Furthermore, Asian populations represent a key demographic group in metropolitan and suburban areas where COVID-19 impacts were particularly pronounced, making their inclusion crucial for understanding prediction accuracy in these important contexts.

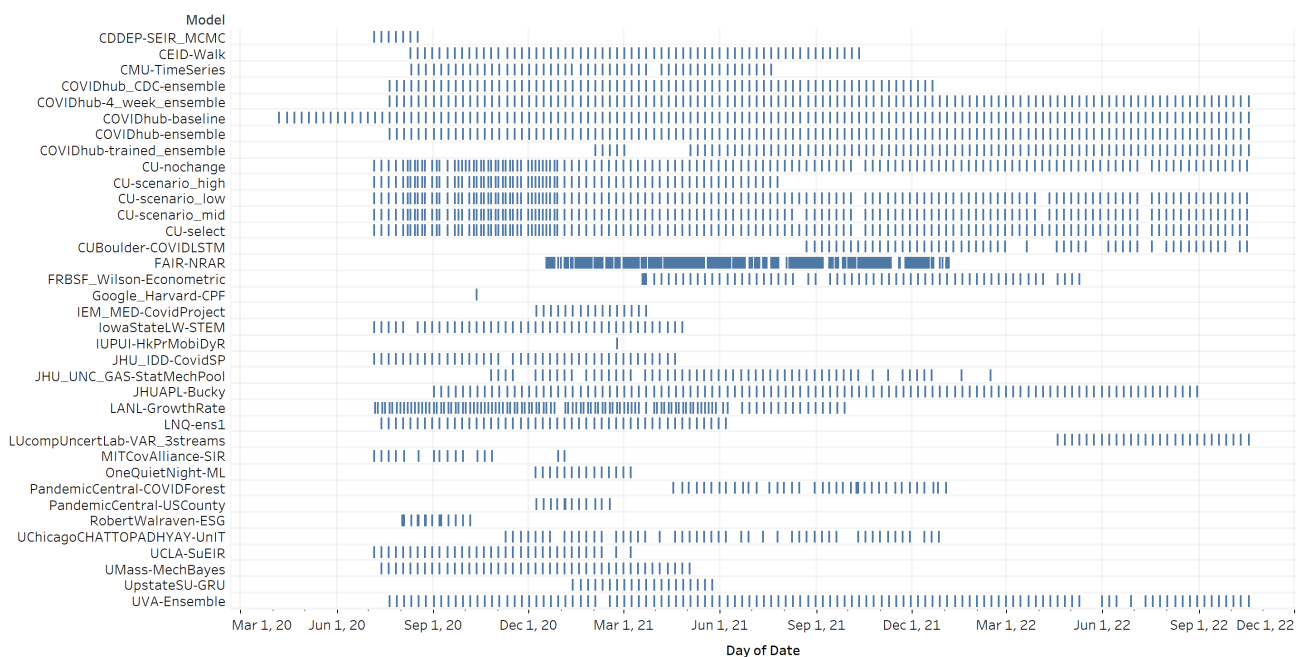


Figure B.1: Timeline Overview of COVID-19 Model Forecasts by County. This Gantt chart captures the forecast submission frequency from the 36 research teams' models analyzed in this paper, delineating the evolution of modeling efforts throughout the pandemic.

Model Name	Model Type	Mobility Used?
CU-nochange	Compartmental	Yes (Census)
CMU-TimeSeries	Statistical	No
COVIDhub_CDC-ensemble	Ensemble	Mixed
COVIDhub-trained_ensemble	Ensemble	Mixed
COVIDhub-ensemble	Ensemble	Mixed
IEM_MED-CovidProject	Compartmental	No
LNQ-ens1	Ensemble	No
JHU_IDD-CovidSP	Compartmental	No
CUBoulder-COVIDLSTM	Deep Learning	Yes (Facebook)
Google_Harvard-CPF	Deep Learning	Yes (Descartes Labs)
FRBSF_Wilson-Econometric	Compartmental	Yes (Google and Safegraph)
UMass-MechBayes	Statistical	No
LANL-GrowthRate	Compartmental	No
UCLA-SuEIR	Compartmental	No
COVIDhub-4_week_ensemble	Ensemble	Mixed
CEID-Walk	Statistical	No
RobertWalraven-ESG	Statistical	No
OneQuietNight-ML	Statistical	Yes(Apple, Google)
CU-scenario_high	Compartmental	Yes (Census)
IUPUI-HkPrMobiDyR	Statistical	Yes (Google)
CDDEP-SEIR_MCMC	Compartmental	No
UpstateSU-GRU	Deep Learning	Yes (Google)
CU-scenario_mid	Compartmental	Yes (Census)
CU-scenario_low	Compartmental	Yes (Census)
UChicagoCHATTOPADHYAY-UnIT	Statistical	No
PandemicCentral-USCounty	Statistical	Yes (Facebook)
JHUAPL-Bucky	Compartmental	Yes (Safegraph)
COVIDhub-baseline	Baseline Models	Not Applicable
PandemicCentral-COVIDForest	Statistical	Yes (Facebook)
LUcompUncertLab-VAR_3streams	Statistical	No
FAIR-NRAR	Deep Learning	Yes (Facebook, Google)
JHU_UNC_GAS-StatMechPool	Ensemble	Yes
CU-select	Compartmental	Yes (Census)
IowaStateLW-STEM	Compartmental	Yes (Descartes Labs, US DOT)
MITCovAlliance-SIR	Compartmental	Yes (Safegraph)
UVA-Ensemble	Ensemble	Yes

Table B.1: List of the **36 COVID-19 forecast models** analyzed in this paper, together with their model type and whether they used mobility data or not, including the source of that mobility data. Information about model type and the use of mobility data was manually extracted from the paper(s) for each model by the authors of this paper.

B.1.2 Description of the model type classification

The authors of this paper created a model taxonomy (see below) after reading the papers for each of the 36 models under study. Using this taxonomy, each COVID-19 forecast model was labeled by two authors of this paper. Whenever there was a labeling disagreement, the model was discussed until a final agreement was reached. For transparency purposes, we provide the complete classification of all models along with their descriptions in Table [B.1](#) in this Appendix.

We propose the following refined model taxonomy based on the primary modeling approach used by each team:

- **Compartmental Models:** Models that primarily use variations of epidemiological compartmental frameworks (e.g., SIR, SEIR) as their core predictive mechanism. While these models may incorporate additional computational techniques for parameter estimation, they are classified based on their fundamental epidemiological structure. For example, UCLA-SuEIR uses BFGS optimization with an SEIR model, but its primary predictive mechanism remains compartmental. Similarly, while the IEM_MED-CovidProject model uses AI techniques for parameter fitting, we classify it as compartmental because its core predictive mechanism relies on a SEIR framework, with AI serving an auxiliary role in parameter optimization. This classification prioritizes the fundamental epidemiological structure over supplementary computational techniques used for implementation.
- **Statistical Models:** Models that rely predominantly on statistical methods including time series analysis, regression, and classical statistical approaches. For instance, CMU-TimeSeries uses quantile regression, while CEID-Walk employs random walk methods.

- **Deep Learning Models:** Models that primarily leverage neural network architectures for prediction. For example, CUBoulder-COVIDLSTM uses LSTM networks as its core prediction mechanism, while Google.Harvard-CPF employs deep learning with mobility data.
- **Ensemble Models:** Models that systematically combine predictions from multiple forecasting approaches. We identify two types of ensembles in the Forecast Hub: (1) Hub-level ensembles (e.g., COVIDhub_CDC_ensemble, COVIDhub-ensemble, etc) that aggregate predictions across different participating teams' models, and (2) team-level ensembles (e.g., UVA-Ensemble, LNQ-ens1) that combine multiple methodological approaches within a single submission. For example UVA-Ensemble computes a bayesian model averaging of autoregressive models with mobility data, LSTM networks, and modified SEIR models. The common thread across both types is their systematic integration of multiple predictive approaches to generate final forecasts.
- **Baseline Models:** Models that serve as fundamental comparison benchmarks. In the Forecast Hub, the primary baseline model (COVIDhub-baseline) uses a specific procedure where the median prediction equals the most recent observed incidence, with uncertainty derived from historical week-to-week changes. This approach assumes future incidence will be similar to current levels, with variations based on historically observed changes. This provides a simple but statistically grounded reference point against which to compare more sophisticated approaches, and is classified as a Baseline model in our approach.

This taxonomy acknowledges that modern epidemiological models often incorporate multiple computational techniques, but classifies them based on their primary predictive mechanism rather than auxiliary methods used for implementation or optimization. We recognize that some models may exhibit hybrid characteristics, but we assign them to categories based on their dominant modeling

paradigm to enable meaningful comparative analysis.

B.1.3 Description of Phases

We divided the pandemic timeline into distinct phases P using a data-driven approach based on local minima (valleys) and local maxima (peaks) in COVID-19 case volumes. This segmentation was motivated by several methodological and epidemiological considerations:

- **P0 (June 3, 2020 - September 9, 2020):** Captures the initial study period with relatively stable low case numbers. The phase ends at a clear valley (around 40,000 cases) before the first major wave, providing a baseline period for model performance analysis.
- **P1 (September 9, 2020 - March 7, 2021):** Encompasses the first major wave reaching approximately 250,000 weekly cases. The start and end points are defined by distinct valleys, capturing a complete epidemic cycle from rise through peak to decline, allowing for the analysis of model behavior during sustained transmission increase.
- **P2 (March 7, 2021 - June 23, 2021):** Represents an inter-wave period with moderate case levels. Both boundaries are marked by clear valleys, isolating a period of relative stability between major waves, which enables assessment of model performance during more stable epidemic conditions.
- **P3 (June 23, 2021 - November 3, 2021):** Contains a distinct wave reaching around 170,000 cases. The phase boundaries capture the full progression from valley through peak to subsequent valley, encompassing the Delta variant wave period.
- **P4 (November 3, 2021 - March 30, 2022):** Defined to contain the most dramatic surge in the

dataset, reaching approximately 800,000 cases. This phase captures the Omicron wave from its onset through peak to initial decline, allowing analysis of model performance during extreme case volumes.

- **P5 (March 30, 2022 - June 22, 2022):** Represents the sharp post-Omicron decline phase. The boundaries isolate this distinct period of rapid case reduction, enabling analysis of model behavior during sustained decrease.
- **P6 (June 22, 2022 - October 20, 2022):** Final phase containing a smaller but distinct wave (Around 130,000 cases). The boundaries capture this complete wave cycle, allowing assessment of model performance during later pandemic dynamics.

This valley-based segmentation offers several analytical advantages. Each phase represents a complete epidemiological pattern (rise-peak-fall or relative stability), with phase transitions occurring at natural breakpoints in case dynamics. The segmentation captures varying intensities of transmission, from baseline to extreme peaks, and phases align with known variant waves while remaining data-driven rather than variant-defined. Additionally, the approach provides sufficient samples within each phase for robust statistical analysis. This systematic segmentation enables us to evaluate how model fairness metrics vary not only across different case volumes but also across distinct epidemic patterns and trajectories throughout the pandemic period.

B.2 Supporting Information for Preprocessing & Model selection

B.2.1 Data Preprocessing

Initial exploratory analysis of the Pinball Loss norm (PBL) distribution revealed several characteristics that influenced our modeling choices. The original distribution exhibited extreme right-skewness with a heavy tail, as evidenced by the substantial difference between the bulk of the distribution and its maximum value (see Figure B.2, top left) and Table B.2. This pattern is typical of error metrics but poses challenges for standard linear regression assumptions.

To address these distributional characteristics, we explored two key preprocessing steps:

1. **Data Trimming:** We observed that the top 1% (14,672 out of 1,541,541 observations) of PBL values were substantially higher than the rest of the distribution (difference between the 99th percentile (1.47×10^{-3}) and the maximum value (5.18×10^{-1})), potentially exerting excessive leverage in regression models. Trimming these extreme values while preserving 99% of observations (Figure B.2, bottom left) improved the stability of the distribution without sacrificing the majority of the data points.
2. **Square Root Transformation:** To further address the remaining right-skewness, we applied a square root transformation to the PBL values. This transformation significantly improved the distribution's properties, producing a more symmetric shape that better satisfies regression assumptions (Figure B.2, bottom right).

Given these distributional characteristics and the strictly positive nature of our dependent variable, we tested both Gamma and Gaussian GLM families with log-link and identity-link functions.

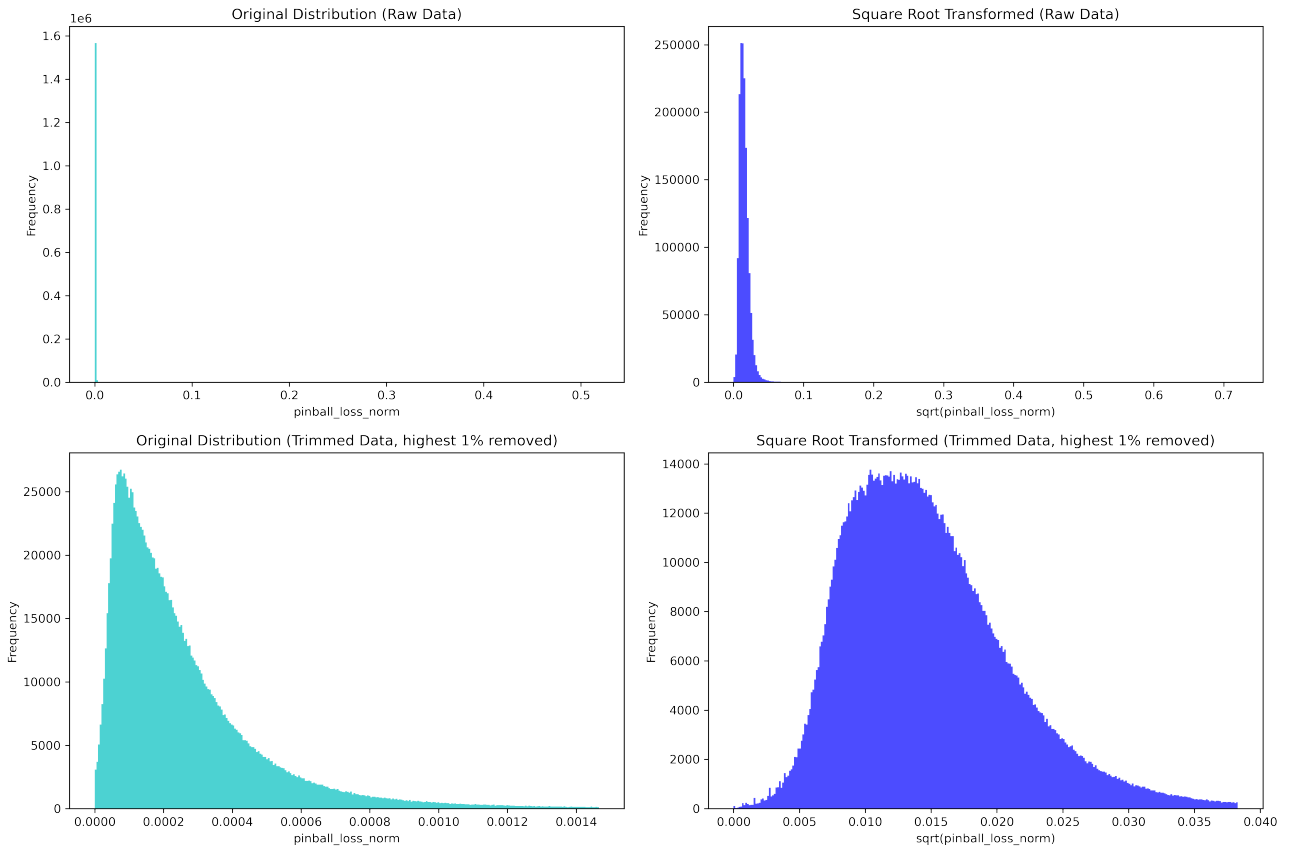


Figure B.2: Distribution of the PBL metric used to evaluate COVID-19 case prediction errors. Top row shows the original (left) and square root transformed (right) distributions of the raw PBL data. Bottom row shows the same distributions after trimming the highest 1% of values, demonstrating improved visualization of the data’s shape and reduction of extreme outliers.

Alternative model specifications, including Gaussian GLMs with identity link and Gamma GLMs with log and inverse link, were evaluated but provided inferior fit based on diagnostic plots, residual patterns, and model performance metrics. The final choice of Gaussian GLM with log link function applied to square root transformed data (with 1% trimming) best addressed both the distributional characteristics of our dependent variable and the heteroscedasticity observed in initial diagnostics. Specific model diagnostics are discussed in Section 2.3.

After evaluating diagnostic plots, residual patterns, and model performance metrics (pseudo- R^2 values), the Gaussian GLM with log-link function applied to the square root transformed data (with 1%

Quantile	Pinball Loss Norm
50th (Median)	1.96×10^{-4}
75th	3.41×10^{-4}
90th	5.53×10^{-4}
95th	7.46×10^{-4}
97th	9.22×10^{-4}
98th	1.09×10^{-3}
99th	1.47×10^{-3}
Maximum	5.18×10^{-1}

Table B.2: Distribution of Pinball Loss Norm Values

trimming) provided the best fit. This combination effectively addressed both the heavy-tailed nature of the original distribution and the heteroscedasticity observed in initial model diagnostics.

	Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5		Iteration 6		Iteration 7	
Variable	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF	Raw_GVIF	A_GVIF
black_perc	11.36818	3.371673	9.767423	3.125288	9.296005	3.048935	6.806963	2.609016	4.407688	2.099449	2.592352	1.610078	1.321822	1.149705
hispanic_perc	8.925986	2.987639	7.617554	2.759992	6.75456	2.598954	6.385277	2.526911	6.37993	2.525852	2.474057	1.572914	1.159437	1.076772
asian_perc	2.06554	1.437199	2.065375	1.437141	1.918433	1.385075	1.918099	1.384955	1.917993	1.384916	1.457816	1.2074	1.456597	1.206896
factor(lookahead)	1.002012	1.000335	1.002011	1.000335	1.002011	1.000335	1.00201	1.000335	1.00201	1.000335	1.00201	1.000335	1.002006	1.000334
factor(phase)	1.113057	1.008966	1.113048	1.008965	1.113046	1.008965	1.113028	1.008964	1.113026	1.008963	1.113015	1.008963	1.112995	1.008961
factor(modelType)	3.461884	1.167923	3.461752	1.167917	3.461675	1.167914	3.461639	1.167912	3.461401	1.167902	3.461395	1.167902	3.461362	1.167901
factor(mobility_used)	3.529834	1.370688	3.529651	1.37067	3.529602	1.370665	3.529574	1.370663	3.529503	1.370656	3.529393	1.370645	3.529174	1.370624
BPHIGH	12.871	3.587617	12.65528	3.557427	10.84353	3.292951	9.612044	3.10033	NA	NA	NA	NA	NA	NA
CANCER	9.636473	3.104267	9.61278	3.100448	9.537255	3.088245	9.537094	3.088219	9.53316	3.087582	NA	NA	NA	NA
CASTHMA	2.572772	1.603986	2.522557	1.588256	2.515252	1.585955	2.430928	1.559143	2.369442	1.539299	2.202915	1.484222	2.050653	1.43201
CHD	29.43185	5.425113	26.35377	5.133592	25.92193	5.091359	NA	NA	NA	NA	NA	NA	NA	NA
DIABETES	33.15382	5.757935	33.09767	5.753058	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
OBESEITY	3.093372	1.758798	3.070124	1.752177	2.772913	1.665207	2.727348	1.651468	2.361238	1.536632	2.357235	1.535329	2.255838	1.501945
STROKE	52.04237	7.21404	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
COPD	13.47433	3.670739	13.12894	3.623388	12.73936	3.569224	7.084034	2.661585	4.404079	2.09859	4.403759	2.098514	2.958887	1.720141
KIDNEY	47.58153	6.897936	20.7619	4.556523	14.69184	3.832994	7.195956	2.682528	7.126606	2.66957	6.302263	2.510431	NA	NA
TOT_POP_perc_65	1.307825	1.143602	1.305161	1.142437	1.296058	1.138446	1.292535	1.136897	1.279676	1.131228	1.23525	1.111418	1.233578	1.110666

Table B.3: Iterative removal of Health Outcome (HO) variables based on adjusted GVIF < 2

B.2.2 Multicollinearity Analysis and Variable Selection

B.2.2.1 GLM-1

To address multicollinearity concerns in our health outcome variables, we employed an iterative variable selection process using the Generalized Variance Inflation Factor (GVIF). Table B.3 shows the progression of this iterative process, where variables of interest were systematically removed based on an adjusted GVIF threshold of 2. The adjusted GVIF (A_GVIF) is calculated as $GVIF^{1/(2Df)}$,

Variable	Raw_GVIF	Adjusted_GVIF
factor(urbanicity)	2.017	1.192
lookahead	1.002	1.000
phase	1.113	1.009
model_type	3.471	1.168
mobility_used	3.534	1.371
CASTHMA	20.030	4.475
STROKE	56.230	7.499
DIABETES	73.127	8.551
CHD	73.595	8.579
BPHIGH	27.348	5.230
CANCER	10.058	3.171
OBESITY	5.198	2.280
KIDNEY	78.350	8.852
COPD	56.263	7.501
TOT_POP_perc_65	1.572	1.254

Table B.4: GVIF Assessment for GLM-2 (Urbanization Main Effects): Initial Assessment satisfies multicollinearity assumption

providing a scale-invariant measure that accounts for the degrees of freedom of each variable.

At the end of this iterative process, our demographic variables of interest (black_perc, hispanic_perc, asian_perc) and data-model characteristic variables (lookahead, phase, model_type, mobility_used) maintained acceptable GVIF values (< 2) throughout the process, supporting their retention in the final model specification. On the other hand, although not strictly necessary ³, the adjusted GVIF values for the control variables (e.g., CASTHMA or COPD) were also below 2. This systematic approach to variable selection helped ensure our model's stability while retaining the most important health outcome controls that showed minimal collinearity with our variables of interest.

B.2.2.2 GLM-2

Table B.4 presents the GVIF assessment for our urbanization model (GLM-2). Most critically, our key variable of interest, urbanization level (factor(urbanicity)), shows a well-contained adjusted GVIF of 1.192, indicating that multicollinearity does not compromise our ability to assess urban-rural disparities. Similarly, all model characteristic variables maintain low adjusted GVIF values during the first iteration: lookahead (1.000), phase (1.009), model_type (1.168), and mobility_used (1.371).

While the health outcome variables exhibit substantial multicollinearity - with *Chronic Kidney Disease* showing the highest adjusted GVIF (8.852), followed by *CHD* (8.579) and *DIABETES* (8.551) - we retain these variables in the model since the coefficients of the variables of interest (urbanization and data-model characteristics) are not affected (their GVIF values are low), and the performance of the control variables as controls is not impaired.

B.2.3 Model Diagnostics

B.2.3.1 GLM-1

The diagnostic plots for our main effects model (GLM-1) indicate generally satisfactory model performance while highlighting some limitations that warrant discussion (Figure B.3):

Model Fit and Prediction Accuracy: Observed vs. predicted values (Panel A of Figure B.3) shows the distribution of observed values at each prediction level through vertical density strips. The dashed line represents perfect prediction ($y=x$). *Expected:* Short, symmetric vertical distributions centered on the diagonal line with consistent length. *Observed:* Reasonable symmetric clustering around the diagonal at lower values with increasing dispersion at higher predictions, suggesting the

³<https://statisticalhorizons.com/multicollinearity/>

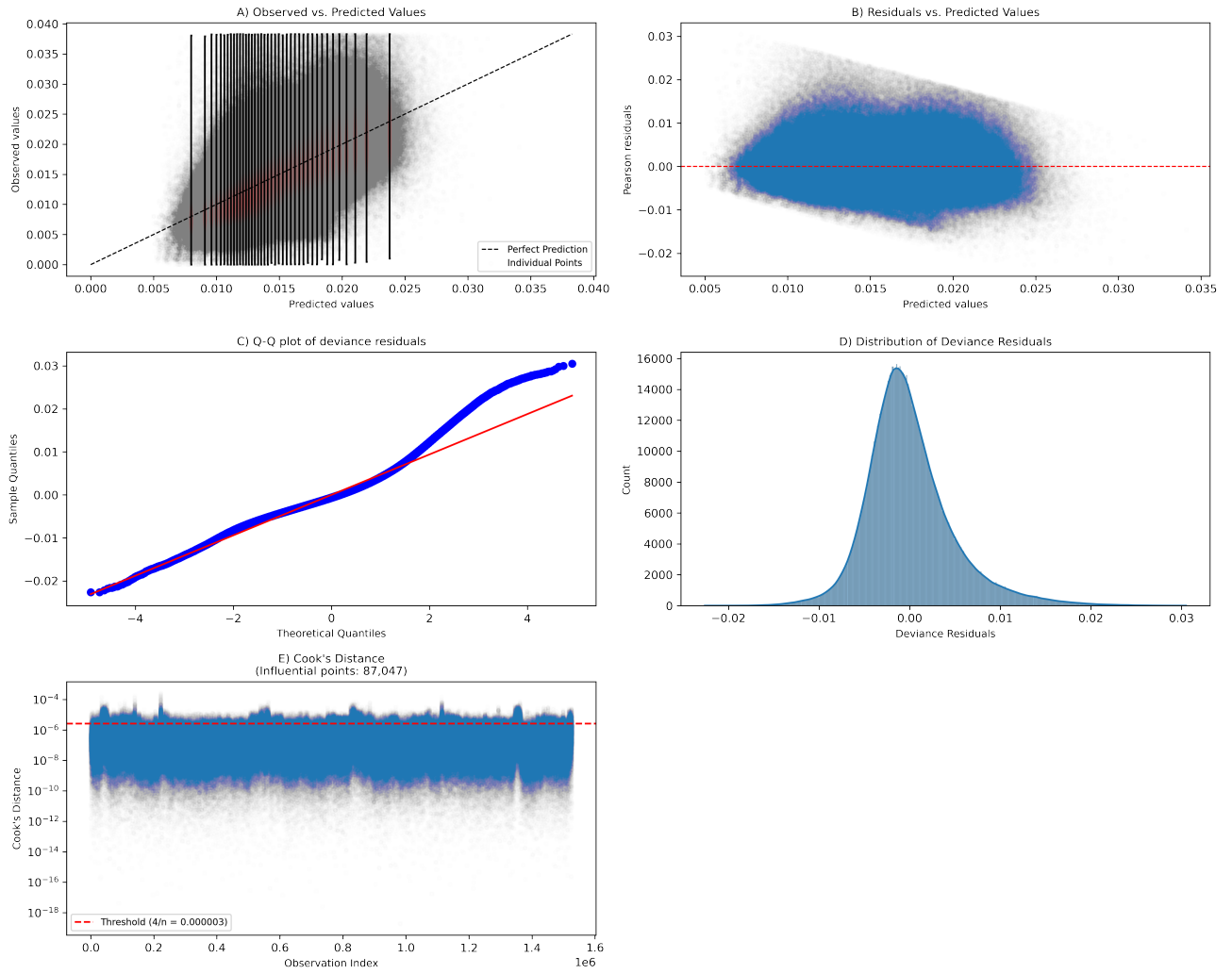


Figure B.3: Diagnostic Plots for GLM-1 (Race and Ethnicity Main Effects Model)

model performs better for smaller values of the transformed PBL.

Residual Analysis: Residuals vs. predicted values (Panel B of Figure B.3) with zero reference line (dashed red). *Expected:* Random scatter around zero with constant spread. *Observed:* Symmetric pattern around zero with mild heteroscedasticity, showing some variation in spread across predicted values.

Normality Assessment: Q-Q plot of deviance residuals (Panel C of Figure B.3) comparing empirical quantiles against theoretical normal distribution. *Expected:* Points following the diagonal line closely if residuals are normally distributed. *Observed:* Good alignment in central region with some

deviation in tails, indicating approximate normality with slightly heavier tails. Even for the distribution of deviance residuals (Panel D of Figure B.3) we see similar patterns of mild heteroscedasticity in the tails. *Expected*: Symmetric, bell-shaped curve centered at zero for normal distribution. *Observed*: Approximately symmetric distribution with slight departure from normality in the tails.

Influence Analysis: Cook's distance plot (Panel E of Figure B.3) identifies 87,047 potentially influential points (approximately 5.7% of observations) above the standard threshold of $4/n$. While this number might seem high, it is not unexpected given our large sample size ($N = 1,526,869$) and the complex nature of COVID-19 forecast errors. Further investigation revealed that removing these points did not substantially alter our main findings regarding racial and ethnic disparities.

These diagnostic results collectively suggest that while our model exhibits some departures from ideal GLM assumptions - particularly in terms of residual heteroscedasticity and influential observations - these limitations are reasonably well-managed through our modeling choices (square root transformation and log-link function) and do not compromise the validity of our main conclusions. The symmetric distribution of residuals, reasonable prediction accuracy, and robustness of findings to influential points support the use of this model specification for investigating racial and ethnic disparities in COVID-19 forecasts. Moreover, the observed patterns in the diagnostics are consistent with what we might expect given the inherent challenges in modeling forecast errors across diverse geographic and demographic contexts during a dynamic pandemic. Importantly, sensitivity analyses confirmed that our key findings regarding demographic disparities in forecast accuracy remain stable even when accounting for these model limitations.

These diagnostic results indicate that our model provides a suitable framework for investigating disparities in COVID-19 forecasts, despite some expected limitations. The combination of square root transformation and log-link function helped improve the distributional properties of our depen-

dent variable by reducing right-skewness and stabilizing variance, although some heteroscedasticity remains visible in the residual plots. The approximate symmetry in residual distribution and generally good alignment in the Q-Q plot's central region suggest that our modeling choices effectively addressed the most severe violations of standard GLM assumptions. While we observe a relatively high number of influential points (5.7% of observations), sensitivity analyses confirmed that our key findings regarding demographic disparities in forecast accuracy remain stable even when these points are excluded.

B.2.3.2 GLM-2

The diagnostic results for our urbanization model (GLM-2) show patterns highly consistent with those observed in GLM-1, indicating stable model performance across our different specifications (Figure B.4):

Model Fit and Prediction Accuracy: Observed vs. predicted values (Panel A) shows the distribution of observed values through vertical density strips, with the dashed line representing perfect prediction. The pattern reveals good prediction accuracy at lower values with increasing dispersion at higher predictions, closely matching what we observed in our race and ethnicity model.

Residual Analysis: The residuals vs. predicted values plot (Panel B) demonstrates symmetric dispersion around zero, with a similar mild heteroscedastic pattern to GLM-1, where the spread of residuals varies somewhat across predicted values.

Normality Assessment: The Q-Q plot (Panel C) and distribution of deviance residuals (Panel D) reveal patterns nearly identical to GLM-1, with good alignment in the central region and slight deviations in the tails. The residual distribution maintains approximate symmetry with minor departures

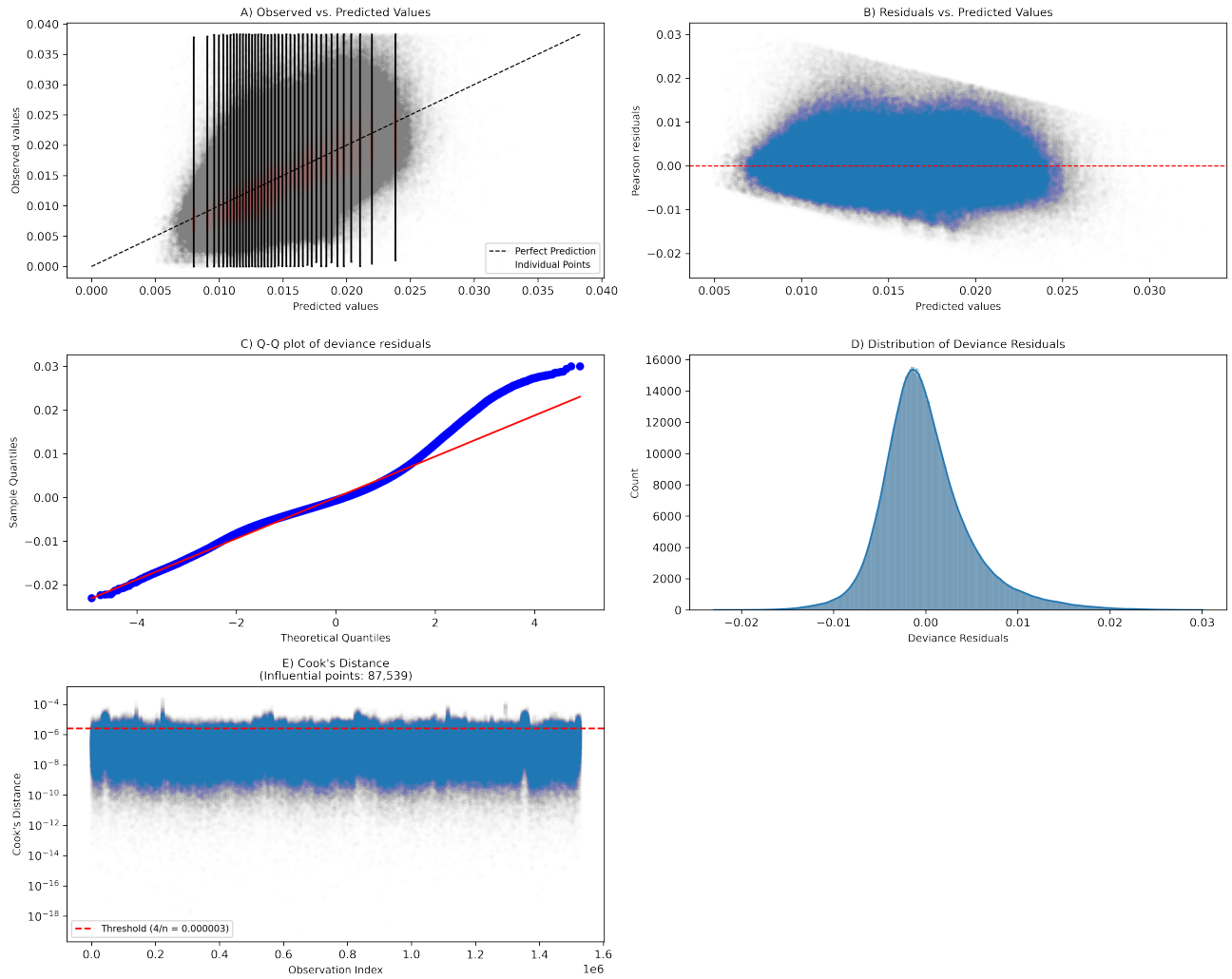


Figure B.4: Diagnostic Plots for GLM-2 (Urbanicity Main Effects Model)

from normality in the extremes.

Influence Analysis: Cook's distance plot (Panel E) identifies 87,539 influential points (5.7% of observations) above the threshold of $4/n$, almost identical to GLM-1's 87,047 points. This consistency across models suggests these influential points reflect inherent characteristics of our COVID-19 forecast data rather than model-specific issues.

These diagnostic results reinforce the appropriateness of our modeling approach. The combination of square root transformation and log-link function effectively addressed the main distributional challenges in our data, as evidenced by the similar patterns across both GLM-1 and GLM-2. While

some heteroscedasticity remains visible in the residual plots, the consistent symmetric distribution of residuals and robustness to influential points provides confidence in our analyses. The striking similarity between GLM-1 and GLM-2 diagnostics suggests that our model specification is handling both race/ethnicity and urbanization predictors in a stable and reliable manner. This consistency across models strengthens our ability to draw meaningful conclusions about disparities in COVID-19 forecast accuracy across different demographic and geographic contexts.

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.301 (0.004)	1.351	[1.342, 1.361]	84.05***
AL	0.091 (0.003)	1.095	[1.088, 1.101]	28.98***
AR	0.074 (0.003)	1.077	[1.070, 1.084]	22.45***
AZ	0.105 (0.004)	1.111	[1.101, 1.120]	24.81***
CA	0.053 (0.003)	1.054	[1.047, 1.061]	15.53***
CO	0.229 (0.003)	1.257	[1.250, 1.265]	73.51***
DE	0.054 (0.009)	1.055	[1.036, 1.074]	5.87***
GA	0.089 (0.003)	1.094	[1.088, 1.100]	31.55***
HI	0.025 (0.011)	1.025	[1.004, 1.047]	2.32*
IA	0.134 (0.003)	1.144	[1.137, 1.151]	44.51***
ID	0.123 (0.003)	1.131	[1.124, 1.138]	39.78***
IL	0.167 (0.003)	1.182	[1.175, 1.188]	60.01***
IN	0.066 (0.003)	1.068	[1.062, 1.074]	23.28***
KS	0.227 (0.003)	1.255	[1.248, 1.262]	83.67***
KY	0.115 (0.003)	1.122	[1.116, 1.128]	42.27***
LA	0.084 (0.003)	1.088	[1.080, 1.095]	24.90***
MA	-0.019 (0.005)	0.981	[0.970, 0.991]	-3.59***
MD	-0.054 (0.004)	0.947	[0.939, 0.955]	-12.61***
ME	-0.059 (0.005)	0.943	[0.934, 0.952]	-11.88***
MI	0.092 (0.003)	1.096	[1.090, 1.102]	34.04***
MN	0.185 (0.003)	1.203	[1.196, 1.209]	64.50***
MO	0.062 (0.003)	1.064	[1.057, 1.070]	19.54***
MS	0.116 (0.003)	1.123	[1.116, 1.130]	35.47***
MT	0.221 (0.003)	1.247	[1.240, 1.254]	77.02***
NC	0.060 (0.003)	1.062	[1.056, 1.068]	20.19***
ND	0.230 (0.003)	1.258	[1.250, 1.266]	68.57***
NE	0.123 (0.003)	1.131	[1.123, 1.138]	36.15***
NH	0.044 (0.006)	1.045	[1.034, 1.057]	7.79***
NJ	0.035 (0.004)	1.036	[1.027, 1.044]	8.14***
NM	0.041 (0.004)	1.042	[1.035, 1.050]	11.62***
NV	0.037 (0.004)	1.037	[1.028, 1.046]	8.33***
OH	-0.031 (0.003)	0.969	[0.963, 0.975]	-10.21***
OK	0.145 (0.003)	1.156	[1.150, 1.162]	53.63***
OR	-0.007 (0.004)	0.993	[0.986, 1.000]	-1.97*
PA	0.007 (0.003)	1.007	[1.001, 1.013]	2.32*
RI	0.236 (0.007)	1.267	[1.250, 1.284]	34.20***
SC	0.076 (0.004)	1.079	[1.072, 1.087]	21.10***
SD	0.177 (0.003)	1.194	[1.187, 1.201]	56.13***
TN	0.146 (0.003)	1.157	[1.151, 1.164]	50.35***
TX	0.113 (0.003)	1.119	[1.113, 1.125]	39.15***
UT	-0.256 (0.005)	0.774	[0.766, 0.782]	-51.03***
VA	0.075 (0.003)	1.078	[1.072, 1.083]	28.32***
VT	0.022 (0.005)	1.022	[1.012, 1.033]	4.31***
WA	0.046 (0.003)	1.047	[1.040, 1.054]	13.45***
WI	0.179 (0.003)	1.196	[1.189, 1.202]	64.20***
WV	0.055 (0.003)	1.056	[1.050, 1.062]	18.12***
WY	0.184 (0.004)	1.202	[1.194, 1.211]	49.17***

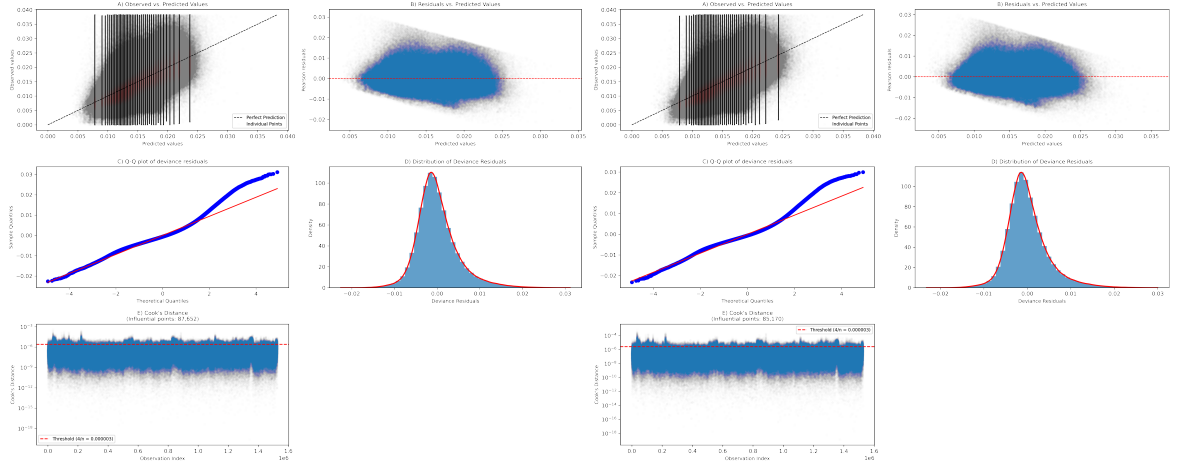
Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL (trimmed).

Table B.5: State Fixed Effects for GLM-1 (Reference State: NY)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.181 (0.005)	1.199	[1.188, 1.209]	39.74***
AL	0.076 (0.004)	1.079	[1.071, 1.086]	21.16***
AR	0.030 (0.004)	1.030	[1.022, 1.038]	7.50***
AZ	0.095 (0.005)	1.099	[1.089, 1.110]	18.86***
CA	0.027 (0.004)	1.027	[1.019, 1.034]	7.20***
CO	0.222 (0.004)	1.249	[1.241, 1.258]	62.84***
DE	0.039 (0.009)	1.040	[1.021, 1.059]	4.24***
GA	0.065 (0.004)	1.067	[1.059, 1.074]	18.11***
HI	-0.203 (0.011)	0.816	[0.799, 0.833]	-19.02***
IA	0.106 (0.004)	1.112	[1.104, 1.120]	29.47***
ID	0.091 (0.003)	1.095	[1.088, 1.102]	26.39***
IL	0.149 (0.003)	1.160	[1.153, 1.167]	47.61***
IN	0.085 (0.003)	1.089	[1.082, 1.095]	28.47***
KS	0.183 (0.003)	1.200	[1.193, 1.207]	60.22***
KY	0.148 (0.003)	1.160	[1.153, 1.167]	50.60***
LA	0.081 (0.004)	1.084	[1.076, 1.092]	20.73***
MA	-0.032 (0.005)	0.969	[0.958, 0.979]	-5.80***
MD	-0.069 (0.005)	0.933	[0.925, 0.941]	-15.19***
ME	-0.080 (0.005)	0.924	[0.914, 0.933]	-15.69***
MI	0.070 (0.003)	1.072	[1.066, 1.079]	23.41***
MN	0.115 (0.003)	1.122	[1.114, 1.129]	33.32***
MO	0.062 (0.003)	1.064	[1.057, 1.071]	18.01***
MS	0.077 (0.004)	1.080	[1.072, 1.089]	18.94***
MT	0.172 (0.003)	1.187	[1.179, 1.195]	50.67***
NC	0.031 (0.003)	1.031	[1.024, 1.038]	8.88***
ND	0.142 (0.004)	1.153	[1.143, 1.162]	33.67***
NE	0.071 (0.004)	1.074	[1.065, 1.082]	17.77***
NH	0.035 (0.006)	1.035	[1.024, 1.047]	6.03***
NJ	0.008 (0.005)	1.008	[0.999, 1.017]	1.84
NM	0.041 (0.004)	1.042	[1.034, 1.050]	10.60***
NV	0.020 (0.005)	1.020	[1.011, 1.029]	4.23***
OH	-0.043 (0.003)	0.958	[0.952, 0.964]	-13.39***
OK	0.117 (0.003)	1.124	[1.117, 1.131]	37.54***
OR	-0.032 (0.004)	0.969	[0.962, 0.976]	-8.50***
PA	-0.023 (0.003)	0.977	[0.971, 0.984]	-7.17***
RI	0.246 (0.007)	1.279	[1.261, 1.297]	34.90***
SC	0.059 (0.004)	1.061	[1.053, 1.070]	14.42***
SD	0.083 (0.004)	1.086	[1.077, 1.095]	19.85***
TN	0.167 (0.003)	1.182	[1.175, 1.189]	53.57***
TX	0.127 (0.003)	1.135	[1.128, 1.142]	38.50***
UT	-0.295 (0.005)	0.744	[0.737, 0.752]	-56.03***
VA	0.091 (0.003)	1.096	[1.089, 1.102]	30.99***
VT	-0.020 (0.005)	0.980	[0.970, 0.990]	-3.80***
WA	0.033 (0.004)	1.033	[1.026, 1.041]	9.01***
WI	0.117 (0.003)	1.125	[1.117, 1.132]	35.19***
WV	0.060 (0.004)	1.062	[1.055, 1.069]	17.08***
WY	0.156 (0.004)	1.169	[1.160, 1.178]	38.44***

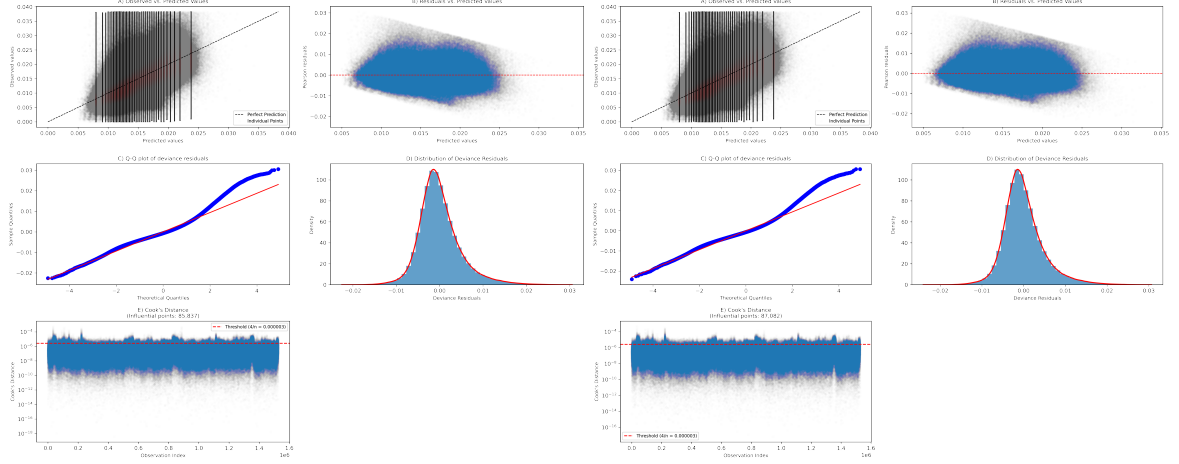
Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL (trimmed).

Table B.6: State Fixed Effects for GLM-2 (Reference State: NY)



(a) Race and Forecast Lookahead

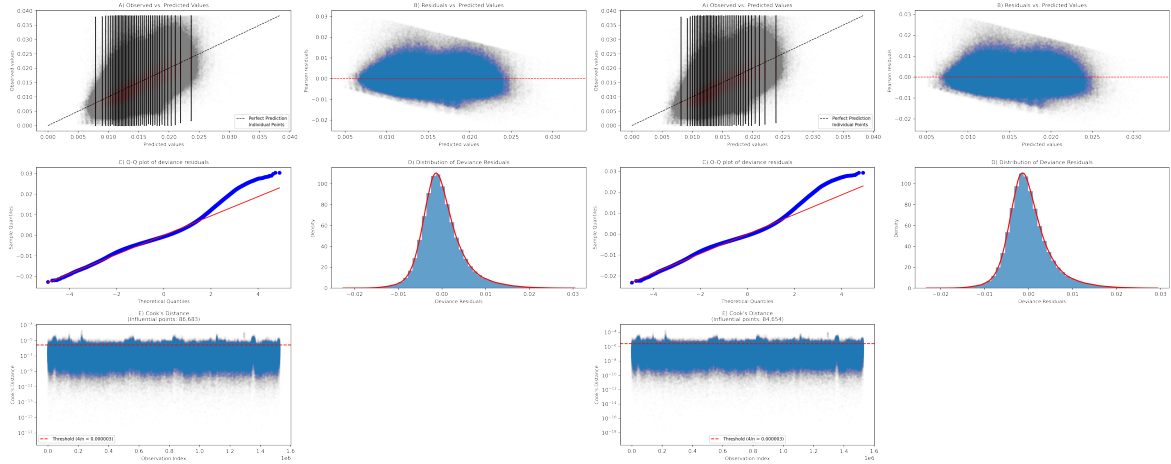
(b) Race and Pandemic Phase



(c) Race and Model Type

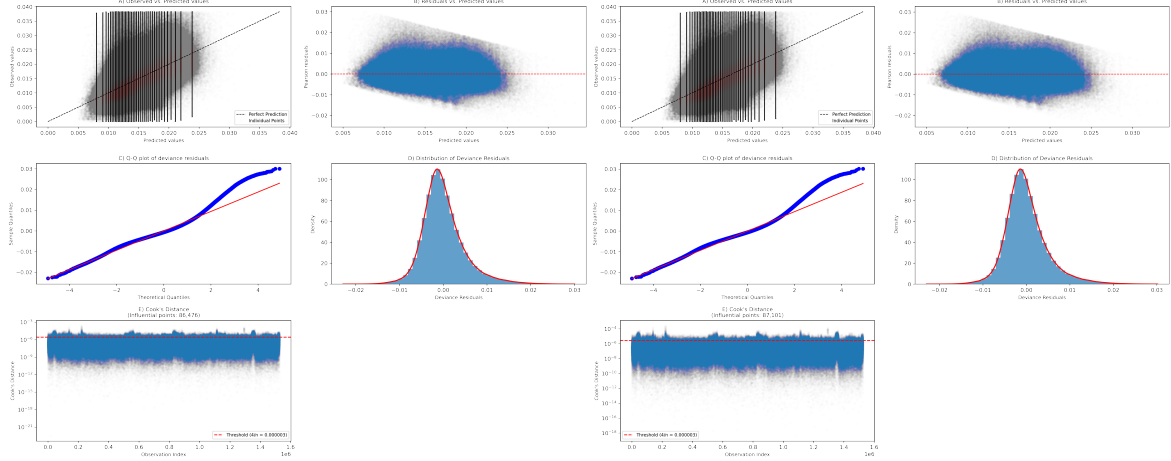
(d) Race and Mobility Usage

Figure B.5: Regression diagnostics for the Generalized Linear Models (GLM) analyzing interactions between race/ethnicity and model characteristics: (a) Forecast lookahead interactions (GLM-1a) showing diagnostic plots for models analyzing how prediction horizons (7-28 days) affect racial disparities; (b) Pandemic phase interactions (GLM-1b) presenting diagnostic assessments for models examining how racial disparities vary across different COVID-19 phases; (c) Model type interactions (GLM-1c) showing residual patterns for analyzing racial disparities across different model architectures (compartmental, deep learning, statistical, etc.); and (d) Mobility data usage interactions (GLM-1d) displaying diagnostics for models examining how the incorporation of mobility data impacts prediction fairness across racial groups. Each subplot contains standard regression diagnostic plots: observed vs. fitted values (top left) residuals vs. fitted values (top right), normal Q-Q plot (middle left), scale-location plot (bottom left), and distribution of deviance residuals (bottom right) to validate model assumptions of normality, homoscedasticity, and identify potential influential observations.



(a) Urbanicity and Lookahead

(b) Urbanicity and Pandemic Phase



(c) Urbanicity and Model Type

(d) Urbanicity and Mobility Usage

Figure B.6: Regression diagnostics for the Generalized Linear Models (GLM) analyzing interactions between urbanicity and model characteristics: (a) Forecast lookahead interactions (GLM-1a) showing diagnostic plots for models analyzing how prediction horizons (7-28 days) affect racial disparities; (b) Pandemic phase interactions (GLM-1b) presenting diagnostic assessments for models examining how racial disparities vary across different COVID-19 phases; (c) Model type interactions (GLM-1c) showing residual patterns for analyzing racial disparities across different model architectures (compartmental, deep learning, statistical, etc.); and (d) Mobility data usage interactions (GLM-1d) displaying diagnostics for models examining how the incorporation of mobility data impacts prediction fairness across racial groups. Each subplot contains standard regression diagnostic plots: observed vs. fitted values (top left) residuals vs. fitted values (top right), normal Q-Q plot (middle left), scale-location plot (bottom left), and distribution of deviance residuals (bottom right) to validate model assumptions of normality, homoscedasticity, and identify potential influential observations.

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.108 (0.001)	1.114	[1.111, 1.116]	93.23***
21-day ahead	0.183 (0.001)	1.201	[1.199, 1.204]	162.93***
28-day ahead	0.251 (0.001)	1.285	[1.283, 1.288]	230.21***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.114 (0.002)	1.121	[1.118, 1.124]	73.02***
Deep Learning	0.037 (0.001)	1.037	[1.035, 1.039]	40.15***
Ensemble	0.007 (0.001)	1.007	[1.005, 1.009]	7.97***
Statistical	0.073 (0.001)	1.076	[1.074, 1.078]	89.84***
<i>Mobility Used (ref: No)</i>				
Mixed	-0.146 (0.001)	0.864	[0.862, 0.866]	-124.31***
Yes	-0.004 (0.001)	0.996	[0.994, 0.997]	-5.55***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.445	[1.442, 1.448]	377.45***
Phase 2	-0.116 (0.001)	0.891	[0.888, 0.893]	-99.57***
Phase 3	0.251 (0.001)	1.285	[1.283, 1.288]	236.62***
Phase 4	0.440 (0.001)	1.553	[1.550, 1.556]	417.08***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.76***
Phase 6	0.141 (0.001)	1.152	[1.149, 1.154]	112.44***
<i>Health Outcomes</i>				
Asthma	-0.041 (0.001)	0.959	[0.958, 0.961]	-43.32***
Obesity	0.003 (0.0001)	1.003	[1.003, 1.004]	25.28***
COPD	0.045 (0.0005)	1.046	[1.045, 1.047]	92.63***
Age 65+	0.190 (0.007)	1.209	[1.193, 1.225]	28.05***
Intercept	-4.686 (0.009)	0.009	[0.009, 0.009]	-508.57***
Notes: Race $\times$ Lookahead interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.7: Main Effects for GLM-1a (Race $\times$ Lookahead Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.299 (0.004)	1.349	[1.339, 1.358]	83.52***
AL	0.091 (0.003)	1.095	[1.088, 1.102]	29.04***
AR	0.075 (0.003)	1.077	[1.070, 1.084]	22.65***
AZ	0.104 (0.004)	1.110	[1.101, 1.119]	24.70***
CA	0.054 (0.003)	1.056	[1.049, 1.063]	16.03***
CO	0.228 (0.003)	1.256	[1.248, 1.263]	73.17***
DE	0.054 (0.009)	1.056	[1.037, 1.075]	5.96***
GA	0.090 (0.003)	1.094	[1.088, 1.100]	31.60***
HI	0.020 (0.011)	1.020	[0.999, 1.042]	1.84
IA	0.135 (0.003)	1.144	[1.138, 1.151]	44.69***
ID	0.123 (0.003)	1.131	[1.124, 1.138]	39.64***
IL	0.167 (0.003)	1.182	[1.175, 1.188]	60.12***
IN	0.066 (0.003)	1.068	[1.062, 1.074]	23.35***
KS	0.227 (0.003)	1.255	[1.249, 1.262]	83.85***
KY	0.116 (0.003)	1.123	[1.117, 1.129]	42.39***
LA	0.084 (0.003)	1.088	[1.081, 1.095]	25.02***
MA	-0.019 (0.005)	0.981	[0.971, 0.991]	-3.57***
MD	-0.054 (0.004)	0.948	[0.940, 0.956]	-12.48***
ME	-0.059 (0.005)	0.942	[0.933, 0.951]	-12.05***
MI	0.092 (0.003)	1.096	[1.090, 1.102]	34.01***
MN	0.185 (0.003)	1.203	[1.196, 1.210]	64.60***
MO	0.062 (0.003)	1.064	[1.058, 1.071]	19.79***
MS	0.116 (0.003)	1.123	[1.116, 1.130]	35.51***
MT	0.220 (0.003)	1.246	[1.239, 1.253]	76.92***
NC	0.061 (0.003)	1.062	[1.056, 1.069]	20.31***
ND	0.230 (0.003)	1.259	[1.250, 1.267]	68.76***
NE	0.123 (0.003)	1.131	[1.124, 1.139]	36.40***
NH	0.044 (0.006)	1.045	[1.033, 1.056]	7.73***
NJ	0.036 (0.004)	1.037	[1.028, 1.046]	8.45***
NM	0.041 (0.004)	1.042	[1.034, 1.049]	11.49***
NV	0.037 (0.004)	1.038	[1.029, 1.047]	8.50***
OH	-0.031 (0.003)	0.970	[0.964, 0.976]	-10.03***
OK	0.145 (0.003)	1.156	[1.150, 1.162]	53.68***
OR	-0.007 (0.004)	0.993	[0.986, 1.000]	-2.05*
PA	0.007 (0.003)	1.007	[1.001, 1.013]	2.34*
RI	0.236 (0.007)	1.266	[1.249, 1.283]	34.10***
SC	0.077 (0.004)	1.080	[1.072, 1.087]	21.21***
SD	0.178 (0.003)	1.194	[1.187, 1.202]	56.27***
TN	0.146 (0.003)	1.158	[1.151, 1.164]	50.44***
TX	0.113 (0.003)	1.120	[1.113, 1.126]	39.37***
UT	-0.257 (0.005)	0.774	[0.766, 0.781]	-51.19***
VA	0.075 (0.003)	1.078	[1.072, 1.083]	28.39***
VT	0.021 (0.005)	1.021	[1.011, 1.031]	4.11***
WA	0.045 (0.003)	1.047	[1.040, 1.054]	13.35***
WI	0.178 (0.003)	1.195	[1.189, 1.202]	64.17***
WV	0.055 (0.003)	1.057	[1.050, 1.063]	18.24***
WY	0.184 (0.004)	1.202	[1.193, 1.211]	49.10***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.8: State Fixed Effects for GLM-1a (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.112 (0.001)	1.119	[1.117, 1.121]	146.51***
21-day ahead	0.191 (0.001)	1.210	[1.208, 1.212]	254.30***
28-day ahead	0.261 (0.001)	1.299	[1.297, 1.301]	359.63***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.113 (0.002)	1.120	[1.117, 1.123]	74.07***
Deep Learning	0.036 (0.001)	1.037	[1.035, 1.039]	40.93***
Ensemble	0.007 (0.001)	1.007	[1.005, 1.008]	7.71***
Statistical	0.072 (0.001)	1.074	[1.073, 1.076]	89.95***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.146 (0.001)	0.864	[0.862, 0.866]	-126.55***
Yes	-0.005 (0.001)	0.995	[0.994, 0.997]	-6.17***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.481 (0.001)	1.618	[1.613, 1.623]	328.46***
Phase 2	0.009 (0.002)	1.009	[1.006, 1.013]	5.29***
Phase 3	0.373 (0.002)	1.453	[1.448, 1.457]	235.62***
Phase 4	0.618 (0.002)	1.856	[1.850, 1.862]	401.03***
Phase 5	0.090 (0.002)	1.094	[1.090, 1.098]	45.57***
Phase 6	0.252 (0.002)	1.286	[1.282, 1.291]	136.28***
<i>Health Outcomes</i>				
Asthma	-0.041 (0.001)	0.960	[0.958, 0.962]	-43.72***
Obesity	0.003 (0.0001)	1.003	[1.003, 1.003]	22.83***
COPD	0.044 (0.0005)	1.045	[1.044, 1.046]	93.22***
Age 65+	0.156 (0.007)	1.169	[1.154, 1.184]	23.51***
Intercept	-4.779 (0.009)	0.008	[0.008, 0.009]	-528.25***

Notes: Race $\times$ Phase interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.9: Main Effects for GLM-1b (Race $\times$ Phase Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.290 (0.004)	1.337	[1.328, 1.346]	82.71***
AL	0.084 (0.003)	1.088	[1.082, 1.095]	27.64***
AR	0.071 (0.003)	1.073	[1.066, 1.080]	21.95***
AZ	0.099 (0.004)	1.104	[1.095, 1.112]	23.86***
CA	0.044 (0.003)	1.045	[1.039, 1.052]	13.38***
CO	0.217 (0.003)	1.242	[1.235, 1.250]	71.21***
DE	0.052 (0.009)	1.054	[1.035, 1.072]	5.82***
GA	0.089 (0.003)	1.093	[1.087, 1.099]	32.08***
HI	0.002 (0.010)	1.002	[0.982, 1.023]	0.21
IA	0.125 (0.003)	1.133	[1.127, 1.140]	42.33***
ID	0.112 (0.003)	1.119	[1.112, 1.125]	37.06***
IL	0.160 (0.003)	1.173	[1.167, 1.180]	58.96***
IN	0.058 (0.003)	1.060	[1.054, 1.066]	21.23***
KS	0.221 (0.003)	1.248	[1.241, 1.254]	83.66***
KY	0.109 (0.003)	1.116	[1.110, 1.122]	41.13***
LA	0.088 (0.003)	1.092	[1.085, 1.099]	26.87***
MA	-0.017 (0.005)	0.983	[0.973, 0.993]	-3.28**
MD	-0.059 (0.004)	0.943	[0.935, 0.951]	-13.91***
ME	-0.061 (0.005)	0.941	[0.932, 0.950]	-12.76***
MI	0.088 (0.003)	1.092	[1.086, 1.097]	33.35***
MN	0.179 (0.003)	1.196	[1.189, 1.203]	64.10***
MO	0.055 (0.003)	1.057	[1.051, 1.063]	17.94***
MS	0.112 (0.003)	1.118	[1.111, 1.125]	35.08***
MT	0.209 (0.003)	1.232	[1.225, 1.239]	74.63***
NC	0.056 (0.003)	1.057	[1.051, 1.063]	19.11***
ND	0.219 (0.003)	1.245	[1.237, 1.253]	66.95***
NE	0.116 (0.003)	1.123	[1.116, 1.131]	35.04***
NH	0.047 (0.005)	1.048	[1.037, 1.060]	8.64***
NJ	0.041 (0.004)	1.042	[1.033, 1.051]	9.80***
NM	0.023 (0.003)	1.023	[1.016, 1.030]	6.53***
NV	0.033 (0.004)	1.033	[1.025, 1.042]	7.67***
OH	-0.035 (0.003)	0.965	[0.960, 0.971]	-11.88***
OK	0.138 (0.003)	1.148	[1.142, 1.154]	52.25***
OR	-0.013 (0.003)	0.987	[0.980, 0.994]	-3.82***
PA	0.004 (0.003)	1.004	[0.998, 1.009]	1.27
RI	0.233 (0.007)	1.262	[1.246, 1.279]	34.62***
SC	0.077 (0.004)	1.081	[1.073, 1.088]	21.92***
SD	0.167 (0.003)	1.181	[1.174, 1.189]	54.02***
TN	0.138 (0.003)	1.148	[1.142, 1.155]	48.83***
TX	0.110 (0.003)	1.116	[1.110, 1.122]	39.11***
UT	-0.268 (0.005)	0.765	[0.757, 0.772]	-54.66***
VA	0.071 (0.003)	1.073	[1.068, 1.079]	27.49***
VT	0.020 (0.005)	1.020	[1.010, 1.030]	4.02***
WA	0.040 (0.003)	1.040	[1.034, 1.047]	11.89***
WI	0.170 (0.003)	1.185	[1.178, 1.191]	62.46***
WV	0.051 (0.003)	1.052	[1.046, 1.058]	17.29***
WY	0.173 (0.004)	1.189	[1.181, 1.198]	47.32***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.10: State Fixed Effects for GLM-1b (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.113 (0.001)	1.119	[1.117, 1.121]	143.58***
21-day ahead	0.191 (0.001)	1.211	[1.209, 1.213]	249.41***
28-day ahead	0.262 (0.001)	1.300	[1.298, 1.302]	352.74***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.103 (0.002)	1.109	[1.104, 1.113]	51.37***
Deep Learning	0.044 (0.001)	1.045	[1.042, 1.048]	33.29***
Ensemble	0.014 (0.001)	1.014	[1.012, 1.016]	12.54***
Statistical	0.080 (0.001)	1.084	[1.081, 1.086]	70.34***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.147 (0.001)	0.864	[0.862, 0.866]	-124.45***
Yes	-0.005 (0.001)	0.996	[0.994, 0.997]	-5.90***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.445	[1.442, 1.447]	376.90***
Phase 2	-0.116 (0.001)	0.890	[0.888, 0.892]	-99.91***
Phase 3	0.251 (0.001)	1.285	[1.282, 1.288]	236.23***
Phase 4	0.440 (0.001)	1.552	[1.549, 1.556]	416.68***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.96***
Phase 6	0.141 (0.001)	1.151	[1.148, 1.154]	112.17***
<i>Health Outcomes</i>				
Asthma	-0.042 (0.001)	0.959	[0.957, 0.961]	-43.77***
Obesity	0.003 (0.0001)	1.003	[1.003, 1.004]	25.92***
COPD	0.045 (0.0005)	1.046	[1.045, 1.047]	93.07***
Age 65+	0.194 (0.007)	1.214	[1.199, 1.231]	28.74***
Intercept	-4.697 (0.009)	0.009	[0.009, 0.009]	-510.25***
Notes: Race $\times$ Model Type interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.11: Main Effects for GLM-1c (Race $\times$ Model Type Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.301 (0.004)	1.351	[1.342, 1.361]	84.05***
AL	0.090 (0.003)	1.095	[1.088, 1.101]	28.98***
AR	0.074 (0.003)	1.077	[1.070, 1.084]	22.44***
AZ	0.105 (0.004)	1.110	[1.101, 1.120]	24.80***
CA	0.053 (0.003)	1.054	[1.047, 1.062]	15.62***
CO	0.229 (0.003)	1.258	[1.250, 1.265]	73.56***
DE	0.054 (0.009)	1.055	[1.036, 1.074]	5.85***
GA	0.089 (0.003)	1.093	[1.087, 1.100]	31.52***
HI	0.025 (0.011)	1.025	[1.004, 1.047]	2.31*
IA	0.134 (0.003)	1.144	[1.137, 1.150]	44.46***
ID	0.123 (0.003)	1.131	[1.124, 1.138]	39.72***
IL	0.167 (0.003)	1.181	[1.175, 1.188]	59.98***
IN	0.065 (0.003)	1.068	[1.062, 1.074]	23.24***
KS	0.227 (0.003)	1.255	[1.248, 1.261]	83.65***
KY	0.115 (0.003)	1.122	[1.116, 1.128]	42.23***
LA	0.084 (0.003)	1.087	[1.080, 1.095]	24.83***
MA	-0.019 (0.005)	0.981	[0.970, 0.991]	-3.58***
MD	-0.055 (0.004)	0.947	[0.939, 0.955]	-12.68***
ME	-0.059 (0.005)	0.943	[0.934, 0.952]	-11.92***
MI	0.092 (0.003)	1.096	[1.090, 1.102]	33.99***
MN	0.184 (0.003)	1.202	[1.196, 1.209]	64.44***
MO	0.062 (0.003)	1.064	[1.057, 1.070]	19.50***
MS	0.116 (0.003)	1.123	[1.116, 1.130]	35.46***
MT	0.220 (0.003)	1.247	[1.240, 1.254]	76.96***
NC	0.060 (0.003)	1.062	[1.056, 1.068]	20.16***
ND	0.229 (0.003)	1.258	[1.249, 1.266]	68.51***
NE	0.123 (0.003)	1.130	[1.123, 1.138]	36.15***
NH	0.044 (0.006)	1.045	[1.033, 1.057]	7.77***
NJ	0.035 (0.004)	1.036	[1.027, 1.044]	8.14***
NM	0.041 (0.004)	1.042	[1.035, 1.049]	11.56***
NV	0.036 (0.004)	1.037	[1.028, 1.046]	8.29***
OH	-0.031 (0.003)	0.969	[0.963, 0.975]	-10.25***
OK	0.145 (0.003)	1.156	[1.150, 1.162]	53.60***
OR	-0.007 (0.004)	0.993	[0.986, 1.000]	-2.03*
PA	0.007 (0.003)	1.007	[1.001, 1.013]	2.27*
RI	0.236 (0.007)	1.267	[1.249, 1.284]	34.18***
SC	0.076 (0.004)	1.079	[1.072, 1.087]	21.08***
SD	0.177 (0.003)	1.194	[1.186, 1.201]	56.09***
TN	0.146 (0.003)	1.157	[1.151, 1.164]	50.32***
TX	0.113 (0.003)	1.119	[1.113, 1.125]	39.16***
UT	-0.256 (0.005)	0.774	[0.766, 0.782]	-51.05***
VA	0.075 (0.003)	1.077	[1.072, 1.083]	28.24***
VT	0.022 (0.005)	1.022	[1.012, 1.032]	4.30***
WA	0.046 (0.003)	1.047	[1.040, 1.054]	13.36***
WI	0.178 (0.003)	1.195	[1.189, 1.202]	64.15***
WV	0.055 (0.003)	1.056	[1.050, 1.062]	18.08***
WY	0.184 (0.004)	1.202	[1.193, 1.211]	49.13***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.12: State Fixed Effects for GLM-1c (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.113 (0.001)	1.119	[1.118, 1.121]	143.64***
21-day ahead	0.191 (0.001)	1.211	[1.209, 1.213]	249.53***
28-day ahead	0.262 (0.001)	1.300	[1.298, 1.302]	352.84***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.114 (0.002)	1.121	[1.118, 1.125]	73.05***
Deep Learning	0.037 (0.001)	1.037	[1.036, 1.039]	40.30***
Ensemble	0.007 (0.001)	1.007	[1.005, 1.009]	8.11***
Statistical	0.073 (0.001)	1.076	[1.074, 1.077]	89.57***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.137 (0.001)	0.872	[0.870, 0.875]	-92.95***
Yes	0.010 (0.001)	1.010	[1.008, 1.012]	9.76***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.444	[1.442, 1.447]	376.83***
Phase 2	-0.117 (0.001)	0.890	[0.888, 0.892]	-100.06***
Phase 3	0.250 (0.001)	1.285	[1.282, 1.287]	236.15***
Phase 4	0.440 (0.001)	1.552	[1.549, 1.555]	416.73***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.94***
Phase 6	0.141 (0.001)	1.151	[1.148, 1.154]	111.97***
<i>Health Outcomes</i>				
Asthma	-0.042 (0.001)	0.959	[0.957, 0.961]	-43.86***
Obesity	0.003 (0.0001)	1.003	[1.003, 1.004]	25.84***
COPD	0.045 (0.0005)	1.046	[1.045, 1.047]	93.25***
Age 65+	0.194 (0.007)	1.214	[1.198, 1.230]	28.63***
Intercept	-4.704 (0.009)	0.009	[0.009, 0.009]	-510.35***

Notes: Race $\times$ Mobility interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.13: Main Effects for GLM-1d (Race $\times$ Mobility Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.301 (0.004)	1.351	[1.342, 1.361]	84.17***
AL	0.090 (0.003)	1.094	[1.088, 1.101]	28.83***
AR	0.074 (0.003)	1.076	[1.069, 1.083]	22.34***
AZ	0.104 (0.004)	1.110	[1.101, 1.119]	24.68***
CA	0.052 (0.003)	1.053	[1.046, 1.060]	15.22***
CO	0.229 (0.003)	1.257	[1.249, 1.265]	73.44***
DE	0.053 (0.009)	1.055	[1.036, 1.074]	5.83***
GA	0.089 (0.003)	1.093	[1.087, 1.099]	31.48***
HI	0.027 (0.011)	1.028	[1.006, 1.049]	2.55*
IA	0.134 (0.003)	1.143	[1.136, 1.150]	44.32***
ID	0.123 (0.003)	1.131	[1.124, 1.138]	39.69***
IL	0.166 (0.003)	1.181	[1.175, 1.188]	59.90***
IN	0.065 (0.003)	1.067	[1.061, 1.073]	23.14***
KS	0.227 (0.003)	1.254	[1.248, 1.261]	83.62***
KY	0.115 (0.003)	1.122	[1.116, 1.128]	42.13***
LA	0.084 (0.003)	1.087	[1.080, 1.094]	24.82***
MA	-0.019 (0.005)	0.981	[0.971, 0.992]	-3.48***
MD	-0.055 (0.004)	0.947	[0.939, 0.955]	-12.70***
ME	-0.058 (0.005)	0.943	[0.934, 0.952]	-11.86***
MI	0.091 (0.003)	1.096	[1.090, 1.102]	33.96***
MN	0.184 (0.003)	1.202	[1.196, 1.209]	64.44***
MO	0.061 (0.003)	1.063	[1.057, 1.070]	19.37***
MS	0.116 (0.003)	1.123	[1.116, 1.130]	35.42***
MT	0.220 (0.003)	1.246	[1.239, 1.253]	76.88***
NC	0.060 (0.003)	1.062	[1.056, 1.068]	20.10***
ND	0.229 (0.003)	1.257	[1.249, 1.266]	68.42***
NE	0.122 (0.003)	1.130	[1.122, 1.138]	36.02***
NH	0.044 (0.006)	1.045	[1.034, 1.057]	7.85***
NJ	0.035 (0.004)	1.036	[1.027, 1.045]	8.20***
NM	0.041 (0.004)	1.041	[1.034, 1.049]	11.41***
NV	0.036 (0.004)	1.037	[1.028, 1.046]	8.27***
OH	-0.032 (0.003)	0.969	[0.963, 0.975]	-10.33***
OK	0.145 (0.003)	1.156	[1.150, 1.162]	53.51***
OR	-0.007 (0.004)	0.993	[0.986, 1.000]	-2.00*
PA	0.007 (0.003)	1.007	[1.001, 1.013]	2.28*
RI	0.236 (0.007)	1.267	[1.250, 1.284]	34.19***
SC	0.076 (0.004)	1.079	[1.072, 1.087]	21.11***
SD	0.177 (0.003)	1.193	[1.186, 1.201]	55.94***
TN	0.146 (0.003)	1.157	[1.150, 1.163]	50.20***
TX	0.112 (0.003)	1.119	[1.113, 1.125]	39.08***
UT	-0.257 (0.005)	0.773	[0.766, 0.781]	-51.24***
VA	0.075 (0.003)	1.077	[1.072, 1.083]	28.23***
VT	0.022 (0.005)	1.022	[1.012, 1.033]	4.32***
WA	0.046 (0.003)	1.047	[1.040, 1.054]	13.48***
WI	0.178 (0.003)	1.195	[1.189, 1.202]	64.13***
WV	0.054 (0.003)	1.056	[1.050, 1.062]	18.03***
WY	0.184 (0.004)	1.202	[1.193, 1.211]	49.07***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.14: State Fixed Effects for GLM-1d (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.142 (0.002)	1.152	[1.147, 1.157]	58.71***
21-day ahead	0.249 (0.002)	1.283	[1.277, 1.288]	106.78***
28-day ahead	0.340 (0.002)	1.404	[1.398, 1.411]	151.23***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.114 (0.002)	1.121	[1.118, 1.125]	73.19***
Deep Learning	0.037 (0.001)	1.037	[1.035, 1.039]	40.18***
Ensemble	0.007 (0.001)	1.007	[1.005, 1.009]	8.03***
Statistical	0.073 (0.001)	1.076	[1.074, 1.078]	90.13***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.147 (0.001)	0.864	[0.862, 0.866]	-124.59***
Yes	-0.004 (0.001)	0.996	[0.994, 0.997]	-5.40***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.445	[1.442, 1.448]	377.92***
Phase 2	-0.116 (0.001)	0.890	[0.888, 0.892]	-100.11***
Phase 3	0.250 (0.001)	1.285	[1.282, 1.287]	236.51***
Phase 4	0.440 (0.001)	1.552	[1.549, 1.555]	417.30***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.76***
Phase 6	0.141 (0.001)	1.151	[1.148, 1.154]	112.47***
<i>Health Outcomes</i>				
Asthma	-0.045 (0.001)	0.956	[0.954, 0.958]	-34.87***
Obesity	0.004 (0.0001)	1.004	[1.004, 1.005]	31.91***
High Blood Pressure	-0.002 (0.0003)	0.998	[0.997, 0.998]	-7.73***
COPD	-0.000 (0.001)	1.000	[0.998, 1.002]	-0.01
Stroke	-0.019 (0.003)	0.981	[0.974, 0.987]	-5.65***
Cancer	-0.016 (0.003)	0.984	[0.978, 0.989]	-5.74***
Coronary Heart Disease	0.085 (0.003)	1.089	[1.084, 1.094]	33.35***
Diabetes	-0.005 (0.001)	0.995	[0.993, 0.997]	-5.34***
Kidney Disease	0.042 (0.006)	1.043	[1.030, 1.056]	6.56***
Age 65+	0.106 (0.007)	1.112	[1.097, 1.127]	15.48***
Intercept	-4.755 (0.018)	0.009	[0.008, 0.009]	-264.84***
Notes: Urbanicity $\times$ Lookahead interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.15: Main Effects for GLM-2a (Urbanicity $\times$ Lookahead Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.181 (0.005)	1.199	[1.188, 1.210]	39.86***
AL	0.076 (0.004)	1.079	[1.071, 1.086]	21.19***
AR	0.030 (0.004)	1.030	[1.022, 1.038]	7.53***
AZ	0.095 (0.005)	1.099	[1.089, 1.110]	18.88***
CA	0.027 (0.004)	1.027	[1.020, 1.035]	7.34***
CO	0.222 (0.004)	1.249	[1.240, 1.257]	62.85***
DE	0.039 (0.009)	1.040	[1.021, 1.059]	4.27***
GA	0.064 (0.004)	1.067	[1.059, 1.074]	18.10***
HI	-0.204 (0.011)	0.816	[0.799, 0.833]	-19.09***
IA	0.106 (0.004)	1.112	[1.104, 1.120]	29.51***
ID	0.091 (0.003)	1.095	[1.088, 1.102]	26.42***
IL	0.148 (0.003)	1.160	[1.153, 1.167]	47.64***
IN	0.085 (0.003)	1.088	[1.082, 1.095]	28.46***
KS	0.183 (0.003)	1.200	[1.193, 1.208]	60.35***
KY	0.148 (0.003)	1.160	[1.153, 1.166]	50.66***
LA	0.081 (0.004)	1.084	[1.076, 1.092]	20.76***
MA	-0.031 (0.005)	0.969	[0.959, 0.980]	-5.73***
MD	-0.069 (0.005)	0.933	[0.925, 0.942]	-15.16***
ME	-0.080 (0.005)	0.924	[0.914, 0.933]	-15.71***
MI	0.070 (0.003)	1.072	[1.066, 1.079]	23.46***
MN	0.115 (0.003)	1.122	[1.114, 1.129]	33.39***
MO	0.062 (0.003)	1.064	[1.057, 1.071]	18.06***
MS	0.077 (0.004)	1.080	[1.072, 1.089]	18.97***
MT	0.172 (0.003)	1.187	[1.179, 1.195]	50.77***
NC	0.031 (0.003)	1.031	[1.024, 1.038]	8.87***
ND	0.142 (0.004)	1.153	[1.143, 1.163]	33.80***
NE	0.071 (0.004)	1.074	[1.066, 1.082]	17.89***
NH	0.034 (0.006)	1.035	[1.023, 1.047]	6.00***
NJ	0.008 (0.005)	1.008	[1.000, 1.017]	1.87
NM	0.041 (0.004)	1.041	[1.034, 1.049]	10.57***
NV	0.020 (0.005)	1.020	[1.011, 1.029]	4.24***
OH	-0.043 (0.003)	0.958	[0.952, 0.964]	-13.38***
OK	0.117 (0.003)	1.124	[1.117, 1.131]	37.61***
OR	-0.031 (0.004)	0.969	[0.962, 0.976]	-8.51***
PA	-0.023 (0.003)	0.978	[0.971, 0.984]	-7.13***
RI	0.246 (0.007)	1.278	[1.261, 1.296]	34.90***
SC	0.059 (0.004)	1.061	[1.053, 1.070]	14.41***
SD	0.083 (0.004)	1.086	[1.078, 1.095]	19.94***
TN	0.167 (0.003)	1.182	[1.175, 1.189]	53.58***
TX	0.126 (0.003)	1.135	[1.128, 1.142]	38.50***
UT	-0.295 (0.005)	0.745	[0.737, 0.752]	-56.06***
VA	0.091 (0.003)	1.096	[1.089, 1.102]	30.94***
VT	-0.020 (0.005)	0.980	[0.970, 0.990]	-3.83***
WA	0.033 (0.004)	1.033	[1.026, 1.041]	8.97***
WI	0.117 (0.003)	1.125	[1.117, 1.132]	35.25***
WV	0.060 (0.004)	1.062	[1.055, 1.069]	17.10***
WY	0.156 (0.004)	1.169	[1.159, 1.178]	38.47***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.16: State Fixed Effects for GLM-2a (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.113 (0.001)	1.119	[1.118, 1.121]	144.19***
21-day ahead	0.191 (0.001)	1.211	[1.209, 1.213]	250.41***
28-day ahead	0.262 (0.001)	1.300	[1.298, 1.302]	354.09***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.114 (0.002)	1.121	[1.118, 1.125]	73.38***
Deep Learning	0.037 (0.001)	1.037	[1.036, 1.039]	40.48***
Ensemble	0.007 (0.001)	1.007	[1.006, 1.009]	8.39***
Statistical	0.073 (0.001)	1.076	[1.074, 1.078]	90.26***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.147 (0.001)	0.863	[0.861, 0.865]	-125.11***
Yes	-0.004 (0.001)	0.996	[0.994, 0.997]	-5.51***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.379 (0.003)	1.461	[1.453, 1.470]	130.08***
Phase 2	0.014 (0.003)	1.014	[1.008, 1.021]	4.25***
Phase 3	0.245 (0.003)	1.278	[1.270, 1.286]	76.61***
Phase 4	0.522 (0.003)	1.686	[1.676, 1.696]	170.89***
Phase 5	0.133 (0.004)	1.143	[1.134, 1.151]	35.21***
Phase 6	0.144 (0.004)	1.154	[1.146, 1.163]	37.40***
<i>Health Outcomes</i>				
Asthma	-0.045 (0.001)	0.956	[0.953, 0.958]	-35.31***
Obesity	0.004 (0.0001)	1.004	[1.004, 1.004]	31.23***
High Blood Pressure	-0.002 (0.0003)	0.998	[0.997, 0.998]	-7.78***
COPD	0.0002 (0.001)	1.000	[0.998, 1.002]	0.18
Stroke	-0.019 (0.003)	0.982	[0.975, 0.988]	-5.43***
Cancer	-0.016 (0.003)	0.984	[0.979, 0.990]	-5.60***
Coronary Heart Disease	0.084 (0.003)	1.088	[1.083, 1.093]	33.06***
Diabetes	-0.005 (0.001)	0.995	[0.993, 0.997]	-5.10***
Kidney Disease	0.043 (0.006)	1.044	[1.031, 1.057]	6.66***
Age 65+	0.102 (0.007)	1.107	[1.093, 1.122]	14.92***
Intercept	-4.746 (0.018)	0.009	[0.008, 0.009]	-263.71***
Notes: Urbanicity main effects and Urbanicity $\times$ Phase interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.17: Main Effects for GLM-2b (Urbanicity $\times$ Phase Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.176 (0.005)	1.193	[1.182, 1.204]	38.81***
AL	0.073 (0.004)	1.075	[1.068, 1.083]	20.37***
AR	0.028 (0.004)	1.028	[1.020, 1.036]	7.05***
AZ	0.091 (0.005)	1.096	[1.085, 1.107]	18.27***
CA	0.021 (0.004)	1.022	[1.014, 1.029]	5.86***
CO	0.219 (0.004)	1.245	[1.236, 1.254]	62.14***
DE	0.040 (0.009)	1.041	[1.023, 1.060]	4.41***
GA	0.061 (0.004)	1.063	[1.055, 1.070]	17.12***
HI	-0.207 (0.011)	0.813	[0.796, 0.830]	-19.45***
IA	0.104 (0.004)	1.109	[1.101, 1.117]	28.83***
ID	0.088 (0.003)	1.092	[1.085, 1.099]	25.75***
IL	0.146 (0.003)	1.157	[1.150, 1.164]	47.03***
IN	0.082 (0.003)	1.085	[1.079, 1.092]	27.59***
KS	0.181 (0.003)	1.198	[1.191, 1.205]	59.89***
KY	0.146 (0.003)	1.157	[1.150, 1.164]	49.95***
LA	0.078 (0.004)	1.081	[1.072, 1.089]	19.99***
MA	-0.030 (0.005)	0.970	[0.960, 0.981]	-5.55***
MD	-0.069 (0.005)	0.933	[0.925, 0.942]	-15.25***
ME	-0.083 (0.005)	0.920	[0.911, 0.930]	-16.40***
MI	0.068 (0.003)	1.070	[1.064, 1.076]	22.85***
MN	0.113 (0.003)	1.119	[1.112, 1.127]	32.80***
MO	0.059 (0.003)	1.061	[1.054, 1.068]	17.34***
MS	0.074 (0.004)	1.077	[1.069, 1.086]	18.31***
MT	0.169 (0.003)	1.185	[1.177, 1.192]	50.20***
NC	0.028 (0.003)	1.028	[1.021, 1.035]	8.04***
ND	0.140 (0.004)	1.150	[1.141, 1.160]	33.30***
NE	0.068 (0.004)	1.070	[1.062, 1.079]	17.03***
NH	0.033 (0.006)	1.033	[1.022, 1.045]	5.76***
NJ	0.013 (0.004)	1.013	[1.004, 1.022]	2.92**
NM	0.037 (0.004)	1.038	[1.030, 1.046]	9.65***
NV	0.017 (0.005)	1.017	[1.007, 1.026]	3.55***
OH	-0.045 (0.003)	0.956	[0.950, 0.962]	-14.15***
OK	0.115 (0.003)	1.122	[1.115, 1.129]	37.16***
OR	-0.034 (0.004)	0.967	[0.960, 0.974]	-9.08***
PA	-0.024 (0.003)	0.976	[0.970, 0.982]	-7.68***
RI	0.252 (0.007)	1.287	[1.270, 1.305]	36.23***
SC	0.056 (0.004)	1.058	[1.049, 1.066]	13.66***
SD	0.080 (0.004)	1.084	[1.075, 1.093]	19.37***
TN	0.164 (0.003)	1.178	[1.171, 1.185]	52.72***
TX	0.123 (0.003)	1.131	[1.123, 1.138]	37.46***
UT	-0.299 (0.005)	0.742	[0.734, 0.749]	-56.89***
VA	0.089 (0.003)	1.093	[1.087, 1.100]	30.33***
VT	-0.025 (0.005)	0.975	[0.965, 0.985]	-4.75***
WA	0.031 (0.004)	1.031	[1.024, 1.039]	8.45***
WI	0.115 (0.003)	1.122	[1.115, 1.129]	34.65***
WV	0.058 (0.003)	1.060	[1.053, 1.067]	16.67***
WY	0.153 (0.004)	1.166	[1.156, 1.175]	37.90***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.18: State Fixed Effects for GLM-2b (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.113 (0.001)	1.119	[1.117, 1.121]	143.69***
21-day ahead	0.191 (0.001)	1.211	[1.209, 1.213]	249.56***
28-day ahead	0.262 (0.001)	1.299	[1.298, 1.301]	352.89***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.125 (0.004)	1.133	[1.125, 1.141]	34.04***
Deep Learning	0.055 (0.003)	1.056	[1.051, 1.062]	20.41***
Ensemble	0.017 (0.002)	1.017	[1.013, 1.021]	8.45***
Statistical	0.050 (0.002)	1.051	[1.046, 1.056]	22.23***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.147 (0.001)	0.863	[0.861, 0.865]	-124.79***
Yes	-0.005 (0.001)	0.995	[0.994, 0.997]	-6.21***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.445	[1.442, 1.448]	377.50***
Phase 2	-0.117 (0.001)	0.890	[0.888, 0.892]	-100.13***
Phase 3	0.250 (0.001)	1.284	[1.282, 1.287]	236.13***
Phase 4	0.439 (0.001)	1.552	[1.549, 1.555]	416.75***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.75***
Phase 6	0.141 (0.001)	1.151	[1.148, 1.154]	112.27***
<i>Health Outcomes</i>				
Asthma	-0.045 (0.001)	0.956	[0.954, 0.958]	-34.84***
Obesity	0.004 (0.0001)	1.004	[1.004, 1.005]	31.87***
High Blood Pressure	-0.002 (0.0003)	0.998	[0.997, 0.998]	-7.68***
COPD	-0.0001 (0.001)	1.000	[0.998, 1.002]	-0.06
Stroke	-0.019 (0.003)	0.981	[0.974, 0.988]	-5.60***
Cancer	-0.016 (0.003)	0.984	[0.978, 0.989]	-5.69***
Coronary Heart Disease	0.085 (0.003)	1.089	[1.084, 1.095]	33.34***
Diabetes	-0.005 (0.001)	0.995	[0.993, 0.997]	-5.41***
Kidney Disease	0.042 (0.006)	1.043	[1.030, 1.056]	6.59***
Age 65+	0.105 (0.007)	1.110	[1.096, 1.125]	15.29***
Intercept	-4.710 (0.018)	0.009	[0.009, 0.009]	-262.75***
Notes: Urbanicity main effects and Urbanicity $\times$ Model Type interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.19: Main Effects for GLM-2c (Urbanicity $\times$ Model Type Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.181 (0.005)	1.198	[1.188, 1.209]	39.68***
AL	0.076 (0.004)	1.079	[1.071, 1.086]	21.12***
AR	0.030 (0.004)	1.030	[1.022, 1.038]	7.50***
AZ	0.095 (0.005)	1.099	[1.088, 1.110]	18.85***
CA	0.026 (0.004)	1.026	[1.019, 1.034]	7.08***
CO	0.222 (0.004)	1.249	[1.240, 1.257]	62.76***
DE	0.039 (0.009)	1.040	[1.021, 1.059]	4.26***
GA	0.064 (0.004)	1.066	[1.059, 1.074]	18.04***
HI	-0.204 (0.011)	0.816	[0.799, 0.833]	-19.08***
IA	0.106 (0.004)	1.112	[1.104, 1.120]	29.46***
ID	0.090 (0.003)	1.095	[1.087, 1.102]	26.34***
IL	0.148 (0.003)	1.160	[1.153, 1.167]	47.60***
IN	0.085 (0.003)	1.088	[1.082, 1.095]	28.43***
KS	0.183 (0.003)	1.200	[1.193, 1.207]	60.22***
KY	0.148 (0.003)	1.160	[1.153, 1.166]	50.58***
LA	0.081 (0.004)	1.084	[1.076, 1.092]	20.70***
MA	-0.032 (0.005)	0.969	[0.959, 0.979]	-5.78***
MD	-0.069 (0.005)	0.933	[0.925, 0.941]	-15.21***
ME	-0.080 (0.005)	0.923	[0.914, 0.933]	-15.73***
MI	0.070 (0.003)	1.072	[1.066, 1.079]	23.39***
MN	0.115 (0.003)	1.122	[1.114, 1.129]	33.30***
MO	0.062 (0.003)	1.064	[1.057, 1.071]	18.01***
MS	0.077 (0.004)	1.080	[1.072, 1.089]	18.90***
MT	0.171 (0.003)	1.187	[1.179, 1.195]	50.62***
NC	0.031 (0.003)	1.031	[1.024, 1.038]	8.84***
ND	0.142 (0.004)	1.153	[1.143, 1.162]	33.67***
NE	0.071 (0.004)	1.074	[1.065, 1.082]	17.79***
NH	0.034 (0.006)	1.035	[1.023, 1.047]	5.99***
NJ	0.009 (0.005)	1.009	[1.000, 1.018]	1.91
NM	0.040 (0.004)	1.041	[1.033, 1.049]	10.50***
NV	0.020 (0.005)	1.020	[1.010, 1.029]	4.19***
OH	-0.043 (0.003)	0.958	[0.952, 0.964]	-13.38***
OK	0.117 (0.003)	1.124	[1.117, 1.131]	37.54***
OR	-0.032 (0.004)	0.969	[0.962, 0.976]	-8.55***
PA	-0.023 (0.003)	0.977	[0.971, 0.984]	-7.16***
RI	0.246 (0.007)	1.279	[1.261, 1.296]	34.90***
SC	0.059 (0.004)	1.061	[1.053, 1.070]	14.38***
SD	0.083 (0.004)	1.086	[1.077, 1.095]	19.86***
TN	0.167 (0.003)	1.182	[1.175, 1.189]	53.55***
TX	0.126 (0.003)	1.135	[1.127, 1.142]	38.44***
UT	-0.295 (0.005)	0.745	[0.737, 0.752]	-56.01***
VA	0.091 (0.003)	1.096	[1.089, 1.102]	30.91***
VT	-0.020 (0.005)	0.980	[0.970, 0.990]	-3.86***
WA	0.033 (0.004)	1.033	[1.026, 1.041]	8.97***
WI	0.117 (0.003)	1.124	[1.117, 1.132]	35.16***
WV	0.060 (0.004)	1.062	[1.055, 1.069]	17.09***
WY	0.156 (0.004)	1.169	[1.159, 1.178]	38.38***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.20: State Fixed Effects for GLM-2c (Reference State: NY)

Variable	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
<i>Lookahead (ref: 7-day)</i>				
14-day ahead	0.113 (0.001)	1.119	[1.117, 1.121]	143.72***
21-day ahead	0.191 (0.001)	1.211	[1.209, 1.213]	249.59***
28-day ahead	0.262 (0.001)	1.300	[1.298, 1.301]	352.91***
<i>Model Type (ref: Compartmental)</i>				
Baseline Models	0.114 (0.002)	1.121	[1.118, 1.124]	73.05***
Deep Learning	0.037 (0.001)	1.037	[1.035, 1.039]	40.28***
Ensemble	0.007 (0.001)	1.007	[1.005, 1.009]	7.94***
Statistical	0.073 (0.001)	1.076	[1.075, 1.078]	90.18***
<i>Mobility Used (ref: No)</i>				
Not Applicable	-0.127 (0.003)	0.881	[0.876, 0.885]	-49.05***
Yes	0.023 (0.002)	1.023	[1.019, 1.027]	11.60***
<i>Phase (ref: Phase 0)</i>				
Phase 1	0.368 (0.001)	1.445	[1.442, 1.448]	377.52***
Phase 2	-0.117 (0.001)	0.890	[0.888, 0.892]	-100.12***
Phase 3	0.250 (0.001)	1.284	[1.282, 1.287]	236.12***
Phase 4	0.440 (0.001)	1.552	[1.549, 1.555]	416.76***
Phase 5	-0.031 (0.001)	0.969	[0.967, 0.972]	-22.80***
Phase 6	0.141 (0.001)	1.151	[1.148, 1.154]	112.26***
<i>Health Outcomes</i>				
Asthma	-0.045 (0.001)	0.956	[0.954, 0.958]	-34.87***
Obesity	0.004 (0.0001)	1.004	[1.004, 1.005]	31.88***
High Blood Pressure	-0.002 (0.0003)	0.998	[0.997, 0.998]	-7.66***
COPD	-0.0001 (0.001)	1.000	[0.998, 1.002]	-0.06
Stroke	-0.019 (0.003)	0.981	[0.975, 0.988]	-5.56***
Cancer	-0.016 (0.003)	0.984	[0.978, 0.989]	-5.72***
Coronary Heart Disease	0.085 (0.003)	1.089	[1.084, 1.095]	33.37***
Diabetes	-0.005 (0.001)	0.995	[0.993, 0.997]	-5.45***
Kidney Disease	0.042 (0.006)	1.043	[1.030, 1.056]	6.58***
Age 65+	0.105 (0.007)	1.111	[1.096, 1.126]	15.31***
Intercept	-4.729 (0.018)	0.009	[0.009, 0.009]	-263.34***
Notes: Urbanicity main effects and Urbanicity $\times$ Mobility interactions shown in main manuscript table. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.				

Table B.21: Main Effects for GLM-2d (Urbanicity $\times$ Mobility Interactions)

State	Coef. (Std. Error)	exp(Coef.)	95% CI	z-value
AK	0.181 (0.005)	1.198	[1.187, 1.209]	39.65***
AL	0.076 (0.004)	1.078	[1.071, 1.086]	21.11***
AR	0.029 (0.004)	1.030	[1.022, 1.038]	7.46***
AZ	0.094 (0.005)	1.099	[1.088, 1.110]	18.82***
CA	0.026 (0.004)	1.026	[1.019, 1.034]	7.06***
CO	0.222 (0.004)	1.249	[1.240, 1.257]	62.72***
DE	0.039 (0.009)	1.040	[1.021, 1.059]	4.23***
GA	0.064 (0.004)	1.066	[1.059, 1.074]	18.03***
HI	-0.204 (0.011)	0.815	[0.798, 0.833]	-19.10***
IA	0.106 (0.004)	1.112	[1.104, 1.120]	29.44***
ID	0.091 (0.003)	1.095	[1.087, 1.102]	26.35***
IL	0.148 (0.003)	1.160	[1.153, 1.167]	47.58***
IN	0.085 (0.003)	1.088	[1.082, 1.095]	28.42***
KS	0.183 (0.003)	1.200	[1.193, 1.207]	60.21***
KY	0.148 (0.003)	1.160	[1.153, 1.166]	50.58***
LA	0.081 (0.004)	1.084	[1.076, 1.092]	20.69***
MA	-0.032 (0.005)	0.969	[0.959, 0.979]	-5.77***
MD	-0.069 (0.005)	0.933	[0.925, 0.941]	-15.16***
ME	-0.080 (0.005)	0.923	[0.914, 0.932]	-15.74***
MI	0.070 (0.003)	1.072	[1.066, 1.078]	23.37***
MN	0.115 (0.003)	1.122	[1.114, 1.129]	33.29***
MO	0.062 (0.003)	1.064	[1.057, 1.071]	18.00***
MS	0.077 (0.004)	1.080	[1.071, 1.089]	18.88***
MT	0.171 (0.003)	1.187	[1.179, 1.195]	50.62***
NC	0.031 (0.003)	1.031	[1.024, 1.038]	8.81***
ND	0.142 (0.004)	1.153	[1.143, 1.162]	33.66***
NE	0.071 (0.004)	1.074	[1.065, 1.082]	17.75***
NH	0.034 (0.006)	1.035	[1.023, 1.047]	6.01***
NJ	0.009 (0.005)	1.009	[1.000, 1.018]	1.92
NM	0.040 (0.004)	1.041	[1.033, 1.049]	10.50***
NV	0.020 (0.005)	1.020	[1.010, 1.029]	4.17***
OH	-0.043 (0.003)	0.958	[0.952, 0.964]	-13.40***
OK	0.117 (0.003)	1.124	[1.117, 1.131]	37.53***
OR	-0.032 (0.004)	0.969	[0.962, 0.976]	-8.55***
PA	-0.023 (0.003)	0.977	[0.971, 0.983]	-7.20***
RI	0.246 (0.007)	1.278	[1.261, 1.296]	34.87***
SC	0.059 (0.004)	1.061	[1.052, 1.070]	14.36***
SD	0.083 (0.004)	1.086	[1.077, 1.095]	19.83***
TN	0.167 (0.003)	1.182	[1.175, 1.189]	53.52***
TX	0.126 (0.003)	1.135	[1.127, 1.142]	38.42***
UT	-0.295 (0.005)	0.744	[0.737, 0.752]	-56.02***
VA	0.091 (0.003)	1.096	[1.089, 1.102]	30.92***
VT	-0.020 (0.005)	0.980	[0.970, 0.990]	-3.87***
WA	0.033 (0.004)	1.033	[1.026, 1.041]	8.97***
WI	0.117 (0.003)	1.124	[1.117, 1.132]	35.15***
WV	0.060 (0.004)	1.062	[1.055, 1.069]	17.09***
WY	0.156 (0.004)	1.169	[1.159, 1.178]	38.38***

Notes: New York (NY) is the reference state. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Model uses Gaussian GLM with log link function. Dependent variable is square root of PBL.

Table B.22: State Fixed Effects for GLM-2d (Reference State: NY)

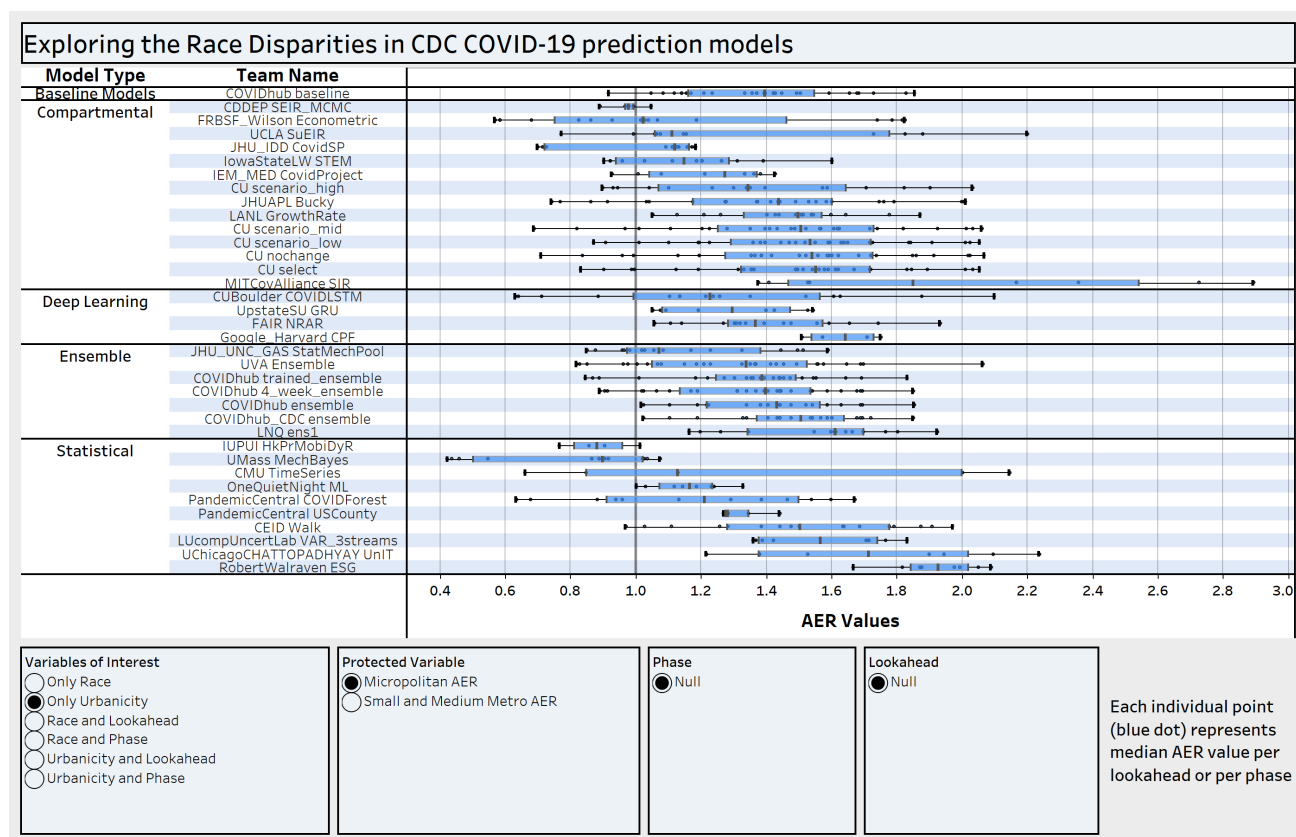


Figure B.7: Forecast Hub Fairness Dashboard showing AER distribution for urbanization levels. Models are organized by type, with teams sorted within each type by median AER values. The AER values compare prediction errors between Micropolitan areas and Large Metropolitan areas (reference group), where values above 1.0 indicate higher prediction errors for Micropolitan regions.

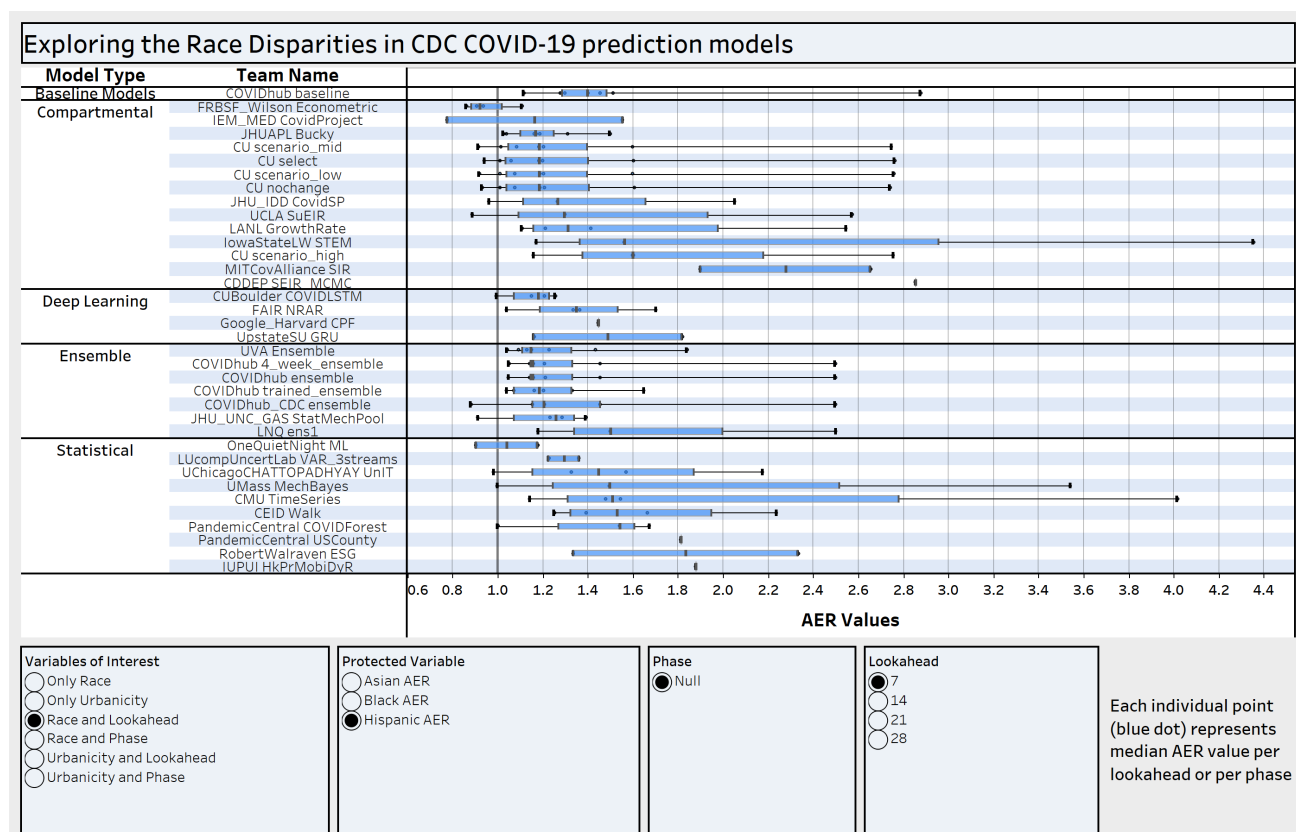


Figure B.8: Forecast Hub Fairness Dashboard displaying the interaction between race and forecast horizon (lookahead). The visualization shows how Hispanic-White AER values vary across different forecast periods (7, 14, 21, and 28 days), allowing assessment of how prediction fairness changes with forecast horizon length(lookaheads).

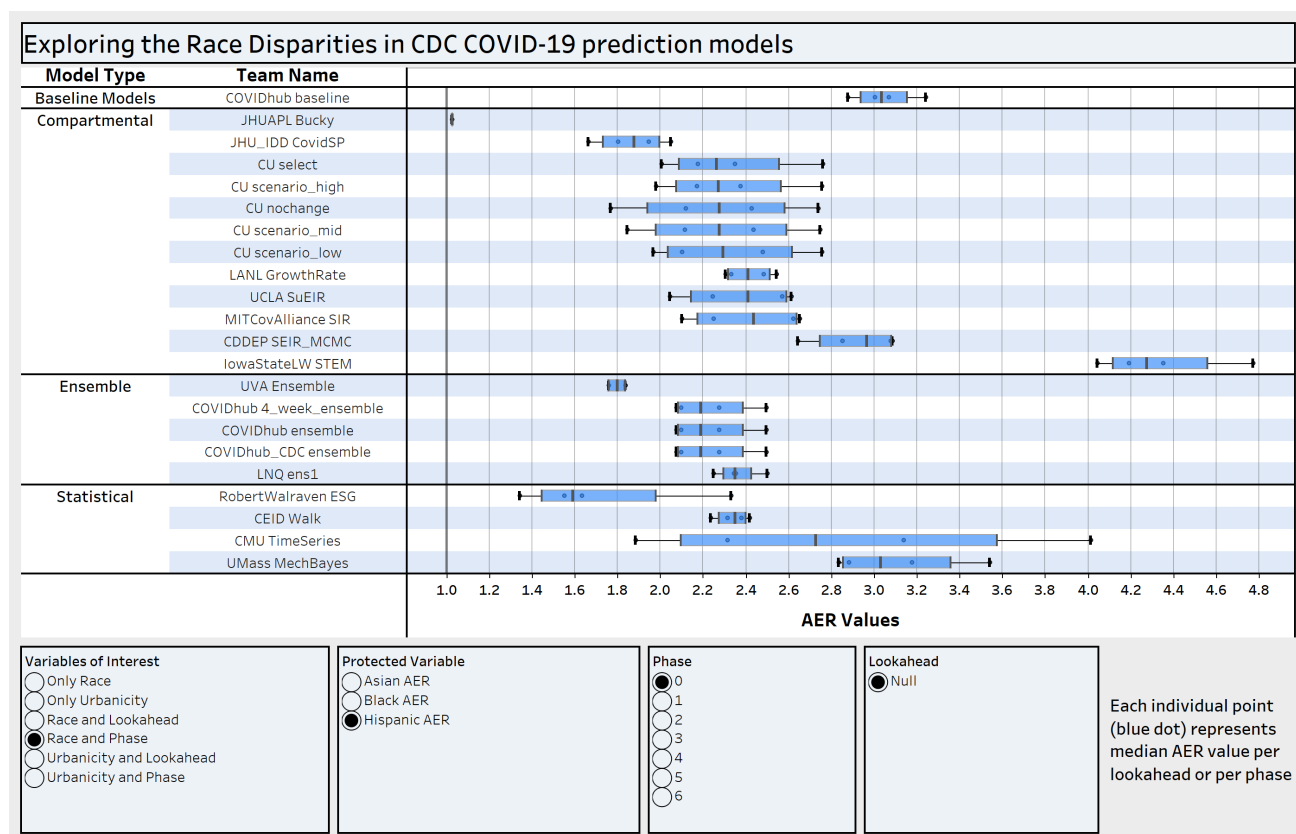


Figure B.9: Forecast Hub Fairness Dashboard illustrating the relationship between race and pandemic phases. The visualization reveals how Hispanic-White AER values evolved across different pandemic phases (0-6), enabling analysis of how prediction fairness varied during different stages of the pandemic.

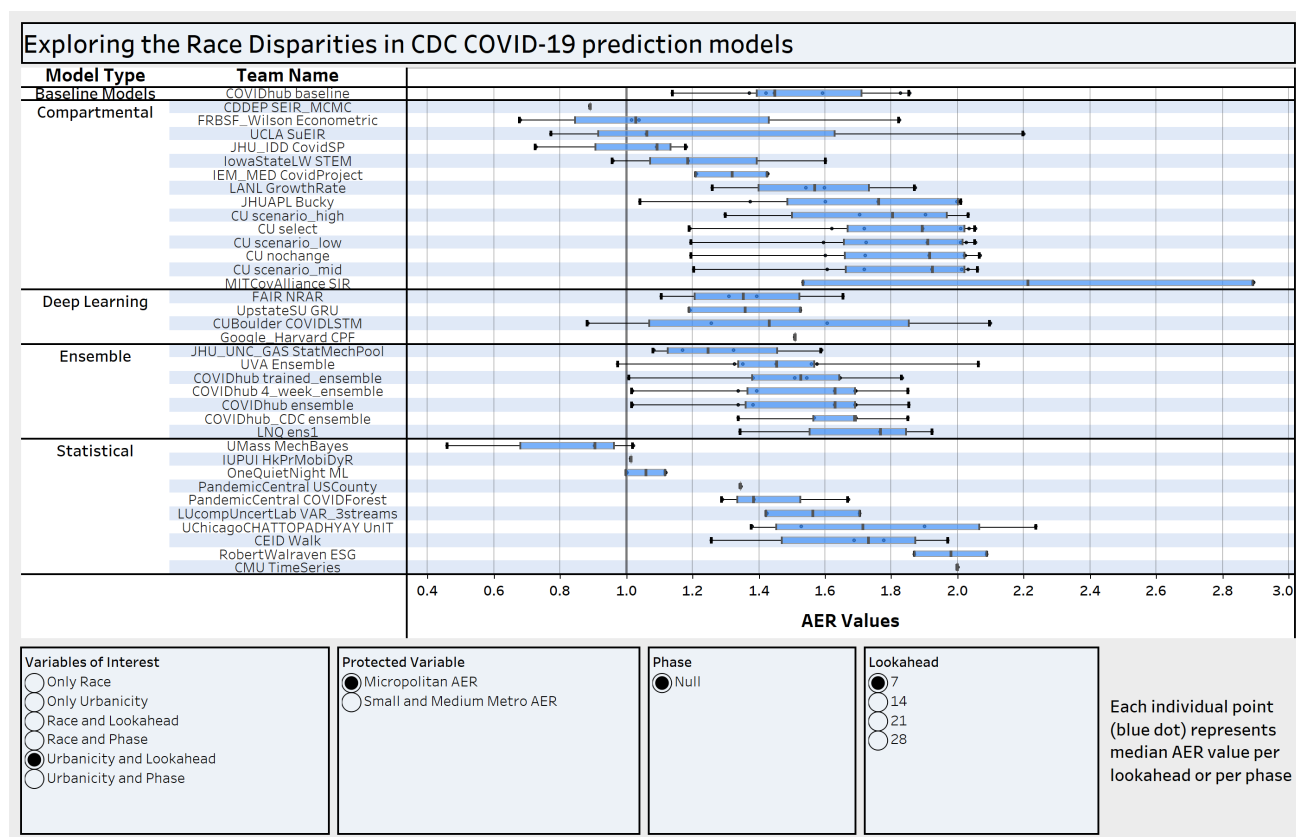


Figure B.10: Forecast Hub Fairness Dashboard showing the interaction between urbanization levels and forecast horizon. The visualization demonstrates how Micropolitan-Large Metropolitan AER values change across different forecast periods (7, 14, 21, and 28 days), highlighting the relationship between prediction fairness and forecast horizon lengths (lookaheads).

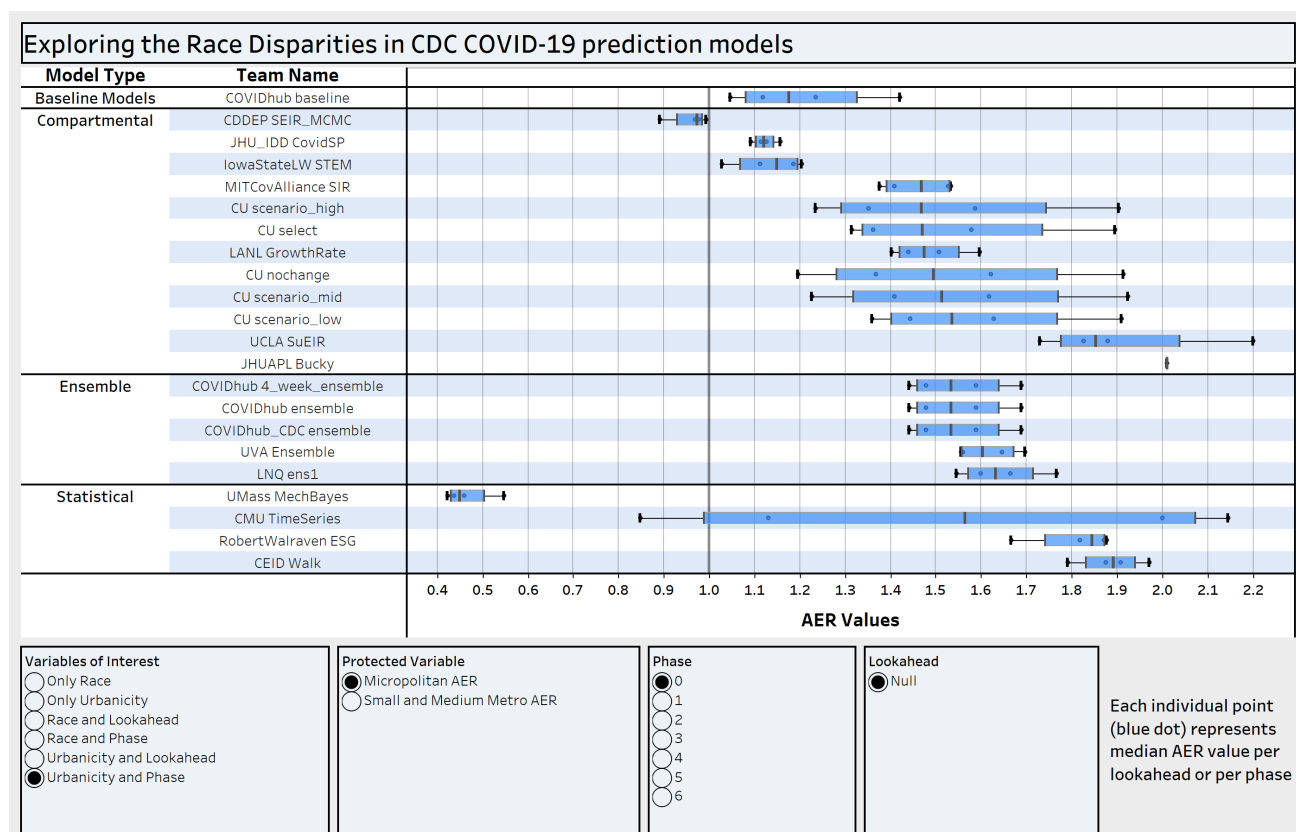


Figure B.11: Forecast Hub Fairness Dashboard depicting the relationship between urbanization levels and pandemic phases. The visualization shows how Micropolitan-Large Metropolitan AER values varied across different pandemic phases (0-6), revealing temporal patterns in prediction fairness across the urban-rural spectrum.

Appendix C: Do Footfalls Reliably Mirror Economic Activity?

C.1 Visit Attribution using Cuebiq and Safegraph Places

Our visit attribution methodology in Algorithm 1 transforms raw mobile device location data from Cuebiq into POI-level visit counts. The process integrates two key datasets: Cuebiq mobility data, providing geocoded device stops (s_j) with associated dwell times, and SafeGraph Places data, which supplies precise POI polygons ($\mathcal{P}$), centroids (c), and business category classifications for establishments across our six target metropolitan areas ($\mathcal{G}$). The attribution process operates in two main stages:

- 1. Data Preprocessing and Spatial Join:** First, we filter both datasets. As a foundational step to protect user privacy in line with Cuebiq’s SPOI policy¹, our POI database is pre-filtered to exclude privacy-sensitive places. These include health-related places, places where vulnerable populations are located, military-related facilities, religious facilities, places related to sexual-orientation, and adult-oriented places. We retain only POIs that match our target business categories and fall within the target MSA boundaries ($\mathcal{P}_{filtered}$). We then filter the device stops ($\mathcal{S}_{filtered}$) to include only those within the MSAs and meeting a minimum dwell time between 5 minutes to 4 hours (τ_{min}, τ_{max}), a threshold set to capture legitimate visits while excluding transient traffic. For each filtered stop, we perform a spatial join to identify the set of all POI polygons that spatially contain the stop’s location ($\mathcal{C}_j$).

- 2. Attribution and Tie-Breaking:** Attribution is determined based on the resulting candidate set

¹<https://spectus.ai/spoi-policy/>

($\mathcal{C}_j$). If a stop falls within only one POI polygon ($|\mathcal{C}_j| = 1$), attribution is straightforward. However, when multiple POIs contain the same stop ($|\mathcal{C}_j| > 1$) — a common scenario in multi-tenant buildings or malls — we employ a distance-based tie-breaking rule. We calculate the Euclidean distance from the stop’s coordinates to the centroid (c_i) of each candidate POI. The visit is then attributed to the single POI with the closest centroid (p_j^*).

Finally, these successfully attributed visits are used to create individual visit records ($\mathcal{V}$) before being aggregated into a POI-week panel. We aggregate by placekey (for POI p_i) and a custom week (w_t), counting the unique devices ($visits_{i,t}$) for each POI-week.

This aggregation uses a non-standard weekly boundary running from Wednesday through Tuesday. This custom period is specifically chosen to address documented reporting delays in SafeGraph Spend data, ensuring temporal alignment between our visit and spending measures. The resulting dataset, comprising 12.5 million POI-week observations from 52,051 unique establishments, is used in the models for analysis.

C.2 Representation check of SafeGraph Spend

SafeGraph Spend is constructed from a permissioned cardholder panel rather than a census of all transactions. To assess geographic representativeness, we compare each state’s share of panel unique customers to its share of U.S. population using yearly Census population estimates and define monthly state bias as the percentage-point difference between observed and expected shares. We report two national diagnostics:

- (i) the monthly state-level Pearson correlation between population and panel customers shown in Figure C.1, and

Algorithm 1: Visit Attribution Algorithm at POI-Level establishments

Input: POI polygons $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ with centroids $\{c_1, c_2, \dots, c_N\}$

Input: Mobile device stops $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$ with coordinates (lat_j, lng_j)

Input: Minimum dwell time threshold τ_{min} (minutes)

Input: Geographic boundaries for target MSAs $\mathcal{G}$

Output: Visit attribution dataset $\mathcal{V}$ for analysis

Data Preprocessing;

$\mathcal{P}_{filtered} \leftarrow$ Filter POIs by business categories and MSA boundaries $\mathcal{G}$;

$\mathcal{S}_{filtered} \leftarrow$ Filter stops where $\tau_{min} \leq dwell_time \leq \tau_{max}$ and location $\in \mathcal{G}$;

Spatial Containment;

foreach stop $s_j \in \mathcal{S}_{filtered}$ **do**

$\mathcal{C}_j \leftarrow \emptyset$ // Container POIs for stop s_j

foreach POI $p_i \in \mathcal{P}_{filtered}$ **do**

if s_j is spatially contained within polygon p_i **then**

$\mathcal{C}_j \leftarrow \mathcal{C}_j \cup \{p_i\}$;

Distance-Based Tie-Breaking;

foreach stop s_j with $|\mathcal{C}_j| > 1$ **do**

$distances_j \leftarrow \{\}$;

foreach $p_i \in \mathcal{C}_j$ **do**

$d_{ij} \leftarrow$ Euclidean distance from s_j to centroid c_i ;

$distances_j[p_i] \leftarrow d_{ij}$;

$p_j^* \leftarrow \arg \min_{p_i \in \mathcal{C}_j} distances_j[p_i]$ // Closest POI

Visit Record Creation;

foreach attributed stop s_j to POI p_j^* **do**

 Create visit record v_j ;

$v_j.device_id \leftarrow s_j.device_id$;

$v_j.placekey \leftarrow p_j^*.placekey$;

$v_j.visit_date \leftarrow s_j.timestamp$;

$v_j.business_category \leftarrow p_j^*.category$;

$v_j.cbg_id \leftarrow s_j.cbg_id$;

$v_j.distance_to_centroid \leftarrow d_{ij}$;

$v_j.is_shared_polygon \leftarrow |\mathcal{C}_j| > 1$;

$\mathcal{V} \leftarrow \mathcal{V} \cup \{v_j\}$;

Temporal Aggregation;

Aggregate $\mathcal{V}$ by $(placekey, week)$ to create POI-week panel;

foreach (p_i, w_t) combination **do**

$visits_{i,t} \leftarrow$ Count unique devices visiting POI p_i in week w_t ;

return Panel dataset ready for elasticity estimation:

$\{(i, t, visits_{i,t}, spend_{i,t}, demographics_i)\}$

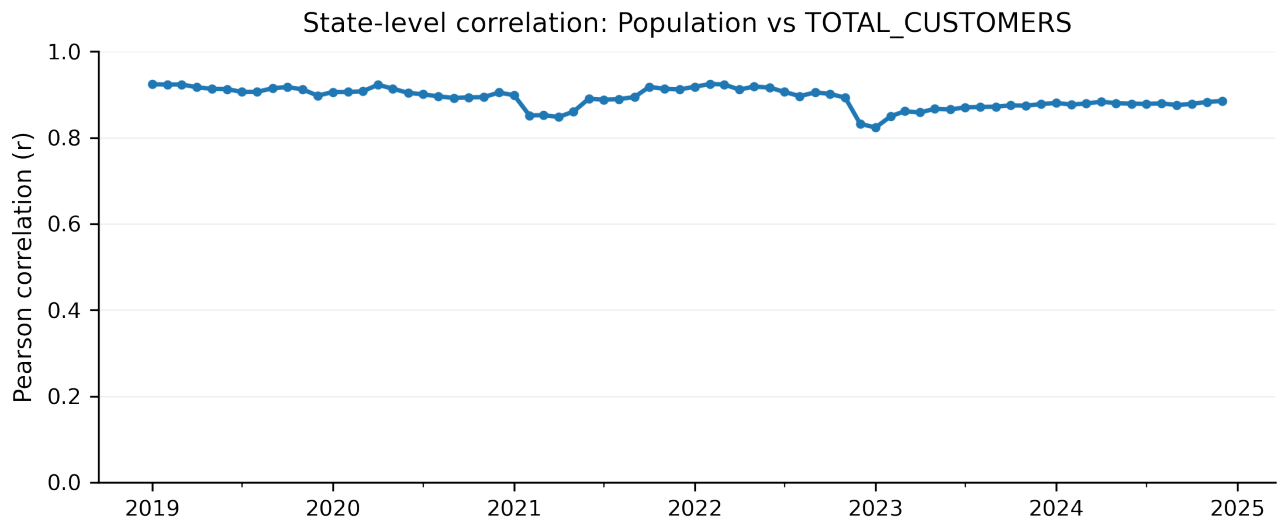


Figure C.1: **State-level alignment between population and SafeGraph Spend panel customers (monthly).** This figure plots the monthly Pearson correlation coefficient computed across U.S. states between (i) each state’s Census Bureau annual population estimate (assigned to all months within that year) and (ii) SafeGraph Spend’s TOTAL_CUSTOMERS (the number of unique panel customers assigned to each state in that month). Higher values indicate that the geographic density of the Spend panel more closely mirrors population density at the state level. Correlations remain consistently high throughout the study period (minimum $r \geq 0.82$), suggesting that the panel’s customer distribution tracks population density closely overall.

(ii) Figure C.2 shows the distribution of absolute state bias over time, (median, interquartile range, and 95th percentile).

These diagnostics indicate that the Spend customer panel tracks population density closely overall, while still exhibiting slight representation biases (i.e., modest over-/under-indexing) in some states and months. This is consistent with SafeGraph’s own documentation, which reports that panel density closely mirrors population density with an overall average percentage-point difference below 1 pp and a maximum of approximately ± 4 pp per state in their analysis, and provides a detailed discussion in their “Quantifying Sampling Bias in SafeGraph Spend” resource.² In addition, SafeGraph reports external validity checks comparing aggregated Spend metrics (rolled up to parent brands) to

²<https://colab.research.google.com/drive/16eBT1dqnUK76GnBMgPMY5rJO4vBD2gdQ?usp=sharing>

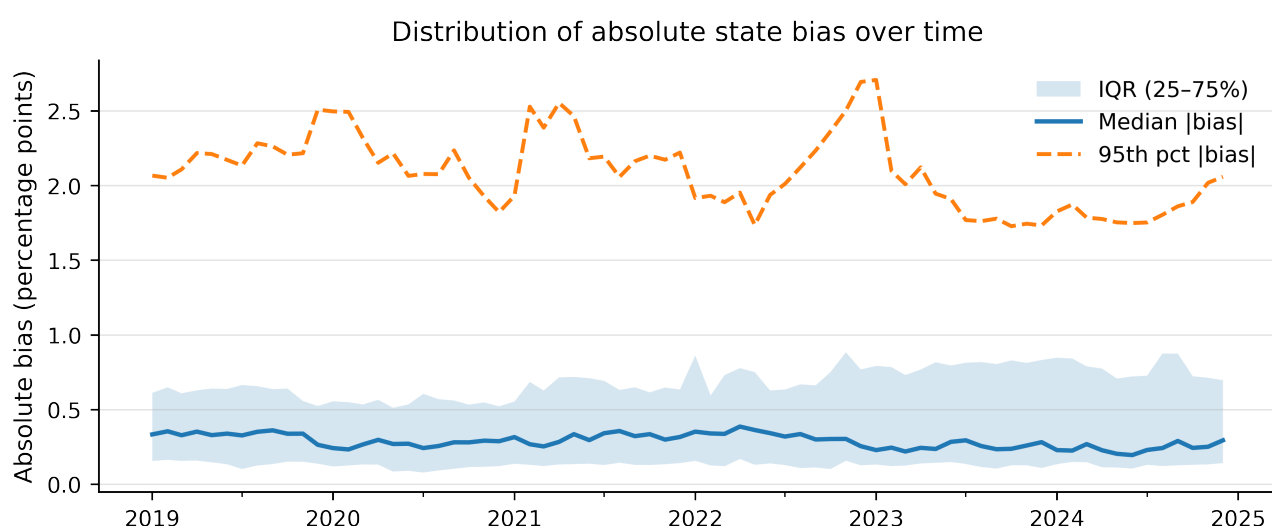


Figure C.2: **Distribution of geographic sampling bias in the SafeGraph Spend panel over time.** This plot summarizes the monthly distribution of absolute state bias in SafeGraph Spend panel coverage, where state bias is defined as the percentage-point difference between a state’s share of SafeGraph Spend panel unique customers (observed share) and its share of the U.S. population from ACS annual population estimates (expected share), and we plot $|bias|$. For each month, the solid line shows the median $|bias|$ across states, the shaded band shows the interquartile range (25th-75th percentiles), and the dashed line shows the 95th percentile, capturing tail deviations. The median and IQR remain low and stable over time, indicating modest typical state-level deviations, while the 95th percentile highlights periods when a small number of states exhibit larger deviations from proportionality.

publicly reported quarterly revenues for major brands (e.g., McDonald’s, Chipotle, Target), finding strong positive correlations (reported in the ~ 0.7 – 0.9 range) ³.

³<https://www.safegraph.com/blog/validating-spend-data-for-brands-against-company-reporting>

	POI Category	MSA	n_pois	n_obs	n_cbgs	Income (mean, sd)	Education (mean, sd)	% White (mean, sd)	% Black (mean, sd)	% Hispanic (mean, sd)	% Asian (mean, sd)
Essential Goods	Grocery/Convenience	CHI	1164	318053	980	49.98 (28.41)	25.65 (13.33)	57.57 (24.99)	9.49 (17.45)	20.87 (20.96)	8.51 (10.26)
		HOU	577	141141	527	45.6 (29.24)	20.54 (14.51)	35.68 (23.98)	14.84 (15.6)	37.47 (24)	8.33 (11.41)
		ATL	513	139078	449	51.84 (29.91)	26.15 (14.97)	46.68 (25.37)	28.88 (24.48)	12.31 (11.67)	7.35 (9.95)
		MIA	715	204445	607	49.06 (29.05)	21.19 (11.17)	32.82 (26.11)	16.02 (20.46)	44.33 (28.16)	2.87 (2.57)
		WAS	911	258302	734	50.6 (28.79)	25.05 (11.03)	41.43 (24.52)	22.6 (22.68)	19.9 (16.34)	10.43 (10.09)
		PHI	749	209062	658	47.79 (28.87)	23.05 (12.33)	64.23 (23.21)	14.25 (17.78)	9.77 (10.58)	7.13 (8.31)
Routine Services	Personal Care Services	CHI	1287	260247	871	55.12 (28.8)	28.03 (12.9)	62.39 (22.37)	9.29 (17.81)	15.59 (15)	9.07 (9.91)
		HOU	884	169638	561	56.23 (28.33)	26.18 (14.3)	42.39 (20.92)	13.82 (12.65)	28.68 (17.11)	10.92 (12.53)
		ATL	856	174581	522	54.41 (28.41)	28.25 (14.1)	45.97 (23.48)	28.17 (23.55)	12.21 (10.63)	8.69 (10.53)
		MIA	743	147434	493	55.2 (27.87)	22.72 (10.61)	37.4 (24.9)	15.99 (19.86)	38.54 (24.9)	3.5 (2.59)
		WAS	738	149253	425	53.64 (28.13)	27.61 (11.14)	47.48 (22.29)	18.34 (18.66)	16.48 (13.14)	11.81 (9.62)
		PHI	763	153565	518	55.09 (27.41)	26.5 (11.3)	70.28 (17.91)	9.89 (12.16)	7.71 (7.2)	7.75 (8.26)
	Drycleaning/Laundry	CHI	118	23896	116	52.17 (31.6)	30.66 (15.25)	55.58 (25.25)	12.52 (20.68)	18.22 (18.87)	9.88 (9.6)
		HOU	85	15831	84	54.49 (29.16)	26.96 (14.1)	40.91 (21.79)	15.98 (14.02)	27.97 (17.1)	10.73 (10.27)
		ATL	35	6847	34	59.7 (28.39)	35.68 (13.15)	53.63 (20.47)	19.82 (15.17)	12.09 (10.34)	9.11 (10.86)
		MIA	56	11425	55	56.75 (26.55)	24.46 (11.84)	41.38 (26.6)	11.75 (16.93)	39.73 (28.08)	3.25 (2.3)
		WAS	74	15116	70	58.9 (28.9)	30.31 (12.09)	48.08 (23.76)	15.58 (17.74)	16.84 (13.89)	14.04 (10.93)
		PHI	28	4978	28	38.96 (29.4)	21.39 (13.08)	60.32 (29.64)	15.55 (21.97)	10 (15.15)	10.29 (14.09)
Social & Experiential	Restaurants	CHI	7752	1940009	2602	48.7 (29.05)	26.02 (13.51)	55.65 (24.94)	11.07 (19.21)	20.74 (20.48)	9 (10.94)
		HOU	5227	1328043	1573	48.94 (28.63)	23.24 (14.77)	37.61 (22.09)	15.71 (14.97)	33.41 (20.52)	9.31 (11.59)
		ATL	4924	1266019	1314	48.75 (28.98)	26.45 (14.37)	43.13 (24.57)	30.89 (24.64)	13.3 (12.3)	7.87 (9.86)
		MIA	4226	1054881	1362	49.49 (28.8)	21.96 (11.18)	34.26 (25.31)	15.64 (19.8)	42.94 (27.14)	3.04 (2.62)
		WAS	4148	1012813	1148	50.21 (29.15)	26.74 (11.81)	43.88 (23.66)	21.45 (22.15)	17.24 (13.14)	11.78 (10.51)
		PHI	3589	866413	1437	49.38 (29.05)	24.99 (12.44)	64.89 (22.37)	13.33 (17.07)	8.83 (8.6)	8.45 (10.13)
	Drinking Places	CHI	282	58832	235	55.31 (28.96)	32.42 (16.28)	61.65 (20.67)	6.83 (11.57)	19.4 (18.16)	8.1 (8.98)
		HOU	181	39381	130	55.07 (27.34)	29.16 (15.3)	47.98 (16.82)	12.8 (12.18)	26.44 (13.59)	8.18 (7.27)
		ATL	118	28454	91	55.8 (29.66)	32.55 (14.21)	53.25 (24.6)	23.2 (19.95)	11.29 (9.32)	6.76 (8.65)
		MIA	113	23869	87	51.66 (28.73)	24.74 (13.18)	44.36 (25.28)	9.96 (11.86)	38.2 (23.87)	2.86 (1.89)
		WAS	117	24196	91	61.92 (30.14)	30.79 (12.01)	54.57 (22.26)	16.01 (18.84)	13.29 (13.09)	10.77 (9.86)
		PHI	109	22752	98	50.15 (30.19)	28.94 (13.46)	67.1 (19.31)	10.61 (13.31)	9.97 (10.54)	7.66 (7.96)
	Specialty Food Stores	CHI	124	26551	118	52.27 (28.62)	28.26 (14.26)	56.54 (24.72)	8.15 (15.91)	20.86 (20.86)	10.84 (12.81)
		HOU	106	23082	101	40.32 (27.35)	18.91 (13.39)	34.3 (22.83)	15.15 (14.37)	38.69 (23.08)	8.47 (12.93)
		ATL	97	20788	90	43.51 (30.36)	24.75 (14.68)	41.18 (24.08)	28.56 (22.38)	18.12 (19.3)	7.49 (9.93)
		MIA	74	15609	68	45.63 (31.42)	22.6 (13.16)	34.42 (29.26)	17.15 (24.83)	42.3 (28.59)	2.31 (1.99)
		WAS	63	13538	57	56.27 (31.13)	27.65 (11.6)	52.05 (23.75)	15.66 (17.45)	15.97 (13.93)	10.36 (11.63)
		PHI	105	24049	95	50.17 (27.82)	24.02 (11.15)	67.64 (23.69)	13.42 (19.55)	8.48 (8.5)	5.97 (7.4)
Discretionary Retail	Clothing	CHI	766	180389	225	50.53 (28.24)	30.06 (12.83)	56.71 (21.53)	10.44 (16.64)	16.89 (15.97)	12.37 (12.2)
		HOU	554	131990	148	54.23 (29.41)	26.88 (13.14)	45.42 (20.91)	12.66 (11.41)	25.57 (17.09)	12.11 (12.79)
		ATL	480	116925	104	51.19 (26.66)	29.69 (14.24)	44.63 (21.71)	27.45 (21.91)	11.65 (8.32)	11.67 (13.63)
		MIA	536	127787	165	50.26 (29.17)	23.27 (11.38)	33.71 (24.36)	17.36 (22.61)	40.75 (24.91)	3.42 (2.67)
		WAS	467	109300	136	44.07 (26.28)	28.77 (12.19)	40.94 (21.64)	23.08 (21.68)	16.59 (10.89)	13.61 (9.37)
		PHI	350	84060	138	48.19 (28.57)	26.36 (10.61)	60.94 (22.73)	13.4 (14.34)	7.42 (5.5)	13.66 (14.27)
	Electronics/Appliance	CHI	80	19354	67	59.56 (24.74)	31.5 (12.85)	61.85 (18.21)	7.89 (11.36)	12.69 (9.91)	13.88 (13.49)
		HOU	45	9910	36	56.6 (25.95)	29.25 (12.63)	41.11 (18.96)	14.45 (12.1)	28.6 (18.28)	12.01 (12.7)
		ATL	56	12590	51	53.49 (28.52)	28.39 (15.51)	39.83 (19.92)	27.98 (20.18)	14.94 (11.77)	12.09 (10.3)
		MIA	54	12255	49	48.57 (31.86)	21.61 (11.81)	35.26 (21.97)	14.23 (14.42)	42.29 (24.75)	3.69 (2.82)
		WAS	60	13852	52	45.02 (26.73)	27.56 (11.34)	41.08 (17.19)	21.86 (17.38)	17.35 (10.61)	13.73 (8.68)
		PHI	44	10140	34	56.23 (27.51)	27.58 (11.32)	64.63 (20.21)	8.57 (8.41)	6.46 (6.66)	16.12 (15.07)
	Department Stores	CHI	195	56096	142	49.91 (27.13)	26.84 (12.71)	56.61 (23.01)	11.61 (19.13)	17.9 (17.65)	10.42 (11.02)
		HOU	138	39369	90	52.62 (29.35)	26.97 (14.58)	40.39 (19.15)	15.21 (14.28)	28.05 (16.58)	12.18 (12)
		ATL	146	40237	97	49.47 (26.41)	27.3 (14.81)	41.7 (22.48)	32.43 (23.42)	12.68 (10.17)	8.5 (10.95)
		MIA	165	45545	113	51.36 (28.48)	22.36 (10.99)	33.59 (25.04)	15.57 (19.09)	43.18 (26.85)	3.58 (3.56)
		WAS	102	28660	76	42.97 (25.72)	25.03 (11.64)	37.54 (22.3)	27.47 (24.89)	17.08 (12.16)	11.94 (10.2)
		PHI	100	28488	78	53.96 (27.75)	24.47 (11.55)	64.18 (21.32)	12.3 (16.24)	8.19 (8.32)	11.05 (11.49)
	Home Furnishings	CHI	80	20115	63	59.24 (26.71)	32.76 (12.23)	66.77 (17.58)	6.11 (8.39)	10.9 (8.88)	12.72 (11.45)
		HOU	51	13170	41	56.73 (30.97)	29 (13.43)	40.76 (19.95)	16.57 (15.27)	22.83 (12.58)	15.06 (14.31)
		ATL	56	14825	41	56.01 (24.92)	34.9 (13.77)	47.34 (21.19)	26.23 (20.83)	12.14 (9.17)	9.23 (7.24)
		MIA	52	13126	44	42.77 (31.38)	22.4 (11.88)	43.57 (26.36)	17.3 (23.54)	31.37 (20.66)	3.06 (2.58)
		WAS	63	16090	52	46.99 (27.45)	30.85 (11.67)	47.53 (19.11)	15.77 (12.64)	16.58 (10.81)	14.39 (10.16)
		PHI	39	8578	37	61.77 (22.47)	26.59 (10.9)	68.59 (16.47)	8.51 (8.53)	6.84 (4.81)	11.78 (11.52)
	Shoes	CHI	180	43900	100	47.86 (26.51)	29.13 (12.83)	56.92 (19.01)	8.5 (12.23)	19.83 (15.84)	11.33 (11.47)
		HOU	150	37796	78	52.87 (29.98)	23.83 (12.49)	42.17 (21.64)	14.17 (12.9)	30.58 (19.28)	8.83 (10.32)
		ATL	135	33730	71	50.27 (28.5)	28.49 (13.92)	45.58 (22.65)	28.16 (22.49)	11.31 (7.2)	10.28 (12.15)
		MIA	147	35124	63	49.9 (28.41)	22.05 (11.23)	28.59 (21.57)	20.64 (23.39)	42.55 (25.31)	3.62 (2.87)
		WAS	101	23569	48	46.17 (25.51)	27.31 (12.72)	35.92 (19.18)	25.47 (20.63)	17.57 (10.74)	15.2 (9.71)
		PHI	68	16740	41	44.84 (32.24)	24.73 (11.36)	60.86 (21.75)	13.98 (14.21)	8.64 (5.43)	11.65 (12.65)
	Sporting/Hobby/Musical	CHI	346	85085	232	54.69 (27.08)	29.43 (11.93)	62.29 (18.8)	7.43 (11.15)	15.44 (13.82)	11.14 (11.31)
		HOU	211	52046	149	50.11 (28.58)	26.02 (14.02)	39.65 (19.95)	14.79 (13.3)	29.77 (17.71)	11.63 (13.98)
		ATL	236	59474	153	53.32 (28.49)	29.47 (14.15)	44.85 (21.86)	28.1 (22.54)	12.49 (9.92)	9.73 (10.93)
		MIA	185	43162	143	48.61 (28.33)	21.56 (9.99)	34.71 (23.97)	16.34 (19.17)	41.32 (26)	3.16 (2.56)
		WAS	183	43086	124	41.55 (28.26)	26.5 (12.7)	44.07 (22.3)	20.2 (20.09)	17.67 (14.79)	12.16 (9.91)
		PHI	157	37671	114	56.13 (27.02)	26.28 (11.41)	69.78 (16.81)	8.22 (8.52)	6.74 (5.34)	10.93 (11.18)

Table C.1: **RQ2: POI and neighborhood socio-demographics by economic group, category, and MSA.** Sample sizes are smaller than in RQ1 because some CBGs lack socio-demographic data (e.g., around airports). Cells shaded red denote strata with insufficient variation for estimation ($SD < 5$); these are excluded from reported regressions. Summary statistics are shown as *mean (sd)*. Abbreviations: n_cbgs = number of CBGs represented; n_pois = number of POIs; n_obs = number of observations per MSA $\times$ POI category.

	POI Type	MSA	R ² (Within)	R ² (Overall)	2019	2020	2021	2022	2023	2024						
Essential Goods	Grocery/ Convenience Stores	ATL	0.033	0.175	0.1344***	[0.1062, 0.1626]	0.1265***	[0.0961, 0.1569]	0.1342***	[0.1068, 0.1615]	0.1061***	[0.0758, 0.1364]	0.0999***	[0.0699, 0.1300]	0.1123***	[0.0834, 0.1413]
		CHI	0.153	0.194	0.1301***	[0.1066, 0.1536]	0.1402***	[0.1094, 0.1710]	0.0537***	[0.0260, 0.0813]	0.0659***	[0.0323, 0.0996]	0.0199	[-0.0079, 0.0476]	-0.0126	[-0.0393, 0.0142]
		HOU	0.031	0.217	0.1157***	[0.0872, 0.1442]	0.1228***	[0.0907, 0.1548]	0.0844***	[0.0534, 0.1153]	0.0849***	[0.0529, 0.1168]	0.0803***	[0.0432, 0.1174]	0.1138***	[0.0839, 0.1437]
		MIA	0.057	0.082	0.0762***	[0.0470, 0.1054]	0.0842***	[0.0469, 0.1215]	0.0526**	[0.0201, 0.0852]	0.0560***	[0.0272, 0.0847]	0.0309*	[0.0073, 0.0545]	0.0529***	[0.0265, 0.0793]
		PHI	0.024	0.147	0.0866***	[0.0450, 0.1283]	0.0981***	[0.0462, 0.1499]	0.0704**	[0.0266, 0.1142]	0.0746***	[0.0323, 0.1169]	0.0944***	[0.0490, 0.1398]	0.1109***	[0.0676, 0.1541]
		WAS	0.041	0.229	0.1651***	[0.1474, 0.1829]	0.1583***	[0.1207, 0.1959]	0.0777***	[0.0614, 0.0940]	0.0301*	[0.0063, 0.0539]	-0.0198	[-0.0411, 0.0016]	0.1021	[-0.0056, 0.0314]
Routine Services	Personal Care	ATL	0.028	-0.059	0.0030	[-0.0096, 0.0155]	0.0172*	[0.0006, 0.0338]	0.0194**	[0.0052, 0.0336]	0.0271***	[0.0126, 0.0416]	0.0279***	[0.0090, 0.0469]	0.0258**	[0.0097, 0.0418]
		CHI	0.026	0.029	0.0482***	[0.0370, 0.0595]	0.0371***	[0.0197, 0.0546]	0.0158**	[0.0045, 0.0271]	0.0308***	[0.0193, 0.0422]	0.0051	[-0.0102, 0.0205]	-0.0016	[-0.0142, 0.0111]
		HOU	0.018	0.060	0.0256***	[0.0129, 0.0383]	0.0245**	[0.0098, 0.0391]	0.0370***	[0.0243, 0.0496]	0.0362***	[0.0237, 0.0487]	0.0261*	[0.0053, 0.0469]	0.0428***	[0.0296, 0.0560]
		MIA	0.034	-0.076	0.0150*	[0.0029, 0.0272]	0.0253*	[0.0021, 0.0485]	0.0118	[-0.0005, 0.0241]	0.0201**	[0.0056, 0.0346]	0.0075	[-0.0066, 0.0216]	0.0105	[-0.0026, 0.0235]
		PHI	0.009	0.044	0.0255***	[0.0117, 0.0392]	0.0424***	[0.0233, 0.0616]	0.0446***	[0.0291, 0.0602]	0.0326***	[0.0167, 0.0486]	0.0254*	[0.0032, 0.0475]	0.0398***	[0.0241, 0.0555]
		WAS	0.045	-0.012	0.0527***	[0.0416, 0.0638]	0.0414**	[0.0131, 0.0698]	0.0203***	[0.0083, 0.0323]	0.0142*	[0.0015, 0.0270]	0.0066	[-0.0118, 0.0250]	0.0155**	[0.0010, 0.0301]
Routine Services	Drycleaning/ Laundry	ATL	0.096	-0.306	-0.0583*	[-0.1053, -0.0112]	0.0562*	[0.0008, 0.1117]	0.1258***	[0.0534, 0.1982]	0.2481***	[0.1611, 0.3351]	0.0756*	[0.0158, 0.1353]	0.1514***	[0.0785, 0.2234]
		CHI	0.119	-0.204	0.0009 [-0.0332, 0.0351]	-0.0015 [-0.0503, 0.0473]	0.0173 [-0.0182, 0.0528]	0.0834***	[0.0391, 0.1276]	0.0568***	[0.0178, 0.0958]	0.0611**	[0.0156, 0.1066]			
		HOU	0.067	-0.134	0.0074 [-0.0213, 0.0362]	0.0642**	[0.0169, 0.1116]	0.0458**	[0.0077, 0.0838]	-0.0021 [-0.0365, 0.0323]	0.0478**	[0.0160, 0.0796]	0.0441*	[0.0051, 0.0830]		
		MIA	0.119	-0.037	0.1166***	[0.0823, 0.1510]	0.0718***	[0.0366, 0.1069]	0.0333 [-0.0107, 0.0772]	0.0359 [-0.0186, 0.0904]	0.0592**	[0.0158, 0.1026]	0.0756**	[0.0122, 0.1390]		
		PHI	0.045	-0.018	0.0395 [-0.0212, 0.1002]	0.0112 [-0.0628, 0.0852]	-0.0093 [-0.0637, 0.0451]	-0.0715***	[-0.1258, -0.0172]	0.0856*	[0.0038, 0.1674]	0.1249***	[0.0705, 0.1794]			
		WAS	0.211	-0.311	-0.0182 [-0.0461, 0.0097]	-0.0175 [-0.0534, 0.0184]	0.0087 [-0.0254, 0.0429]	0.0042 [-0.0295, 0.0378]			-0.0064 [-0.0379, 0.0252]	0.0218 [-0.0115, 0.0551]				
Social & Experiential	Restaurants	ATL	0.055	0.096	0.0963***	[0.0603, 0.1323]	0.1177***	[0.0692, 0.1663]	0.0923***	[0.0580, 0.1266]	0.1038***	[0.0677, 0.1400]	0.1251***	[0.0867, 0.1635]	0.1425***	[0.1045, 0.1805]
		CHI	0.082	0.149	0.1043***	[0.0681, 0.1405]	0.1276***	[0.0761, 0.1791]	0.1118***	[0.0739, 0.1498]	0.1046***	[0.0659, 0.1432]	0.1177***	[0.0758, 0.1596]	0.1357***	[0.0966, 0.1748]
		HOU	0.030	0.209	0.1130***	[0.0807, 0.1454]	0.1203***	[0.0773, 0.1632]	0.1069***	[0.0737, 0.1401]	0.0999***	[0.0662, 0.1337]	0.1069***	[0.0699, 0.1439]	0.1159***	[0.0810, 0.1508]
		MIA	0.053	0.052	0.0867***	[0.0477, 0.1256]	0.0976***	[0.0401, 0.1552]	0.0801***	[0.0415, 0.1187]	0.0960***	[0.0572, 0.1347]	0.0966***	[0.0563, 0.1369]	0.1058***	[0.0657, 0.1460]
		PHI	0.022	0.179	0.1205***	[0.0869, 0.1541]	0.1287***	[0.0845, 0.1730]	0.0962***	[0.0630, 0.1293]	0.0939***	[0.0592, 0.1285]	0.0948***	[0.0604, 0.1292]	0.1086***	[0.0745, 0.1427]
		WAS	0.037	0.046	0.0755***	[0.0378, 0.1133]	0.1079***	[0.0568, 0.1591]	0.0923***	[0.0510, 0.1335]	0.0879***	[0.0467, 0.1291]	0.1021***	[0.0602, 0.1440]	0.1281***	[0.0880, 0.1631]
Social & Experiential	Bars/Pubs	ATL	0.109	0.374	0.3445***	[0.3111, 0.3780]	0.2717***	[0.2050, 0.3384]	0.1466***	[0.1126, 0.1806]	0.1945***	[0.1622, 0.2268]	0.1392***	[0.0979, 0.1806]	0.1954***	[0.1600, 0.2308]
		CHI	0.031	0.281	0.2204***	[0.1875, 0.2533]	0.1827***	[0.1355, 0.2300]	0.1048***	[0.0777, 0.1319]	0.0965***	[0.0691, 0.1239]	0.1300***	[0.0907, 0.1692]	0.1435***	[0.1110, 0.1761]
		HOU	0.045	0.279	0.1934***	[0.1647, 0.2221]	0.1924***	[0.1498, 0.2351]	0.1551***	[0.1228, 0.1873]	0.0899***	[0.0567, 0.1232]	0.0950***	[0.0667, 0.1234]	0.1594***	[0.1229, 0.1959]
		MIA	0.086	0.272	0.2623***	[0.2155, 0.3091]	0.2273***	[0.1489, 0.3058]	0.1987***	[0.1295, 0.2289]	0.1987***	[0.1513, 0.2461]	0.2159***	[0.1521, 0.2797]	0.1537***	[0.0953, 0.2120]
		PHI	0.053	0.298	0.2270***	[0.1901, 0.2640]	0.2863***	[0.2278, 0.3448]	0.2528***	[0.2161, 0.2895]	0.2008***	[0.1678, 0.2339]	0.1105***	[0.0756, 0.1454]	0.0978***	[0.0559, 0.1397]
		WAS	0.133	0.004	0.1086***	[0.0615, 0.1557]	0.2343***	[0.1623, 0.3063]	0.1093***	[0.0567, 0.1619]	0.1067***	[0.0630, 0.1504]	0.0952***	[0.0504, 0.1401]	0.0415	[-0.0100, 0.0929]
Social & Experiential	Specialty Food Stores	ATL	0.048	0.160	0.1265***	[0.0890, 0.1641]	0.1612***	[0.1171, 0.2053]	0.1003***	[0.0606, 0.1401]	0.1289***	[0.0914, 0.1664]	0.1339***	[0.0869, 0.1809]	0.1919***	[0.1446, 0.2393]
		CHI	0.033	0.166	0.1208***	[0.0849, 0.1567]	0.1014***	[0.0565, 0.1463]	0.0891***	[0.0480, 0.1302]	0.0967***	[0.0566, 0.1369]	0.1138***	[0.0709, 0.1568]	0.0751***	[0.0372, 0.1131]
		HOU	0.007	0.021	-0.0090 [-0.0502, 0.0323]	0.0199 [-0.0242, 0.0640]	0.0247 [-0.0136, 0.0630]	0.0410* [0.0004, 0.0816]	0.0440 [-0.0089, 0.0969]	0.1459***	[0.0721, 0.2197]					
		MIA	0.036	0.108	0.0987***	[0.0536, 0.1439]	0.1080***	[0.0589, 0.1572]	0.0263 [-0.0186, 0.0712]	0.0911***	[0.0391, 0.1430]	0.0856***	[0.0486, 0.1226]	0.0199	[-0.0346, 0.0744]	
		PHI	0.019	0.248	0.1535***	[0.1222, 0.1848]	0.1542***	[0.1237, 0.1847]	0.1451***	[0.1099, 0.1804]	0.1755***	[0.1397, 0.2112]	0.0754***	[0.0314, 0.1194]	0.0996***	[0.0667, 0.1326]
		WAS	0.024	0.231	0.1914***	[0.1474, 0.2354]	0.2233***	[0.1729, 0.2737]	0.2191***	[0.1825, 0.2557]	0.1602***	[0.1205, 0.2000]	0.0661***	[0.0205, 0.1117]	0.0842*	[0.0200, 0.1484]
Discretionary Retail	Clothing Stores	ATL	0.077	0.120	0.1311***	[0.0817, 0.1805]	0.1563***	[0.1024, 0.2103]	0.1527***	[0.1042, 0.2013]	0.1298***	[0.0815, 0.1781]	0.0962***	[0.0533, 0.1571]	0.1330***	[0.0817, 0.1843]
		CHI	0.053	0.185	0.1547***	[0.1023, 0.2072]	0.1887***	[0.1237, 0.2536]	0.1456***	[0.0884, 0.2028]	0.1316***	[0.0725, 0.1907]	0.1283***	[0.0659, 0.2008]	0.1733***	[0.1149, 0.2317]
		HOU	0.035	0.256	0.1817***	[0.1341, 0.2294]	0.1899***	[0.1376, 0.2423]	0.1783***	[0.1231, 0.2335]	0.1385***	[0.0906, 0.1865]	0.1363***	[0.0567, 0.2059]	0.2099***	[0.1603, 0.2595]
		MIA	0.106	0.186	0.2003***	[0.1488, 0.2518]	0.2250***	[0.1639, 0.2861]	0.1610***	[0.1110, 0.2110]	0.1544***	[0.1079, 0.2009]	0.1344***	[0.0748, 0.1941]	0.1799***	[0.1261, 0.2336]
		PHI	0.029	0.206	0.1579***	[0.1071, 0.2087]	0.1573***	[0.1021, 0.2125]	0.1452***	[0.0902, 0.2001]	0.1284***	[0.0757, 0.1810]	0.1051***	[0.0393, 0.1708]	0.1669***	[0.1160, 0.2178]
		WAS	0.079	0.136	0.1557***	[0.0959, 0.2155]	0.1913***	[0.1150, 0.2675]	0.1399***	[0.0724, 0.2074]	0.1246***	[0.0615, 0.1877]	0.0878*	[0.0086, 0.1671]	0.1299***	[0.0681, 0.1918]
Discretionary Retail	Electronics/ Appliance Stores	ATL	0.083	0.381	0.3249***	[0.2358, 0.4139]	0.3141***	[0.2003, 0.4278]	0.3199***	[0.2332, 0.4066]	0.2102***	[0.1139, 0.3065]	0.1054*	[0.0195, 0.1914]	0.1276* [0.0164, 0.2388]	
		CHI	0.087	0.448	0.3596***	[0.2752, 0.4439]	0.4566***	[0.3436, 0.5696]	0.3309***	[0.2486, 0.4132]	0.2735***	[0.1898, 0.3571]	0.2191***	[0.1400, 0.2983]	0.3132***	[0.2267, 0.3997]
		HOU	0.082	0.448	0.3500***	[0.2053, 0.4947]	0.5060***	[0.3659, 0.6462]	0.4492***	[0.3064, 0.5919]	0.4675***	[0.3215, 0.6134]	0.3448***	[0.1763, 0.5133]	0.4666***	[0.3190, 0.6143]
		MIA	0.190	0.521	0.4839***	[0.4037, 0.5640]	0.4819***	[0.4134, 0.5503]	0.3471***	[0.2700, 0.4242]	0.2418***	[0.1607, 0.3229]	0.1598***	[0.0839, 0.2358]	0.1632***	[0.0770, 0.2494]
		PHI	0.073	0.411	0.3399***	[0.2145, 0.4653]	0.4118***	[0.2144, 0.6092]	0.2966***	[0.1622, 0.4310]	0.2703***	[0.1472, 0.3933]	0.1974***	[0.0796, 0.3152]	0.2372***	[0.1019, 0.3725]
		WAS	0.062	0.269	0.2545***	[0.1593, 0.3496]	0.2875***	[0.1542, 0.4208]	0.1859***	[0.0820, 0.2898]	0.1463**	[0.0459, 0.2467]	0.1498***	[0.0528, 0.2468]	0.1823***	[0.0756, 0.2890]
Discretionary Retail	Department Stores	ATL	0.121	0.412	0.3180***	[0.1935, 0.4425]	0.3680***	[0.1947, 0.5412]	0.2892***	[0.1602, 0.4182]	0.3349***	[0.2058, 0.4639]	0.2091***	[0.0724, 0.3459]	0.2438***	[0.1059, 0.3818]
		CHI	0.193	0.471	0.3253***	[0.2086, 0.4420]	0.4367***	[0.2403, 0.6331]	0.2807***	[0.1538, 0.4076]	0.3339***	[0.1958, 0.4720]	0.2532***	[0.1170, 0.3893]	0.2234***	[0.0960, 0.3507]
		HOU	0.150	0.652	0.4502***	[0.3411, 0.5593]	0.4584***	[0.2865, 0.6302]	0.4004***	[0.2688, 0.5321]	0.2988***	[0.1760, 0.4217]	0.1825*	[0.0432, 0.3217]	0.2378***	[0.1257, 0.3499]
		MIA	0.165	0.437	0.3457***	[0.										

	POI Category	MSA	Mean Resp.	Stability CV	Reliability	COVID Impact	Geo Consist.
Essential Goods	Grocery Stores	ATL	0.119	0.114	1.000	0.000	0.803
		CHI	0.098	0.391	0.667	-0.076	
		HOU	0.100	0.174	1.000	0.000	
		MIA	0.059	0.295	1.000	0.000	
		PHI	0.089	0.155	1.000	0.000	
		WAS	0.108	0.524	0.667	-0.087	
Routine Services	DryCleaning/Laundry	ATL	0.100	0.940	1.000	0.149	0.736
		CHI	—	—	0.500	—	
		HOU	0.051	0.160	0.667	—	
		MIA	0.081	0.267	0.667	0.000	
		PHI	—	—	0.500	—	
		WAS	—	—	0.000	—	
	Personal Care	ATL	0.024	0.185	0.833	—	0.868
		CHI	0.033	0.356	0.667	-0.033	
		HOU	0.032	0.217	1.000	0.000	
		MIA	—	—	0.500	—	
		PHI	0.035	0.221	1.000	0.000	
		WAS	0.029	0.536	0.833	-0.032	
Social & Experiential	Bars/Pubs	ATL	0.215	0.335	1.000	-0.198	0.810
		CHI	0.146	0.297	1.000	-0.116	
		HOU	0.148	0.282	1.000	0.000	
		MIA	0.206	0.168	1.000	0.000	
		PHI	0.196	0.357	1.000	0.000	
		WAS	0.131	0.397	0.833	0.126	
	Restaurants	ATL	0.113	0.155	1.000	0.000	0.925
		CHI	0.117	0.099	1.000	0.000	
		HOU	0.111	0.061	1.000	0.000	
		MIA	0.094	0.088	1.000	0.000	
		PHI	0.107	0.126	1.000	0.000	
		WAS	0.099	0.168	1.000	0.000	
	Specialty Food Stores	ATL	0.141	0.207	1.000	0.000	0.809
		CHI	0.100	0.152	1.000	0.000	
		HOU	—	—	0.333	—	
		MIA	0.096	0.088	0.667	0.000	
		PHI	0.134	0.260	1.000	0.000	
		WAS	0.157	0.393	1.000	0.000	
	Clothing	ATL	0.133	0.147	1.000	0.000	0.893
		CHI	0.154	0.141	1.000	0.000	
		HOU	0.173	0.155	1.000	0.000	
		MIA	0.176	0.171	1.000	0.000	
		PHI	0.143	0.147	1.000	0.000	
		WAS	0.138	0.227	1.000	0.000	
	Department Stores	ATL	0.294	0.184	1.000	0.000	0.919
		CHI	0.309	0.223	1.000	0.000	
		HOU	0.338	0.312	1.000	0.000	
		MIA	0.331	0.089	1.000	0.000	
		PHI	0.351	0.162	1.000	0.000	
		WAS	0.376	0.205	1.000	0.000	
Discretionary Retail	Electronics/Appliance	ATL	0.234	0.392	1.000	0.000	0.756
		CHI	0.326	0.226	1.000	0.000	
		HOU	0.431	0.142	1.000	0.000	
		MIA	0.313	0.432	1.000	0.000	
		PHI	0.292	0.238	1.000	0.000	
		WAS	0.201	0.261	1.000	0.000	
	Home Furnishings	ATL	0.295	0.190	1.000	0.000	0.854
		CHI	0.239	0.255	1.000	0.000	
		HOU	0.339	0.224	1.000	0.000	
		MIA	0.263	0.173	1.000	0.000	
		PHI	0.258	0.223	1.000	0.000	
		WAS	0.219	0.266	1.000	0.000	
	Shoe Stores	ATL	0.134	0.141	1.000	0.000	0.638
		CHI	0.133	0.209	0.833	0.000	
		HOU	0.283	0.059	1.000	0.000	
		MIA	0.118	0.212	1.000	0.000	
		PHI	0.148	0.215	1.000	0.000	
		WAS	0.128	0.200	1.000	0.000	
	Sporting/Hobby/Music	ATL	0.195	0.116	1.000	0.000	0.842
		CHI	0.237	0.138	1.000	0.000	
		HOU	0.207	0.128	1.000	0.000	
		MIA	0.148	0.268	1.000	0.000	
		PHI	0.214	0.148	1.000	0.000	
		WAS	0.160	0.133	1.000	0.000	

Table C.3: Category-Specific Elasticity Performance Metrics by Metropolitan Area.

Table presents elasticity performance metrics for visit–spend relationships across business categories and metropolitan areas (2019–2024). *Mean Resp.* = average responsiveness (elasticity) across years; *Stability CV* = coefficient of variation measuring temporal consistency, Score = 1 - CV reported in main manuscript; *Reliability* = statistical significance rate; *COVID Impact* = elasticity disruption during 2020–2021 (negative values shaded); *Geo Consist.* = geographic consistency across MSAs. “—” indicates metrics could not be derived due to not meeting the reliability threshold.

	Category	MSA	Elasticity_2019	Elasticity_2020	Change_%	Increased	Sig_2019	Sig_2020
Essential Goods	Grocery Stores	ATL	0.134	0.126	-5.9	false	true	true
		CHI	0.130	0.140	7.8	true	true	true
		HOU	0.116	0.123	6.2	true	true	true
		MIA	0.076	0.084	10.6	true	true	true
		PHI	0.087	0.098	13.2	true	true	true
		WAS	0.165	0.158	-4.1	false	true	true
Routine Services	Personal Care Services	ATL	0.003	0.017	481.0	true	false	true
		CHI	0.048	0.037	-23.0	false	true	true
		HOU	0.026	0.024	-4.3	false	true	true
		MIA	0.015	0.025	68.3	true	true	true
		PHI	0.025	0.042	66.6	true	true	true
		WAS	0.053	0.041	-21.4	false	true	true
	Drycleaning/Laundry	ATL	-0.058	0.056	196.5	true	true	true
		CHI	0.001	-0.001	-260.1	false	false	false
		HOU	0.007	0.064	764.5	true	false	true
		MIA	0.117	0.072	-38.5	false	true	true
		PHI	0.040	0.011	-71.6	false	false	false
		WAS	-0.018	-0.018	3.6	true	false	false
Social & Experiential	Restaurants	ATL	0.096	0.118	22.3	true	true	true
		CHI	0.104	0.128	22.3	true	true	true
		HOU	0.113	0.120	6.4	true	true	true
		MIA	0.087	0.098	12.7	true	true	true
		PHI	0.121	0.129	6.8	true	true	true
		WAS	0.076	0.108	42.9	true	true	true
	Bars/Pubs	ATL	0.345	0.272	-21.1	false	true	true
		CHI	0.220	0.183	-17.1	false	true	true
		HOU	0.193	0.192	-0.5	false	true	true
		MIA	0.262	0.227	-13.3	false	true	true
		PHI	0.227	0.286	26.1	true	true	true
		WAS	0.109	0.234	115.8	true	true	true
	Specialty Food Stores	ATL	0.127	0.161	27.4	true	true	true
		CHI	0.121	0.101	-16.1	false	true	true
		HOU	-0.009	0.020	322.0	true	false	false
		MIA	0.099	0.108	9.5	true	true	true
		PHI	0.154	0.154	0.4	true	true	true
		WAS	0.191	0.223	16.7	true	true	true
Discretionary Retail	Clothing Stores	ATL	0.131	0.156	19.2	true	true	true
		CHI	0.155	0.189	21.9	true	true	true
		HOU	0.182	0.190	4.5	true	true	true
		MIA	0.200	0.225	12.3	true	true	true
		PHI	0.158	0.157	-0.3	false	true	true
		WAS	0.156	0.191	22.8	true	true	true
	Department Stores	ATL	0.318	0.368	15.7	true	true	true
		CHI	0.325	0.437	34.2	true	true	true
		HOU	0.450	0.458	1.8	true	true	true
		MIA	0.346	0.367	6.1	true	true	true
		PHI	0.364	0.443	21.8	true	true	true
		WAS	0.399	0.511	28.0	true	true	true
	Electronics/Appliance Stores	ATL	0.325	0.314	-3.3	false	true	true
		CHI	0.360	0.457	27.0	true	true	true
		HOU	0.350	0.506	44.6	true	true	true
		MIA	0.484	0.482	-0.4	false	true	true
		PHI	0.340	0.412	21.2	true	true	true
		WAS	0.254	0.288	13.0	true	true	true
	Home Furnishings	ATL	0.190	0.371	95.1	true	true	true
		CHI	0.149	0.285	91.3	true	true	true
		HOU	0.375	0.449	19.7	true	true	true
		MIA	0.195	0.306	57.1	true	true	true
		PHI	0.256	0.371	45.0	true	true	true
		WAS	0.138	0.292	111.0	true	true	true
	Shoe Stores	ATL	0.136	0.153	12.7	true	true	true
		CHI	0.134	0.165	23.3	true	true	true
		HOU	0.284	0.276	-2.8	false	true	true
		MIA	0.163	0.135	-17.1	false	true	true
		PHI	0.193	0.143	-26.1	false	true	true
		WAS	0.169	0.148	-12.6	false	true	true
	Sporting/Hobby/Musical Stores	ATL	0.150	0.212	40.7	true	true	true
		CHI	0.196	0.264	34.6	true	true	true
		HOU	0.227	0.249	9.5	true	true	true
		MIA	0.181	0.214	18.5	true	true	true
		PHI	0.158	0.236	49.5	true	true	true
		WAS	0.130	0.172	32.5	true	true	true

Table C.4: % Changes in elasticity estimates in 2020 compared to 2019 across MSAs. The true tag in increased column indicates whether the elasticity estimate increased in 2020 compared to 2019. The Sig_2019 and Sig_2020 indicates whether the estimates were significant for 2019 and 2020 respectively.

Category	MSA	N_Obs	N_poi	R^2 (Within)	R^2 (Overall)	2019	2020	2021	2022	2023	2024
BRANDED											
Restaurants	ATL	933567	3497	0.057	0.121	0.1032*** [0.0636, 0.1429]	0.1261*** [0.0711, 0.1811]	0.0968*** [0.0591, 0.1344]	0.1194*** [0.0809, 0.1580]	0.1414*** [0.0983, 0.1844]	0.1565*** [0.1144, 0.1985]
Restaurants	CHI	1266846	4779	0.157	0.193	0.1165*** [0.0782, 0.1547]	0.1579*** [0.1047, 0.2110]	0.1285*** [0.0905, 0.1664]	0.1243*** [0.0870, 0.1616]	0.1276*** [0.0808, 0.1745]	0.1468*** [0.1065, 0.1872]
Restaurants	HOU	938587	3480	0.039	0.224	0.1152*** [0.0761, 0.1543]	0.1317*** [0.0794, 0.1841]	0.1161*** [0.0761, 0.1560]	0.1135*** [0.0744, 0.1525]	0.1160*** [0.0717, 0.1603]	0.1244*** [0.0824, 0.1664]
Restaurants	MIA	687014	2542	0.039	0.119	0.1059*** [0.0658, 0.1459]	0.1193*** [0.0606, 0.1779]	0.0850*** [0.0467, 0.1233]	0.0982*** [0.0598, 0.1367]	0.0926*** [0.0495, 0.1356]	0.1105*** [0.0681, 0.1529]
Restaurants	PHI	550203	2094	0.045	0.193	0.1179*** [0.0790, 0.1568]	0.1633*** [0.1091, 0.2175]	0.1114*** [0.0744, 0.1484]	0.1121*** [0.0752, 0.1490]	0.0928*** [0.0543, 0.1314]	0.1021*** [0.0639, 0.1402]
Restaurants	WAS	689614	2675	0.026	0.066	0.0667** [0.0240, 0.1094]	0.1150*** [0.0576, 0.1724]	0.0929*** [0.0486, 0.1372]	0.1004*** [0.0574, 0.1434]	0.1095*** [0.0614, 0.1575]	0.1444*** [0.0998, 0.1889]
NON-BRANDED											
Restaurants	ATL	420781	1853	0.065	0.116	0.1356*** [0.1105, 0.1608]	0.1234*** [0.0808, 0.1661]	0.1018*** [0.0759, 0.1276]	0.1193*** [0.0937, 0.1450]	0.0804*** [0.0546, 0.1062]	0.0845*** [0.0574, 0.1117]
Restaurants	CHI	697711	3213	0.042	0.181	0.1631*** [0.1384, 0.1878]	0.1434*** [0.0952, 0.1916]	0.1009*** [0.0730, 0.1289]	0.0979*** [0.0706, 0.1251]	0.0836*** [0.0583, 0.1090]	0.0769*** [0.0505, 0.1033]
Restaurants	HOU	468031	2153	0.028	0.246	0.1532*** [0.1342, 0.1721]	0.1264*** [0.0896, 0.1632]	0.1127*** [0.0925, 0.1330]	0.0995*** [0.0790, 0.1200]	0.0787*** [0.0567, 0.1008]	0.0764*** [0.0551, 0.0978]
Restaurants	MIA	440181	2067	0.094	0.076	0.1381*** [0.1090, 0.1671]	0.1360*** [0.0810, 0.1909]	0.0937*** [0.0633, 0.1242]	0.1038*** [0.0738, 0.1339]	0.0800*** [0.0516, 0.1084]	0.0605*** [0.0284, 0.0927]
Restaurants	PHI	388255	1857	0.024	0.225	0.1756*** [0.1524, 0.1989]	0.1232*** [0.0757, 0.1707]	0.1030*** [0.0782, 0.1278]	0.1084*** [0.0854, 0.1314]	0.0707*** [0.0484, 0.0929]	0.0748*** [0.0507, 0.0988]
Restaurants	WAS	383005	1798	0.085	0.123	0.1757*** [0.1530, 0.1984]	0.1547*** [0.1071, 0.2024]	0.1044*** [0.0782, 0.1305]	0.0796*** [0.0542, 0.1050]	0.0615*** [0.0363, 0.0866]	0.0502*** [0.0248, 0.0756]

Table C.5: **Branded vs. non-branded restaurant elasticities by MSA, 2019–2024.** Annual within-POI spend–visit elasticities estimated separately for branded and non-branded segments using the panel specification in Equation (1) in main manuscript with POI and year fixed effects. Columns report sample size (N_Obs , N_poi), within- and overall- R^2 , and year-specific elasticities with 95% confidence intervals in brackets. Asterisks denote conventional significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

	POI Category	MSA	Income	Education	% Black	% Hispanic	% Asian
Essential Goods	Grocery Stores	ATL	-0.0011***	-0.0007	-0.0013***	-0.0038***	-0.0024***
		CHI	0.0007***	0.0017***	-0.0004	-0.0005**	0.0052***
		HOU	-0.0013***	0.0016**	0.0002	-0.0021***	0.0023***
		MIA	-0.0006***	0.0071***	-0.0020***	0.0000	
		PHI	0.0004	0.0020***	-0.0003	-0.0018***	0.0020***
		WAS	-0.0006***	0.0029***	0.0003	-0.0042***	-0.0019***
Routine Services	Personal Care Services	ATL	-0.0006***	0.0008*	0.0000	-0.0012**	0.0007*
		CHI	0.0005***	0.0019***	-0.0002	0.0004	0.0011**
		HOU	0.0002	0.0023***	0.0002	-0.0005	-0.0009***
		MIA	-0.0002	0.0025***	-0.0013***	-0.0005***	
		PHI	0.0001	0.0002	0.0001	0.0004	0.0014***
		WAS	0.0008***	0.0015***	0.0010***	-0.0005	0.0016***
	Drycleaning/Laundry Services	ATL	0.0001	0.0009	0.0013	-0.0003	-0.0056***
		CHI	0.0009*	0.0021*	-0.0013*	0.0001	0.0027*
		HOU	-0.0014*	0.0039***	0.0006	-0.0010	-0.0012
		MIA	0.0012	0.0000	0.0001	-0.0011	
		PHI	-0.0020	-0.0007	-0.0074***	-0.0055**	0.0044**
		WAS	-0.0026***	-0.0011	-0.0055***	-0.0033***	-0.0032*
Social & Experiential	Restaurants	ATL	0.0004***	0.0015***	0.0004***	-0.0009***	-0.0001
		CHI	0.0003***	0.0035***	-0.0013***	-0.0006***	0.0018***
		HOU	0.0002	0.0020***	-0.0003*	-0.0021***	0.0001
		MIA	-0.0002**	0.0021***	-0.0021***	-0.0006***	
		PHI	-0.0004***	0.0027***	-0.0015***	-0.0015***	0.0003
		WAS	-0.0002**	0.0017***	-0.0010***	-0.0024***	-0.0010***
	Drinking Places (Alcoholic Beverages)	ATL	-0.0018***	0.0021*	-0.0036***	-0.0038**	0.0010
		CHI	0.0015***	0.0045***	0.0014*	0.0008	0.0042***
		HOU	0.0022***	0.0030***	0.0033***	0.0020***	-0.0011
		MIA	0.0017**	-0.0007	0.0019*	-0.0014*	
		PHI	-0.0002	0.0058***	0.0006	0.0063***	-0.0014
		WAS	-0.0014***	0.0027*	-0.0010	-0.0000	-0.0042***
	Specialty Food Stores	ATL	0.0012**	-0.0026*	-0.0008	-0.0064***	-0.0017
		CHI	0.0016**	-0.0016	-0.0015*	-0.0027***	0.0038***
		HOU	-0.0014*	0.0144***	-0.0042**	-0.0035***	-0.0096***
		MIA	0.0017*	-0.0012	-0.0007	-0.0014*	
		PHI	0.0016**	0.0049***	0.0027***	-0.0005	0.0026
		WAS	0.0013**	-0.0030	-0.0003	-0.0066***	-0.0026
	Clothing Stores	ATL	-0.0006**	0.0004	-0.0007	0.0001	-0.0025***
		CHI	0.0002	0.0009**	-0.0025***	-0.0004	0.0015***
		HOU	0.0014***	-0.0008	0.0015***	-0.0021***	-0.0016***
		MIA	0.0006**	0.0025***	-0.0004	0.0006	
		PHI	-0.0021***	0.0012*	-0.0035***	-0.0033**	0.0002
		WAS	0.0001	0.0032***	0.0012***	-0.0007	0.0037***
	Shoe Stores	ATL	-0.0004	0.0003	-0.0005	-0.0044**	-0.0041***
		CHI	-0.0003	-0.0012	-0.0027***	-0.0001	-0.0002
		HOU	0.0006	0.0030***	0.0040***	0.0005	-0.0041***
		MIA	0.0013***	-0.0000	-0.0009	-0.0006	
		PHI	-0.0002	0.0043***	0.0027*	-0.0059	-0.0007
		WAS	0.0013***	-0.0001	0.0022***	-0.0014	0.0069***
Discretionary Retail	Department Stores	ATL	0.0024***	-0.0008	0.0004	-0.0060***	-0.0040***
		CHI	0.0024***	0.0001	-0.0005	-0.0005	0.0007
		HOU	0.0008	0.0023*	0.0032***	0.0002	0.0049***
		MIA	-0.0008	0.0023	-0.0036***	-0.0024***	
		PHI	-0.0001	0.0088***	-0.0016**	0.0030*	-0.0030*
		WAS	0.0017***	-0.0004	0.0001	-0.0029***	0.0008
	Electronics and Appliance Stores	ATL	0.0052***	-0.0070***	-0.0002	0.0098***	0.0063***
		CHI	0.0034***	-0.0029**	0.0035***	0.0056***	0.0053***
		HOU	0.0016	0.0003	-0.0019	0.0034	0.0027
		MIA	0.0022***	0.0056**	-0.0010	0.0033***	
		PHI	0.0006	-0.0043	-0.0091**	0.0018	0.0021*
		WAS	-0.0018***	0.0078***	0.0066***	0.0056**	0.0003
	Home Furnishings Stores	ATL	0.0015	-0.0036**	0.0043***	0.0049	0.0042*
		CHI	-0.0006	-0.0005	-0.0046*	-0.0062***	0.0024**
		HOU	-0.0027**	-0.0002	0.0010	-0.0097***	-0.0095***
		MIA	0.0015	-0.0006	-0.0005	0.0039**	
		PHI	-0.0031**	-0.0024	0.0009		-0.0020
		WAS	0.0003	0.0006	-0.0011	-0.0051***	-0.0017
	Sports/Hobby/Music	ATL	-0.0009*	-0.0011*	-0.0013***	0.0020	-0.0006
		CHI	0.0021***	0.0003	0.0013**	0.0014**	0.0043***
		HOU	0.0008*	0.0005	0.0003	0.0005	0.0012*
		MIA	0.0016***	-0.0067***	-0.0015	0.0003	
		PHI	0.0004	0.0010	-0.0018	-0.0078***	-0.0005
		WAS	0.0001	-0.0002	0.0005	0.0026***	0.0022**

Table C.6: **Demographic Moderation of Visit-Spend Elasticities by Economic Category.** Table shows demographic moderation coefficients from category-specific panel regressions. Each coefficient represents the interaction between log(visits) and the respective demographic variable. Income is measured as median household income percentile, Education as percentage with bachelor’s degrees, and racial/ethnic variables as population percentages. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. All regressions include POI fixed effects and year fixed effects with Driscoll-Kraay standard errors. Statistically significant estimates are shaded green for amplification and red for attenuation.

Appendix D: Evaluating LLM Priors About Neighborhood-Level Mobility

D.1 System Prompts

D.1.1 Appendix: Prompt Used for Zero-Shot Mobility Classification

For the zero-shot mobility outcome predictions from the LLMs, we used the following system prompt.

System Prompt for Mobility Outcome Classification

You are a careful classifier for mobility outcomes.

Output ONLY one XML document that starts with <response> and ends with

</response>, with EXACTLY three child elements in this order:

<reasoning>, <rank>, <answer>.

Do not output anything else.

Primary task: Predict the mobility outcome class in <answer> using ONLY the information provided in the user prompt (outcome definition, CBSA-specific cutpoints, and CBG feature buckets/percentiles).

Keep <reasoning> concise and focused on direction and justification:

explain how the most influential features (the ones you will rank highest) push the mobility outcome up or down, and why they matter more than other features. Avoid listing everything; summarize.

<rank> is an explanation-only section. List the 10 most relevant predictors for THIS CBG's prediction (most important first).

<rank> MUST contain exactly 10 <item/> tags.

Each <item/> MUST have EXACTLY three attributes:

feature="..." val="very_low|low|neutral|high|very_high" push="+"|"-"|"0"

- val MUST be copied exactly from the bucket given in the prompt for that feature.

- push is the direction this feature (given its val) pushes the mobility outcome for THIS CBG: "+" pushes toward higher outcome, "-" toward lower, "0" neutral or unclear.

<answer> MUST be exactly one of: A | B | C.

A = low outcome (lowest tertile),

B = middle tertile,

C = high outcome (highest tertile).

The tertile cutpoints for A/B/C are provided in the prompt.

When deciding <answer>, weigh items by rank: items earlier in <rank>

matter much more than later items. Treat ranks 1-3 as primary evidence, ranks 4-6 as supporting evidence, and ranks 7-10 as weak evidence.

D.1.2 Prompt Used for Directional Alignment Analysis

To elicit predictor-level directional reasoning from the LLM, we used the following system prompt.

System Prompt for Directional Alignment

You are a spatial data analyst. Your task is to provide reasoning for characteristic mobility outcomes within a small geographic region (Census Block Group).

Output ONLY one XML document that starts with <response> and ends with </response>. This document must contain EXACTLY two child elements in this order:
<reasoning>, <assessment>.

Primary Task

Identify the directional push of every predictor on the mobility outcome for this CBG.

Analytical Framework (Two-Step Evaluation)

For each predictor, apply the following logic:

1. Classify the relationship: Is the feature a Facilitator (positive association) or a Restrictor (inverse association) for the outcome?
2. Apply the instance value:
 - If a feature is a Facilitator and its value is high or very_high, then push="+"
 - If a feature is a Facilitator and its value is low or very_low, then push="-"
 - If a feature is a Restrictor and its value is high or very_high, then push="-"
 - If a feature is a Restrictor and its value is low or very_low, then push="+"

<reasoning>

Provide a concise explanation. Explicitly list which predictors you classified as Facilitators and which you classified as Restrictors based on the outcome definition.

</reasoning>

<assessment>

List EVERY predictor provided in the user prompt as an <item/>.

Each <item/> must have these attributes:

- feature: the exact name of the feature as provided
- val: the observed bucket (very_low|low|neutral|high|very_high)
- push: "+", "-", or "0"

</assessment>

Model	Parse success rate	Used in analysis
claude_sonnet_4.5	99.999%	Yes
gemma3_27b	99.980%	Yes
gemma3_12b	99.813%	Yes
qwen3_30b_a3b	99.315%	Yes
qwen3_8b	99.180%	Yes
gptoss_20b_low	98.814%	Yes
qwen3_4b	96.222%	Yes
qwen3_5_35b_a3b	88.769%	No
qwen_3_5_9b	87.350%	No
deepseek_v3.2	72.187%	No
llama_3_1_8b_inst	56.273%	No
gptoss_120b	20.744%	No

Table D.1: Parse success rates across evaluated LLMs. Models with parse success rate below 95% were excluded from downstream analysis.

Bibliography

- [1] Saad Mohammad Abrar, Naman Awasthi, Daniel Smolyak, and Vanessa Frias-Martinez. Analysis of performance improvements and bias associated with the use of human mobility data in covid-19 case prediction models. *ACM Journal on Computing and Sustainable Societies*, 2023.
- [2] Saad Mohammad Abrar, Naman Awasthi, Daniel Smolyak, and Vanessa Frias-Martinez. Auditing the fairness of covid-19 forecast hub case prediction models. *arXiv preprint arXiv:2405.14891*, 2024.
- [3] Saad Mohammad Abrar, Kazi Tasnim Zinat, Naman Awasthi, and Vanessa Frias-Martinez. Covid-19’s unequal toll: An assessment of small business impact disparities with respect to ethnorace in metropolitan areas in the us using mobility data. *arXiv preprint arXiv:2405.11121*, 2024.
- [4] Aniruddha Adiga, Gursharn Kaur, Lijing Wang, Benjamin Hurt, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Lewis, and Madhav V Marathe. Phase-informed bayesian ensemble models improve performance of covid-19 forecasts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15647–15653, 2023.
- [5] Haldun Akoglu. User’s guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.
- [6] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. *Nature*, 587(7834):402–407, 2020.
- [7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.
- [8] Apple. Apple mobility data. <https://covid19.apple.com/mobility>, 2020.
- [9] Irina Arhipova, Gundars Berzins, Edgars Brekis, Juris Binde, Martins Opmanis, Aldis Erglis, and Evija Ansonska. Mobile phone data statistics as a dynamic proxy indicator in assessing regional economic activity and human commuting patterns. *Expert Systems*, 37(5):e12530, 2020.
- [10] Sercan Arik, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long Le, Vikas Menon, Shashank Singh, Leyou Zhang, Martin Nikoltchev, et al. Interpretable sequence learning for covid-19 forecasting. *Advances in Neural Information Processing Systems*, 33:18807–18818, 2020.

- [11] Hikaru Asano, Hiroki Ouchi, Akira Kasuga, and Ryo Yonetani. Mobqa: A benchmark dataset for semantic understanding of human mobility data through question answering. *arXiv preprint arXiv:2508.11163*, 2025.
- [12] Aregu Asmare and Shimelis Zewdie. Omnichannel retailing strategy: a systematic review. *The International Review of Retail, Distribution and Consumer Research*, 32(1):59–79, 2022.
- [13] Hamada S Badr and Lauren M Gardner. Limitations of using mobile phone data to model covid-19 transmission in the usa. *The Lancet Infectious Diseases*, 21(5):e113, 2021.
- [14] Hamada S Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M Squire, and Lauren M Gardner. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.
- [15] Rachel E Baker, Sang Woo Park, Wenchang Yang, Gabriel A Vecchi, C Jessica E Metcalf, and Bryan T Grenfell. The impact of covid-19 nonpharmaceutical interventions on the future dynamics of endemic infections. *Proceedings of the National Academy of Sciences*, 117(48):30547–30553, 2020.
- [16] Scott R Baker, Nicholas Bloom, Steven J Davis, and Stephen J Terry. Covid-induced economic uncertainty. Technical report, National Bureau of Economic Research, 2020.
- [17] Dylan Balla-Elliott, Zoë B Cullen, Edward L Glaeser, Michael Luca, and Christopher Stanton. Determinants of small business reopening decisions after covid restrictions were lifted. *Journal of Policy Analysis and Management*, 41(1):278–317, 2022.
- [18] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.
- [19] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [20] Alexander W Bartik, Marianne Bertrand, Zoe Cullen, Edward L Glaeser, Michael Luca, and Christopher Stanton. The impact of covid-19 on small business outcomes and expectations. *Proceedings of the national academy of sciences*, 117(30):17656–17666, 2020.
- [21] David R Bell, Santiago Gallino, and Antonio Moreno. Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science*, 64(4):1629–1651, 2018.
- [22] Ciro Beneduce, Bruno Lepri, and Massimiliano Luca. Large language models are zero-shot next location predictors. *IEEE Access*, 2025.
- [23] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, 5(1):1–5, 2015.
- [24] Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Urban mobility assessment using llms. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 67–79, 2024.

- [25] Ulrike Boehmer, Nancy R Kressin, Dan R Berlowitz, Cindy L Christiansen, Lewis E Kazis, and Judith A Jones. Self-reported vs administrative race/ethnicity data and study results. *American journal of public health*, 92(9):1471–1472, 2002.
- [26] David Bounie, Youssouf Camara, and John W Galbraith. Consumers’ mobility, expenditure and online-offline substitution response to covid-19: Evidence from french transaction data. *Available at SSRN 3588373*, 2022.
- [27] Mohamed Amine Bouzaghrane, Hassan Obeid, Marta González, and Joan Walker. Human mobility reshaped? deciphering the impacts of the covid-19 pandemic on activity patterns, spatial habits, and schedule habits. *EPJ Data Science*, 13(1):24, 2024.
- [28] Stacey Bricka, Timothy Reuscher, Paul Schroeder, Mitchell Fisher, Justina Beard, and Xiaoyuan Layla Sun. Summary of travel trends: 2022 national household travel survey. 2024.
- [29] Rebecca Brough, Matthew Freedman, and David C Phillips. Understanding socioeconomic disparities in travel behavior during the covid-19 pandemic. *Journal of Regional Science*, 61(4):753–774, 2021.
- [30] Caroline O Buckee, Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, Urs Gasser, Yonatan H Grad, Bryan Grenfell, M Elizabeth Halloran, Moritz UG Kraemer, et al. Aggregated mobility data could help fight covid-19. *Science*, 368(6487):145–146, 2020.
- [31] Raymundo M Campos-Vazquez and Gerardo Esquivel. Consumption and geographic mobility in pandemic times. evidence from mexico. *Review of Economics of the Household*, 19(2):353–371, 2021.
- [32] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.
- [33] CDC. Cdc covid-19 visualization. <https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-math-modeling.html>, 2020.
- [34] CDC. Forecast hub. <https://covid19forecasthub.org/>, 2020.
- [35] CDC. PLACES: Local Data for Better Health (Place Data 2022 Release). https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Place-Data-202/eav7-hnsx/about_data, 2022. Accessed: 2025-01-06.
- [36] CDC. Health disparities, 2023. URL https://www.cdc.gov/nchs/nvss/vsrr/covid19/health_disparities.htm#CountyRaceHispanicOrigin. Last accessed January 2024.
- [37] CDC. Cdc national center for health statistics: Urban-rural classification scheme for counties. https://www.cdc.gov/nchs/data_access/urban_rural.htm#2013_Urban-Rural_Classification_Scheme_for_Counties, 2023.

- [38] CDC et al. Underlying conditions and the higher risk for severe covid-19, 2024.
- [39] CensusBureau. American community survey. <https://www.census.gov/programs-surveys/acs>, 2023.
- [40] Brian Cepparulo. Is mobility a good proxy for economic activity? *Available at SSRN 4638714*, 2023.
- [41] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David B. Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 2020. doi: 10.1038/s41586-020-2923-3.
- [42] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.
- [43] Serina Chang, Mandy L Wilson, Bryan Lewis, Zakaria Mehrab, Komal K Dudakiya, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, et al. Supporting covid-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2632–2642, 2021.
- [44] Raphaël Charron-Chénier, Joshua J Fink, and Lisa A Keister. Race and consumption: black and white disparities in household spending. *Sociology of Race and Ethnicity*, 3(1):50–67, 2017.
- [45] Yile Chen, Yicheng Tao, Yue Jiang, Shuai Liu, Han Yu, and Gao Cong. Enhancing large language models for mobility analytics with semantic location tokenization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 262–273, 2025.
- [46] Raj Chetty, John N Friedman, and Michael Stepner. The economic impacts of covid-19: Evidence from a new public database built using private sector data. *The Quarterly Journal of Economics*, 139(2):829–889, 2024.
- [47] Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022.
- [48] Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E Ho. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for covid-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, 2021.
- [49] Victor Couture, Jonathan I Dingel, Allison Green, Jessie Handbury, and Kevin R Williams. Jue insight: Measuring movement and social contact with smartphone data: a real-time application to covid-19. *Journal of Urban Economics*, 127:103328, 2022.
- [50] Marc Cowling, Weixi Liu, Andrew Ledger, and Ning Zhang. What really happens to small and medium-sized enterprises in a global economic recession? uk evidence on sales and job dynamics. *International Small Business Journal*, 33(5):488–513, 2015.

- [51] Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Ariane Stark, Apurv Shah, Nutch Wattanachit, Martha W Zorn, Nicholas G Reich, and US COVID-19 Forecast Hub Consortium. The united states covid-19 forecast hub dataset. *Scientific Data*, 2022. doi: 10.1101/2021.11.04.21265886. URL <https://doi.org/10.1038/s41597-022-01517-w>.
- [52] David S Curtis, Alessandro Rigolon, Dorothy L Schmalz, and Barbara B Brown. Policy and environmental predictors of park visits during the first months of the covid-19 pandemic: getting out while staying in. *Environment and Behavior*, 54(2):487–515, 2022.
- [53] Tiago Tiburcio da Silva, Rodrigo Francisquini, and Mariá CV Nascimento. Meteorological and human mobility data on predicting covid-19 cases by a novel hybrid decomposition method with anomaly detection analysis: A case study in the capitals of brazil. *Expert Systems with Applications*, 182:115190, 2021.
- [54] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1376, 2013.
- [55] Marina Del Rios, Sylvia Puente, Pamela Vergara-Rodriguez, and Noreen Sugrue. Invisibilidad de los latinos en la pandemia. *AMA Journal of Ethics*, pages 289–295, 2022.
- [56] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.
- [57] Megan D Douglas, Ebony Respress, Anne H Gaglioti, Chaohua Li, Mitchell A Blount, Jammie Hopkins, Peter T Baltrus, Robina Josiah Willock, Lee S Caplan, Daniel E Dawes, et al. Variation in reporting of the race and ethnicity of covid-19 cases and deaths across us states: April 12, 2020, and november 9, 2020. *American Journal of Public Health*, 111(6):1141–1148, 2021.
- [58] Yunhan Du, Takaaki Aoki, and Naoya Fujiwara. A review of human mobility: Linking data, models, and real-world applications. *Journal of Computational Social Science*, 8(4):90, 2025.
- [59] Haga Elimam et al. The role of small businesses (small scale economic projects) in alleviating the acuity of unemployment. *International Business Research*, 10(3):120–132, 2017.
- [60] Abdolmajid Erfani and Vanessa Frias-Martinez. A fairness assessment of mobility-based covid-19 case prediction models. *Plos one*, 18(10):e0292090, 2023.
- [61] Niklas Eriksson and Asle Fagerstrøm. Exploring showrooming behavior in electronics and clothing retail. Research report, Yrkeshögskolan Arcada, 2019.
- [62] Robert Fairlie. The impact of covid-19 on small business owners: Evidence from the first three months after widespread social-distancing restrictions. *Journal of economics & management strategy*, 29(4):727–740, 2020.
- [63] Robert W Fairlie. *The Impacts of COVID-19 on Racial Disparities in Small Business Earnings*. Office of Advocacy, US Small Business Administration, 2022.

- [64] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqu Guo, and Yong Li. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5413–5424, 2025.
- [65] Luzhao Feng, Ting Zhang, Qing Wang, Yiran Xie, Zhibin Peng, Jiandong Zheng, Ying Qin, Muli Zhang, Shengjie Lai, Dayan Wang, et al. Impact of covid-19 outbreaks and interventions on influenza in china and the united states. *Nature communications*, 12(1):1–8, 2021.
- [66] Jesús Fernández-Villaverde and Charles I Jones. Macroeconomic outcomes and covid-19: a progress report. Technical report, National Bureau of Economic Research, 2020.
- [67] Kevin Foster, Claire Greene, and Joanna Stavins. 2023 survey and diary of consumer payment choice: Summary results. *Federal Reserve Bank of Atlanta Research Data Report*, (24-1), 2024.
- [68] John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.
- [69] John Fox and Sanford Weisberg. *An R companion to applied regression*. Sage publications, 2018.
- [70] Vanessa Frias-Martinez and Jesus Virseda. Cell phone analytics: Scaling human behavior studies into the millions. *Information Technologies & International Development*, 9(2):pp–35, 2013.
- [71] Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Socio-economic levels and human mobility. In *Qual meets quant workshop-QMQ*, pages 1–6, 2010.
- [72] Vanessa Frias-Martinez, Victor Soto, Jesus Virseda, and Enrique Frias-Martinez. Computing cost-effective census maps from cell phone traces. In *Workshop on pervasive urban applications*, 2012.
- [73] Vanessa Frias-Martinez, Jesus Virseda, and Aldo Gomero. Mobilizing education: evaluation of a mobile learning tool in a low-income school. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 441–450, 2012.
- [74] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly covid-19 cases in germany. *Scientific Reports*, 12(1):1–18, 2022.
- [75] Cheng Fu, Grant McKenzie, Vanessa Frias-Martinez, and Kathleen Stewart. Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems*, 72:25–37, 2018.
- [76] Joseph Galasso, Duy M Cao, and Robert Hochberg. A random forest model for forecasting regional covid-19 cases utilizing reproduction number estimates and demographic data. *Chaos, Solitons & Fractals*, 156:111779, 2022.

- [77] Michael D Garber, Katie Labgold, and Michael R Kramer. On selection bias in comparison measures of smartphone-generated population mobility: an illustration of no-bias conditions with a commercial data source. *Annals of epidemiology*, 70:16–22, 2022.
- [78] Santi García-Cremades, Juan Morales-García, Rocío Hernández-Sanjaime, Raquel Martínez-España, Andrés Bueno-Crespo, Enrique Hernández-Orallo, José J López-Espín, and José M Cecilia. Improving prediction of covid-19 evolution by fusing epidemiological and mobility data. *Scientific Reports*, 11(1):1–16, 2021.
- [79] Lauren Gardner, Jeremy Ratcliff, Ensheng Dong, and Aaron Katz. A need for open public data standards and sharing in light of covid-19. *The Lancet Infectious Diseases*, 21(4):e80, 2021.
- [80] Oliver Gatalo, Katie K Tseng, Alisa Hamilton, Gary Lin, Eili Y. Klein, and Cdc MInD-Healthcare Program. Associations between phone mobility data and covid-19 cases. *Lancet Infectious Diseases*, 2020. doi: 10.1016/s1473-3099(20)30725-8.
- [81] Jay Ghurye, Gautier Krings, and Vanessa Frias-Martinez. A framework to model human behavior at large scale during natural disasters. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, volume 1, pages 18–27. IEEE, 2016.
- [82] Letian Gong, Yan Lin, Xinyue Zhang, Yiwen Lu, Xuedi Han, Yichen Liu, Shengnan Guo, Youfang Lin, and Huaiyu Wan. Mobility-llm: Learning visiting intentions and travel preference from human mobility data with large language models. *Advances in Neural Information Processing Systems*, 37:36185–36217, 2024.
- [83] Ana Belén Rodríguez González, Javier Burrieza-Galán, Juan José Vinagre Díaz, Inés Peirats de Castro, Mark Richard Wilby, and Oliva Garcia Cantú-Ros. Using app usage data from mobile devices to improve activity-based travel demand models. *IEEE Transactions on Big Data*, 10(5):633–643, 2024.
- [84] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.
- [85] Google. Google mobility data. <https://www.google.com/covid19/mobility/>, 2020.
- [86] Austan Goolsbee and Chad Syverson. Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of public economics*, 193:104311, 2021.
- [87] Cary P Gross, Utibe R Essien, Saamir Pasha, Jacob R Gross, Shi-yi Wang, and Marcella Nunez-Smith. Racial and ethnic disparities in population-level covid-19 mortality. *Journal of general internal medicine*, 35:3097–3099, 2020.
- [88] Grace Guan, Yotam Dery, Matan Yechezkel, Irad Ben-Gal, Dan Yamin, and Margaret L Brandeau. Early detection of covid-19 outbreaks using human mobility data. *PloS one*, 16(7):e0253865, 2021.
- [89] Furkan Gursoy and Ioannis A Kakadiaris. Error parity fairness: Testing for group fairness in regression tasks. *arXiv preprint arXiv:2208.08279*, 2022.

- [90] Torsten Hägerstrand. What about people in regional science. *Transport Sociology: Social aspects of transport planning*, pages 143–158, 1970.
- [91] Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.
- [92] Yuanmo He and Milena Tsvetkova. High socioeconomic status is associated with diverse consumption across brands and price levels. *arXiv preprint arXiv:2506.13840*, 2025.
- [93] Marco Hernandez, Lingzi Hong, Vanessa Frias-Martinez, Andrew Whitby, and Enrique Frias-Martinez. Estimating poverty using cell phone data: evidence from guatemala. *World Bank Policy Research Working Paper*, (7969), 2017.
- [94] Lingzi Hong and Vanessa Frias-Martinez. Modeling and predicting evacuation flows during hurricane irma. *EPJ Data Science*, 9(1):29, 2020.
- [95] Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Topic models to infer socioeconomic maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [96] Lingzi Hong, Cheng Fu, Paul Torrens, and Vanessa Frias-Martinez. Understanding citizens’ and local governments’ digital communications during natural disasters: the case of snowstorms. In *Proceedings of the 2017 ACM on web science conference*, pages 141–150, 2017.
- [97] Md Shahadat Hossain, Mahmudur Rahman Fatmi, and Corrie Elizabeth Thirkell. How will in-person and online grocery shopping and meal consumption activities evolve after covid-19? *Transportation Research Record*, 2678(12):337–348, 2024.
- [98] Xiao Hou, Song Gao, Qin Li, Yuhao Kang, Nan Chen, Kaiping Chen, Jinmeng Rao, Jordan S Ellenberg, and Jonathan A Patz. Intracounty modeling of covid-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences*, 118(24):e2020524118, 2021.
- [99] Justin T Huang, Masha Krupenkin, David Rothschild, and Julia Lee Cunningham. The cost of anti-asian racism during the covid-19 pandemic. *Nature Human Behaviour*, pages 1–14, 2023.
- [100] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
- [101] Cornelia Ilin, Sébastien Annan-Phan, Xiao Hui Tai, Shikhar Mehra, Solomon Hsiang, and Joshua E Blumenstock. Public mobility data enables covid-19 forecasting and management at local and global scales. *Scientific reports*, 11(1):1–11, 2021.
- [102] Sibren Isaacman, Vanessa Frias-Martinez, and Enrique Frias-Martinez. Modeling human migration patterns during drought conditions in la guajira, colombia. In *Proceedings of the 1st ACM SIGCAS conference on computing and sustainable societies*, pages 1–9, 2018.

- [103] Caroline Jaret and Baptiste Meunier. Nowcasting world gdp growth with high-frequency data. *Journal of Forecasting*, 41(6):1181–1200, 2022.
- [104] Farah Kader and Clyde Lanford Smith. Participatory approaches to addressing missing covid-19 race and ethnicity data. *International Journal of Environmental Research and Public Health*, 18(12):6559, 2021.
- [105] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jinmeng Rao, and Jake Kruse. Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic. *Scientific data*, 7(1):1–13, 2020.
- [106] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O’Banion. Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*, 2020.
- [107] Nishant Kishore, Aimee R Taylor, Pierre E Jacob, Navin Vembar, Ted Cohen, Caroline O Buckee, and Nicolas A Menzies. Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for sars-cov-2 transmission in the usa: a population-based study. *The Lancet Digital Health*, 4(1):e27–e36, 2022.
- [108] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.
- [109] Cheng-Pin Kuo and Joshua S Fu. Evaluating the impact of mobility on covid-19 pandemic with machine learning hybrid predictions. *Science of The Total Environment*, 758:144151, 2021.
- [110] Shu-Chen Kuo, Shu-Man Shih, Li-Hsin Chien, and Chao A Hsiung. Collateral benefit of covid-19 control measures on influenza activity, taiwan. *Emerging infectious diseases*, 26(8):1928, 2020.
- [111] Matthew Le, Mark Ibrahim, Levent Sagun, Timothee Lacroix, and Maximilian Nickel. Neural relational autoregression for high-resolution covid-19 forecasting. *Facebook AI Research*, 2020.
- [112] Timothy F Leslie, Hossein Amiri, and Andreas Züfle. Exploring economic sectoral dynamics through high-resolution mobility data. *arXiv preprint arXiv:2506.13985*, 2025.
- [113] Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D Salim. Large language models for next point-of-interest recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1463–1472, 2024.
- [114] Siyu Li, Toan Tran, Haowen Lin, John Krumm, Cyrus Shahabi, Lingyi Zhao, Khurram Shafique, and Li Xiong. Geo-llama: Leveraging llms for human mobility trajectory generation with spatiotemporal constraints. *arXiv preprint arXiv:2408.13918*, 2024.
- [115] Xuchuan Li, Fei Huang, Jianrong Lv, Zhixiong Xiao, Guolong Li, and Yang Yue. Be more real: Travel diary generation using llm agents and individual profiles. *arXiv preprint arXiv:2407.18932*, 2024.

- [116] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5351–5362, 2024.
- [117] Tianming Liu, Manzi Li, and Yafeng Yin. Aligning llm with human travel choices: a persona-based embedding learning approach. *arXiv preprint arXiv:2505.19003*, 2025.
- [118] Yifan Liu, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, Chris Stanford, and Jiaqi Ma. Human mobility modeling with household coordination activities under limited information via retrieval-augmented llms. In *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, pages 951–958. IEEE, 2025.
- [119] Benjamin Lucas, Behzad Vahedi, and Morteza Karimzadeh. A spatiotemporal machine learning approach to forecasting covid-19 incidence at the county level in the usa. *International Journal of Data Science and Analytics*, 15(3):247–266, 2023.
- [120] Yuxiao Luo, Zhongcai Cao, Xin Jin, Kang Liu, and Ling Yin. Deciphering human mobility: Inferring semantics of trajectories with large language models. In *2024 25th IEEE international conference on mobile data management (MDM)*, pages 289–294. IEEE, 2024.
- [121] Haoxuan Ma, Xishun Liao, Yifan Liu, Qinhua Jiang, Chris Stanford, Shangqing Cao, and Jiaqi Ma. Learning universal human mobility patterns with a foundation model for cross-domain data fusion. *Transportation Research Part C: Emerging Technologies*, 180:105311, 2025.
- [122] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023.
- [123] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.
- [124] Marco Marozzi. Nonparametric simultaneous tests for location and scale testing: a comparison of several methods. *Communications in Statistics-Simulation and Computation*, 42(6):1298–1317, 2013.
- [125] Kohei Matsumura, Yusuke Oh, Tomohiro Sugo, and Koji Takahashi. Nowcasting economic activity with mobility data. *Journal of the Japanese and International Economies*, 73:101327, 2024.
- [126] Zakaria Mehrab, Aniruddha Adiga, Madhav V Marathe, Srinivasan Venkatramanan, and Samarth Swarup. Evaluating the utility of high-resolution proximity metrics in predicting the spread of covid-19. *ACM Transactions on Spatial Systems and Algorithms*, 2021.
- [127] Umberto Minora, Stefano Maria Iacus, Filipe Batista e Silva, Francesco Sermi, and Spyridon Spyridon. Nowcasting tourist nights spent using innovative human mobility data. *PloS one*, 18(10):e0287063, 2023.

- [128] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228, 2024.
- [129] Daisuke Moriwaki. Nowcasting unemployment rates with smartphone gps data. In *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories*, pages 21–33. Springer, 2019.
- [130] Yves Nievergelt. The concept of elasticity in economics. *Siam Review*, 25(2):261–265, 1983.
- [131] Behnam Nikparvar, Md Rahman, Faizeh Hatami, Jean-Claude Thill, et al. Spatio-temporal prediction of the covid-19 pandemic in us counties: modeling with a deep lstm neural network. *Scientific reports*, 11(1):1–12, 2021.
- [132] Kristen Nixon, Sonia Jindal, Felix Parker, Maximilian Marshall, Nicholas G Reich, Kimia Ghobadi, Elizabeth C Lee, Shaun Truelove, and Lauren Gardner. Real-time covid-19 forecasting: challenges and opportunities of model performance and translation. *The Lancet Digital Health*, 4(10):e699–e701, 2022.
- [133] OECD. The unequal impact of covid-19: a spotlight on frontline workers, migrants and racial/ethnic minorities, 2022.
- [134] Juhwan Oh, Hwa-Young Lee, Quynh Long Khuong, Jeffrey F Markuns, Chris Bullen, Osvaldo Enrique Artaza Barrios, Seung-sik Hwang, Young Sahng Suh, Judith McCool, S Patrick Kachur, et al. Mobility restrictions were associated with reductions in covid-19 incidence early in the pandemic: evidence from a real-time evaluation in 34 countries. *Scientific reports*, 11(1):1–17, 2021.
- [135] Tatsushi Oka, Wei Wei, and Dan Zhu. The effect of human mobility restrictions on the covid-19 transmission network in china. *PloS one*, 16(7):e0254403, 2021.
- [136] Paul Ong, Andre Comandom, Nicholas DiRago, and Lauren Harper. Covid-19 impacts on minority businesses and systemic inequality. 2020.
- [137] Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92, 2016.
- [138] Sen Pei and Jeffrey Shaman. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *MedRxiv*, pages 2020–03, 2020.
- [139] Nicola Perra. Non-pharmaceutical interventions during the covid-19 pandemic: A review. *Physics Reports*, 913:1–52, 2021.
- [140] Cristian Podesta, Natalie Coleman, Amir Esmalian, Faxi Yuan, and Ali Mostafavi. Quantifying community resilience based on fluctuations in visits to points-of-interest derived from digital trace data. *Journal of the Royal Society Interface*, 18(177):20210158, 2021.
- [141] Matthew Quint, David Rogers, and Rick Ferguson. Showrooming and the rise of the mobile-assisted shopper. *Columbia Business School. Center on Global Brand Leadership*, 2013.

- [142] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
- [143] Chotirat Ann Ralanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. Mining time series data. In *Data mining and knowledge discovery handbook*, pages 1069–1103. Springer, 2005.
- [144] Michael Reibel and Moira Regelson. Neighborhood racial and ethnic change: The time dimension in segregation. *Urban Geography*, 32(3):360–382, 2011.
- [145] Peter Richterich. Severe underestimation of covid-19 case numbers: effect of epidemic growth rate and test restrictions. *MedRxiv*, pages 2020–04, 2020.
- [146] Weston C Roda, Marie B Varughese, Donglin Han, and Michael Y Li. Why is it difficult to accurately predict the covid-19 epidemic? *Infectious disease modelling*, 5:271–281, 2020.
- [147] Can Rong, Jingtao Ding, and Yong Li. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Computing Surveys*, 57(1):1–49, 2024.
- [148] Alberto Rubio, Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. Human mobility in advanced and developing economies: A comparative analysis. In *2010 AAAI Spring Symposium Series*, 2010.
- [149] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [150] Safegraph. Safegraph mobility data. <https://www.safegraph.com/guides/mobility-data>, 2020.
- [151] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
- [152] James Robert Ezequiel Sampi Bravo and Charl Jooste. Nowcasting economic activity in times of covid-19: An approximation from the google community mobility report. *World Bank Policy Research Working Paper*, (9247), 2020.
- [153] Sanja Šćepanović, Igor Mishkovski, Pan Hui, Jukka K Nurminen, and Antti Ylä-Jääski. Mobile phone call data as a regional socio-economic proxy indicator. *PloS one*, 10(4):e0124160, 2015.
- [154] Frank Schlosser, Vedran Sekara, Dirk Brockmann, and Manuel Garcia-Herranz. Biases in human mobility data impact epidemic modeling. *arXiv preprint arXiv:2112.12521*, 2021.
- [155] Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Yong Li. Beyond imitation: Generating human mobility from context-aware reasoning with large language models. *arXiv preprint arXiv:2402.09836*, 2024.

- [156] Jagdish Sheth. Impact of covid-19 on consumer behavior: Will the old habits return or die? *Journal of business research*, 117:280–283, 2020.
- [157] Linxuan Shi and Zhengtian Xu. Dine in or takeout? trends on restaurant service demand amid the covid-19 pandemic. *Service Science*, 16(4):241–271, 2024.
- [158] Jacob M Souch and Jeralynn S Cossman. A commentary on rural-urban disparities in covid-19 testing rates per 100,000 and risk factors. *The Journal of Rural Health*, 37(1):188, 2021.
- [159] Alessandro Spelta and Paolo Pagnottoni. Mobility-based real-time economic monitoring amid the covid-19 pandemic. *Scientific Reports*, 11(1):13069, 2021.
- [160] Ben Sprung-Keyser, Nathaniel Hendren, Sonya Porter, et al. *The radius of economic opportunity: Evidence from migration and local labor markets*. US Census Bureau, Center for Economic Studies, 2022.
- [161] Julia Stoyanovich and Bill Howe. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3), 2019.
- [162] Thomas C Tsai, Sercan Arik, Benjamin H Jacobson, Jinsung Yoon, Nate Yoder, Dario Sava, Margaret Mitchell, Garth Graham, and Tomas Pfister. Algorithmic fairness in pandemic forecasting: lessons from covid-19. *NPJ Digital Medicine*, 5(1):59, 2022.
- [163] Zander S Venter, Adam Sadilek, Charlotte Stanton, David N Barton, Kristin Aunan, Sourangsu Chowdhury, Aaron Schneider, and Stefano Maria Iacus. Mobility in blue-green spaces does not predict covid-19 transmission: a global analysis. *International Journal of Environmental Research and Public Health*, 18(23):12567, 2021.
- [164] Marcos R Vieira, Enrique Frias-Martinez, Petko Bakalov, Vanessa Frias-Martinez, and Vassilis J Tsotras. Querying spatio-temporal patterns in mobile phone-call databases. In *2010 Eleventh International Conference on Mobile Data Management*, pages 239–248. IEEE, 2010.
- [165] Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. Large language models as urban residents: An llm agent framework for personal mobility generation. *Advances in Neural Information Processing Systems*, 37:124547–124574, 2024.
- [166] Lijing Wang, Xue Ben, Aniruddha Adiga, Adam Sadilek, Ashish Tendulkar, Srinivasan Venkataramanan, Anil Vullikanti, Gaurav Aggarwal, Alok Talekar, Jiangzhuo Chen, et al. Using mobility data to understand and forecast covid19 dynamics. *medRxiv*, 2020.
- [167] Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*, 2023.
- [168] Zhaonan Wang, Renhe Jiang, Hao Xue, Flora D Salim, Xuan Song, and Ryosuke Shibasaki. Event-aware multimodal mobility nowcasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4228–4236, 2022.

- [169] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.
- [170] Jiahui Wu, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Spatial sensitivity analysis for urban hotspots using cell phone traces. *Environment and Planning B: Urban Analytics and City Science*, 2021.
- [171] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Enhancing short-term crime prediction with human mobility flows and deep learning architectures. *EPJ Data Science*, 11(1):53, 2022.
- [172] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, and Vanessa Frías-Martínez. Auditing the fairness of place-based crime prediction models implemented with deep learning approaches. *Computers, Environment and Urban Systems*, 102:101967, 2023.
- [173] Lun Wu, Liu Yang, Zhou Huang, Yaoli Wang, Yanwei Chai, Xia Peng, and Yu Liu. Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems*, 77:101368, 2019.
- [174] Neo Wu, Xue Ben, Bradley Green, Kathryn Rough, Srinivasan Venkatramanan, Madhav Marathe, Paul Eastham, Adam Sadilek, and Shawn O’Banion. Predicting onset of covid-19 with mobility-augmented seir model. *medRxiv*, 2020.
- [175] Xinhua Wu and Qi R Wang. Popular llms amplify race and gender disparities in human mobility. *arXiv preprint arXiv:2411.14469*, 2024.
- [176] Fengli Xu, Qi Wang, Esteban Moro, Lin Chen, Arianna Salazar Miranda, Marta C González, Michele Tizzoni, Chaoming Song, Carlo Ratti, Luis Bettencourt, et al. Using human mobility data to quantify experienced urban inequalities. *Nature human behaviour*, 9(4):654–664, 2025.
- [177] Takahiro Yabe, Yunchang Zhang, and Satish V Ukkusuri. Quantifying the economic impact of disasters on businesses using human mobility data: a bayesian causal inference approach. *EPJ Data Science*, 9(1):36, 2020.
- [178] Takahiro Yabe, Bernardo Garcia Bulle Bueno, Morgan Frank, Alex Pentland, and Esteban Moro. Behavior-based dependency networks between places shape urban economic resilience. *arXiv preprint arXiv:2311.18108*, 2023.
- [179] Takahiro Yabe, Bernardo García Bulle Bueno, Morgan R Frank, Alex Pentland, and Esteban Moro. Behaviour-based dependency networks between places shape urban economic resilience. *Nature human behaviour*, 9(3):496–506, 2025.
- [180] Jing Yang, YA Zheng, Xi Gou, Ke Pu, Zhaofeng Chen, Qinghong Guo, Rui Ji, Haojia Wang, Yuping Wang, and Yongning Zhou. Prevalence of comorbidities and its effects in patients infected with sars-cov-2: a systematic review and meta-analysis. *International journal of infectious diseases*, 94:91–95, 2020.

- [181] Yanni Yang, Alex Pentland, and Esteban Moro. Identifying latent activity behaviors and lifestyles using mobility data to describe urban dynamics. *EPJ Data Science*, 12(1):15, 2023.
- [182] Nazar Zaki and Elfadil A Mohamed. The estimations of the covid-19 incubation period: A scoping reviews of the literature. *Journal of infection and public health*, 14(5):638–646, 2021.
- [183] Choujun Zhan, Yufan Zheng, Zhikang Lai, Tianyong Hao, and Bing Li. Identifying epidemic spreading dynamics of covid-19 by pseudocoevolutionary simulated annealing optimizers. *Neural Computing and Applications*, 33(10):4915–4928, 2021.
- [184] Yecheng Zhang, Rong Zhao, Zimu Huang, Xinyu Wang, Yue Ma, and Ying Long. Genai models capture urban science but oversimplify complexity. *arXiv preprint arXiv:2505.13803*, 2025.
- [185] Yanli Zhang-James, Jonathan Hess, Asif Salkin, Dongliang Wang, Samuel Chen, Peter Winkelstein, Christopher P Morley, and Stephen V Faraone. A seq2seq model to forecast the covid-19 cases, deaths and reproductive r numbers in us counties. *Research Square*, 2021.
- [186] Dong Zhao, Adriana-Simona Mihăiță, Yuming Ou, Hanna Grzybowska, and Mo Li. Origin–destination matrix estimation for public transport: A multi-modal weighted graph approach. *Transportation Research Part C: Emerging Technologies*, 165:104694, 2024.
- [187] Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, and Quanquan Gu. Epidemic model guided machine learning for covid-19 forecasts in the united states. *MedRxiv*, pages 2020–05, 2020.